Klieme, Eckhard [Hrsg.]; Leutner, Detlev [Hrsg.]; Kenk, Martina [Hrsg.]

Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes

Weinheim ; Basel : Beltz 2010, 312 S. - (Zeitschrift für Pädagogik, Beiheft; 56)

Zeitschrift für Pädagogik · 56. Beiheft

Zeitschrift für Pädagogik · 56. Beiheft

Kompetenzmodellierung

Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes

Herausgegeben von
Eckhard Klieme, Detlev Leutner und Martina Kenk

BELTZ

Herstellung: Lore Amann
Gesamtherstellung: Druckhaus „Thomas Müntzer“, Bad Langensalza
Printed in Germany
ISSN 0514-2717
Bestell-Nr. 41157

Inhaltsverzeichnis

Mathematische Kompetenzen

Naturwissenschaftliche Kompetenzen

Sprachliche und Lesekompetenzen

Fächerübergreifende Kompetenzen

Lehrerkompetenzen

Eckhard Klieme/Detlev Leutner/Martina Kenk

Kompetenzmodellierung

Eine aktuelle Zwischenbilanz des DFG-Schwerpunktprogramms[1]

Die Vermittlung von Kompetenzen ist ein zentrales Ziel schulischer und beruflicher Bildung. Der Beitrag von Bildung und Ausbildung zur gesellschaftlichen Entwicklung hängt von den erreichten Kompetenzen der Absolventinnen und Absolventen in konkreten Anforderungsbereichen ab. Beispiele sind das Beherrschen von Mutter- und Fremdsprachen, der Gebrauch mathematischer Modelle, naturwissenschaftliches Verständnis oder das Lernen und Problemlösen in alltags- und berufsrelevanten Bereichen. Hinsichtlich der Förderung solcher Kompetenzen weist das deutsche Bildungswesen, wie internationale Vergleichsstudien gezeigt haben, Schwächen auf. In der empirischen Bildungsforschung werden erziehungswissenschaftliche, psychologische und fachdidaktische Grundlagen, psychometrische Modelle sowie konkrete Messverfahren entwickelt, mit denen solche internationale Vergleichsstudien erst möglich werden. Im nationalen Rahmen dient Kompetenzmessung zunehmend als Grundlage für bildungspolitische Steuerung wie auch zur Begründung von pädagogischen und didaktischen Entscheidungen im Einzelfall (z.B. Förderempfehlungen). Der Messung von Kompetenzen kommt somit eine Schlüsselfunktion für die Optimierung von Bildungsprozessen und für die Weiterentwicklung des Bildungswesens zu. Dennoch wird in der Bildungspraxis und Bildungspolitik häufig unterschätzt, wie anspruchsvoll die empirische Erfassung von Kompetenzen aus theoretischer und methodischer Perspektive ist. Die Entwicklung sowohl theoretisch wie auch empirisch fundierter Kompetenzmodelle als Ausgangspunkt für die Entwicklung adäquater Messverfahren stellt immer noch eine Herausforderung dar. Auch im internationalen Rahmen besteht noch Bedarf an interdisziplinärer Forschung, die theoretische Grundlagen, psychometrische Modelle und Messverfahren systematisch verknüpft und damit neue Perspektiven und eine neue Qualität für die Messung von Lernvoraussetzungen und Lernergebnissen schafft.

Das im Herbst 2007 eingerichtete DFG-Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (kurz: SPP „Kompetenzmodelle“) legt seinen Schwerpunkt auf die Modellierung und Messung von Kompetenzen (Klieme/Leutner 2006; Koeppen u.a. 2008). In der insgesamt sechsjährigen Laufzeit untersuchen Expertinnen und Experten aus der Erziehungswissenschaft, der Psychologie und den Fachdidaktiken grundlegende Fragen der Kompetenzmodellierung, der Entwicklung psychometrischer Modelle und der Nut-

1 Diese Veröffentlichung wurde ermöglicht durch Sachbeihilfen der Deutschen Forschungsgemeinschaft (KL 1057/9–1 und DL 645/11–1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

zung von Technologien zur Kompetenzmessung. Vertreten sind Wissenschaftlerinnen und Wissenschaftler aus über 20 deutschen Hochschulen sowie dem Max-Planck-Institut für Bildungsforschung (MPIB), dem Deutschen Institut für Internationale Pädagogische Forschung (DIPF), dem Leibniz Institut für die Pädagogik der Naturwissenschaften und der Mathematik (IPN), dem Institut zur Qualitätsentwicklung im Bildungswesen (IQB) und dem Institut für Schulqualität der Länder Berlin und Brandenburg e.V. (ISQ). Als internationale Expertinnen und Experten zur Beratung des SPP und als „critical friends" konnten Mark Wilson (University of California, USA), Joan Herman (CRESST, USA) und Benő Csapó (Szeged University, Ungarn) gewonnen werden. Das Kompetenzcluster „Technology-Based Assessment" (TBA) am DIPF bietet dem SPP Serviceleistungen bei der Entwicklung und Durchführung von computerbasierten Tests an.

Der für das SPP zentrale Begriff „Kompetenz" wird in heterogenen Zusammenhängen und mit sehr unterschiedlichen Bedeutungen verwendet. Für die Forschungsarbeiten im SPP wurde eine vergleichsweise enge Definition des Kompetenzbegriffs gewählt, um auf dieser gemeinsamen Basis theoretische Modelle und psychometrische Messverfahren entwickeln zu können. Kompetenzen sind im SPP definiert als „kontextspezifische kognitive Leistungsdispositionen, die sich funktional auf Situationen und Anforderungen in bestimmten Domänen beziehen" (Klieme/Leutner 2006, S. 879). Sie werden durch Erfahrung und Lernen erworben und können durch institutionalisierte Bildungsprozesse beeinflusst werden.

Im SPP wird das Ziel verfolgt, in mehreren Domänen (im Sinne von Inhaltsbereichen) theoretisch fundierte und für Diagnostik und Assessment nützliche Kompetenzmodelle zu entwickeln und empirisch zu überprüfen. Zurzeit werden im SPP mathematische, naturwissenschaftliche, sprachlich-kulturelle, berufsbezogene sowie fächerübergreifende Kompetenzen thematisiert. Im Kern steht jeweils die Entwicklung und empirische Prüfung theoretischer Kompetenzmodelle, aus denen psychometrische Messmodelle entwickelt werden. Daraus leiten sich Messverfahren zur empirischen Erfassung von Kompetenzen ab. Abschließend stellt sich die Frage, wie die Nutzung von Diagnostik und Assessment zu fundierten und präzisen Entscheidungen in der pädagogischen und bildungspolitischen Praxis beiträgt.

Nach Abschluss der ersten Zweijahresphase (2007–2009) wird in diesem Themenheft eine erste Zwischenbilanz zu den Forschungsarbeiten der insgesamt 23 Projekte des SPP vorgelegt. Das Heft ist wie folgt gegliedert: In einem einführenden Beitrag untersucht Benő Csapó grundlegende Fragen der Kompetenzmodellierung und der Kompetenzmessung, indem er Bildungsziele, Curricula und Assessment-Ansätze vor dem Hintergrund eines umfassenden Rahmenmodells zueinander in Beziehung setzt. Die darauf folgenden Beiträge aus den SPP-Projekten sind nach den fünf Kompetenzdomänen geordnet: fünf Beiträge mit Schwerpunkt „Mathematik", vier Beiträge mit Schwerpunkt „Naturwissenschaften", fünf Beiträge mit Schwerpunkt „Sprache und Lesen", sechs Beiträge mit Schwerpunkt „Fächerübergreifende Kompetenzen" und drei Beiträge mit Schwerpunkt „Lehrerkompetenzen". An jede Domäne schließt sich ein kurzer Beitrag von Kolleginnen und Kollegen an, die aufgrund ihrer besonderen Expertise im jeweili-

gen Forschungsgebiet eingeladen worden waren, die bisher erzielten Ergebnisse der Projekte einzuschätzen und zu kommentieren.

Literatur

Klieme, E./Leutner, D. (2006): Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. In: Zeitschrift für Pädagogik 52, S. 876–903.

Koeppen, K./Hartig, J./Klieme, E./Leutner, D. (2008): Current issues in competence modeling and assessment. In: Zeitschrift für Psychologie/Journal of Psychology 216, S. 61–73.

Anschrift der Autoren/der Autorin

Prof. Dr. Eckhard Klieme, Deutsches Institut für Internationale Pädagogische Forschung, Schloßstr. 29, D-60486 Frankfurt a.M.
E-Mail: klieme@dipf.de

Prof. Dr. Detlev Leutner, Universität Duisburg-Essen, Fakultät für Bildungswissenschaften, Lehrstuhl für Lehr-Lernpsychologie, Weststadttürme, Berliner Platz 6–8, D-45117 Essen
E-Mail: detlev.leutner@uni-duisburg-essen.de

Dipl.-Päd. Martina Kenk, Deutsches Institut für Internationale Pädagogische Forschung, Schloßstr. 29, D-60486 Frankfurt a.M.
E-Mail: kenk@dipf.de

Benő Csapó

Goals of Learning and the Organization of Knowledge

1. Introduction

Since the beginning of formal schooling, there has been a perennial search for *worthwhile knowledge*. Philosophers who have been posing similar questions on the issue mostly deal with objective knowledge,[1] whereas educators are interested in knowledge possessed by individuals. More specifically, educators' interest is in teaching and learning processes that result in worthwhile knowledge. One of the most recent candidates for this status is *competence*. Although there is no consensual understanding of the term, it has entered the discourse of policy documents. In this paper I outline a framework for interpreting the concept of competence. In so doing I offer a systematic way for comparing educational standards, curricula and assessment practices that will help us to better identify the goals of learning and design curricula.

For more than two millennia, there have been three main types of answers to the question „Why do children have to attend school?": (1) Transmitting knowledge accumulated by scientific inquiry has been a goal since at least Aristotle's time. (2) Cultivating children's developing minds emerged as a goal in ancient times as well, and since then has disappeared and re-emerged in the history of education. (3) Seneca's aphorism *Non scolae, sed vitae discimus* indicates that a social aspect, the external usefulness of knowledge mastered at school, has also been around for quite some time. Over the past centuries, attention has shifted between these three aspects of schooling, with one of them dominating from time to time. The pendulum seems to swing not only between the internal (focusing on children's abilities) and external (content of teaching) poles, but also along a triangle, set by the internal/psychological, content/disciplinary and social needs/application points.

In this paper I argue that these same three aspects identified in the course of the history of education still play a key role, and propose a framework that helps us to better identify goals of learning and contributes to more conscious curriculum design. Previous approaches were often dominated by one of these aspects. I argue that we have to keep all three of them in mind when setting standards, developing curricula and devising assessment frameworks.

Knowledge and *learning* are closely interlinked key concepts of educational science: The way children learn determines the type of resulting knowledge. In educational contexts, the two therefore cannot be conceptualized independently of each other. Conse-

1 Popper (1972) gave this title to his collection of essays, but subjective knowledge is no less interesting from a philosopher's perspective; see Polanyi (1958).

quently, revising these concepts should be perceived as parallel or rather integrated processes. Recent developments in society and economy, greater expectations concerning trained work force, expanding opportunities of learning and especially the accelerating speed of changes have prompted a continuous effort to define not only the knowledge needed by modern societies but also optimal learning and teaching processes. Such a reconceptualizing course of action is clearly indicated by the large number of recent publications on the subject. *Review of Research in Education* devoted its 2006 volume[2] to revisiting the concept of learning and its 2008 volume[3] to the concept of knowledge. The collection of essays edited by Benavot/Braslavsky (2006) examines new approaches to school knowledge and curriculum development from a broader social and global perspective that goes beyond the cognitive point of view.

One aspect of the new approach, as seen in the widespread use of the expression ‚forms of knowledge', indicates that a more differentiated view of knowledge as a product of learning is needed. We may assume that different goals require different methods of learning (and teaching) and that these processes result in different types of knowledge. In previous studies (Csapó 2004) I outlined a model representing various types of knowledge produced by schooling, as a function of pedagogical culture and of the methods of implementing curriculum contents. This model, based on the theoretical generalization of the findings of a series of empirical research projects (Csapó 2002), provided a framework to account for the differences in the quality of knowledge.[4]

In this paper I aim to show how a deeper understanding of learning and knowledge organization can contribute to designing curricula, preparing teaching materials and devising assessment standards that promote both students' development and their social needs more efficiently. I argue that three dimensions of the goals of learning have to be considered and that schooling cannot become more effective unless all three are viewed together. Each of the three dimensions can be targeted as a main goal in itself or can be seen as a prerequisite to or a means of achieving goals in the other two dimensions.

2. Sources of Educational Goals and the Dimensions of Learning

Most of the arguments regarding the goals of education fall under one of the following three approaches: (1) The scientific accumulation of knowledge is accelerating; therefore, an increasing amount of knowledge must be acquired at school. (2) Learning is about cultivating students' intellect and improving their abilities. (3) School must prepare its students for life and provide them with knowledge they can apply beyond school.

---

2 Rethinking Learning: What Counts as Learning and What Learning Counts.

3 What Counts as Knowledge in Educational Settings: Disciplinary Knowledge, Assessment and Curriculum.

4 For a summary of these projects, see Csapó 2007.

These three approaches are deeply rooted in European culture in general and European education in particular. When setting learning goals we have to consider three corresponding aspects or dimensions. First, there is the disciplinary or content dimension. An important source for setting goals is systematically organized external knowledge, accumulated and offered by the arts and scientific disciplines. Next, there is the social and cultural dimension, defined by the context for applying knowledge and by the expectations students need to know and be able to fulfill in order to become active and successful members of a given society. Finally, there is the internal, psychological dimension: how human intellect acquires, processes and applies knowledge, and how education should shape the related capabilities. These three aspects, however, point to three dimensions, as illustrated in Figure 1.

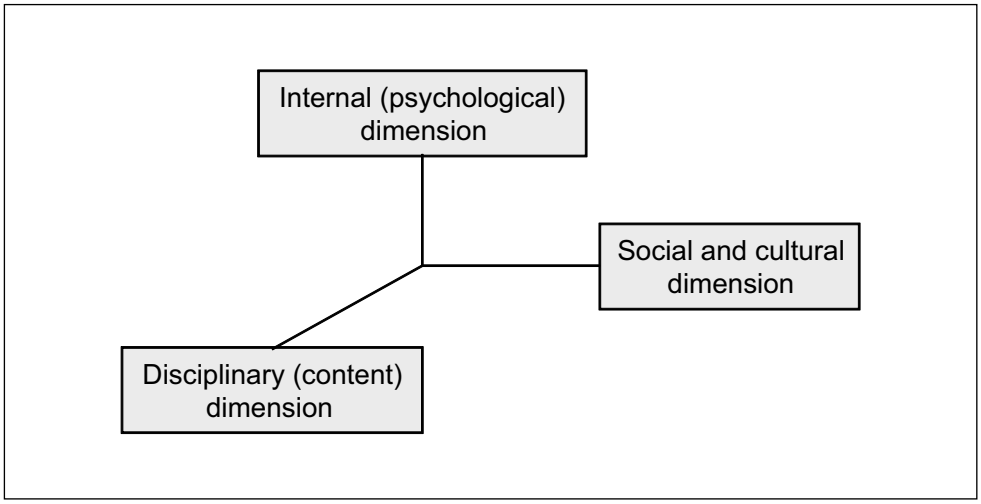


Fig. 1: Dimensions of the goals of learning

A similar three-dimensional model can be applied to knowledge acquired through learning. When inquiries are made about both the integrating principles that incorporate the individual elements of knowledge into an operational system and the justification of the presumption that the acquired knowledge will last, we arrive at the very same three-dimensional scheme (Csapó 2004). In a similar vein, when making efforts to find the sources for setting and accomplishing goals, we also find three categories (i.e., knowledge accumulated by the arts and sciences, results of psychological and methodological research and social needs and expectations). Table 1 provides a summary of this scheme.

Assessment frameworks may also assume three approaches. Discipline-dominated assessments measure knowledge in the same (disciplinary) context in which they are mastered. Assessments focusing on psychological attributes such as general abilities, problem solving or creativity may be free of any specific content. Finally, assessments evaluating how knowledge is applied in a social context have to deal with transferable knowledge and competencies.

Characteristics	Disciplinary, content-based	Internal, psychological	Social, cultural, application
Goals of learning	Acquisition of canonized content (objective, scientific knowledge)	Development of cognitive functions and intellectual abilities	Acquisition of sociocultural codes and modes of behavior and action, preparing the individual for integration into society
Emerging knowledge	Expertise, domain specific skills	Thinking skills, improved general abilities	Literacy, flexible and expandable knowledge applicable in a broad range of contexts
Sources for designing standards, curricula, textbooks, learning materials	A systematic body of knowledge of the arts and sciences	Results of psychological and educational research	Analysis of social needs and contexts of knowledge and skills application
Assessment	Same context as learning	Focus on structures; content plays a secondary role	Transfer from school to everyday context

Tab. 1: Dimensions and structure of knowledge

When examining the individual columns of the table and the resources available for improving teaching and learning in the particular dimension, large discrepancies are found. The following sections examine the most important features of these three dimensions.

3. Disciplinary, Content-Based Dimension of Learning

In early times, schooling focused on the acquisition of knowledge in philosophy, the humanities, and the arts. Then, spectacular developments in the natural sciences demanded their curricular inclusion, leading to the humanities/science dichotomy. In the last century, procedures for the systematic planning of school curriculum contents were also established, and it is no coincidence that the systematic, canonized body of knowledge accumulated by the sciences (and, to a lesser extent, by the arts) became its primary source.

The disciplinary approach has affected the methodology of teaching and learning substantially. Knowledge is rooted in science, with teachers and textbooks as transmitters. Ac-

cording to the simplified approach that neglects the other two dimensions, the teacher ‚delivers' the teaching material, which the student learns. The role model for the students is a researcher or teacher of the given discipline. The school transmits the mathematics of mathematicians, the physics of physicists and the history of historians. Developing an understanding of the relationships offered by the content occurs in the specific context of the given school subject. Gardner (1991) calls this kind of comprehension *disciplinary understanding*. Concepts are anchored in results of scientific research and are shaped by scientific definitions. Learning contents is organized in the way the given discipline structures its knowledge, and the process of teaching follows this logical order. Dealing with formulas for mathematics, physics and chemistry; memorizing in the simplest case; or transforming and linking the formulas take place by mastering a body of disciplinary knowledge.

Some school systems, including many European and Asian ones, have achieved remarkable results in transmitting disciplinary knowledge. Schools of this kind nurture ‚little scientists', who can turn into great researchers when they grow up. This approach seems acceptable for those few who strive to continue their studies at universities in the given discipline and later become experts in the field, earning their living as such. However, without investing effort into developing general intellectual skills (for which science as a learning content offers excellent opportunities), the disciplinary approach in itself is not enough to educate inventive, creative scientists. At the same time, those who do not want to pursue a career related to that particular discipline in research, development or education will hardly benefit from discipline-oriented learning. Research on conceptual change and science misconception has also shown that students' scientific knowledge is often isolated from everyday life and that students tend to apply their naive models generalized from personal experiences rather than their school-created scientific knowledge to interpret phenomena.

Traditional discipline-oriented teaching methodology focuses on transmitting content defined as valuable by the scientific community. This viewpoint is further reinforced in many teacher training systems by allocating instructors' job statuses who teach discipline specific teaching methods (*Fachdidaktik* in Germany) at disciplinary departments. Influential academic communities in this area have been formed, with strong professional associations and journals. Discipline-related teaching methodology journals (especially those of the natural sciences) adopted the norms of scientific publications at an early stage and compiled a considerable amount of scientifically established knowledge on the teaching of the particular disciplines at school.[5]

The disciplinary approach to learning is in a very strong position, having at least a half-century advantage over the other two dimensions in terms of its traditions and infrastructure. Its position is further strengthened by the fact that nearly the entire community of academics identifies with the very same approach and uses it when educating their successors: academics, specialists, experts, or the gifted in general.

5 The fact that the Web of Knowledge (formerly Thomson Scientific) includes these types of journals in the *Science Citation Index* (and not in the *Social Science Citation Index*) also seems to support this statement.

Several trends in psychology and education have contributed to strengthening this approach, the most prominent being the early phase of cognitive psychology,[6] which regarded genuine knowledge primarily as expertise.[7] The development of expertise is studied by comparing the novice and the expert, and progress is defined in terms of the number and differentiation of schemata used in specific contexts. Fully developed expertise comprises thousands of specific schemata, which, once learned, can be used effectively. However, this entails learning a huge amount of facts and data and mastering schemata applicable in the appropriate contexts. Such knowledge is generally reproductive and used under circumstances similar to its acquisition. Both the expert and expertise are defined by the subject, without allowing for transferring knowledge to novel or distant areas. Here, problem-solving is seen as the application of knowledge to (relatively) new situations. In this model, experts are engaged in much less thinking than is usually assumed: they know the answer practically off-hand. If they do think, it is not computation-like logical operation. Rather, it is a search among familiar schemata, the matching of a ready-made solution with the situation.

Despite all its shortcomings, disciplinary learning has yielded much that is valuable and should be preserved. However, it needs to be revisited from time to time (see Ford/Forman 2006; Duschl 2008). The discipline-based approach has little to say about how learners actually reason. Although the study of knowledge as expertise assumes that experts reason when they process information, other paradigms have developed more sophisticated models of how reasoning takes place.

4. Internal, Psychological Dimensions of Learning

References to psychological considerations preceded the establishment of the science of psychology. One of Greek philosophy's major missions was to cultivate the intellect. The virtues or wisdom mentioned by Aristotle do not imply the acquisition of an external entity but rather the development of an internal quality.

No sooner was formal education born than the need to develop thinking, generally meant as logical thinking, was manifested. For a long time, it was thought that it could be fostered by learning mathematics and the grammatical structures of languages. The assumption behind the endless practice of certain grammatical and logical puzzles was that they made students smarter, but without a clear vision of how schooled minds differed from unschooled ones, these efforts produced little success.

As soon as scientific tools for studying the human intellect came into being and psychometrics presented techniques for measuring intelligence, the urge to develop the in-

6 As a prototype of the works on this issue from the early stage of cognitive psychology, see Simon's 1979 study. For later conceptualizations, see Ericsson/Smith 1991.

7 The first and major part of the book ‚How People Learn' – and its extension to mathematics, sciences and history – provides a good description of this approach (Bransford/Brown/Cocking 2000).

tellect based on this new scientific approach accordingly emerged. The question arose to what extent intelligence or any of its components can be learned and taught. Factor analytic studies provided the basis for models of the structure of human intellect and identified the most important intellectual abilities. However, the concept of intelligence became the subject of ideological and political debates; as a result, it fell into disrepute for some time.

Nevertheless, several experiments attempted to improve thinking skills, general cognitive abilities[8] or even intelligence, although most of them adopted the so-called direct approach and yielded controversial results (see Blagg 1991). Failure may be partly due to the fact that intelligence is a complex construct and its measurable manifestations and effective functioning imply the combination of a number of specific abilities in a concerted and coordinated effort. Moreover, the notion of intelligence – particularly in association with hereditariness – became discredited in the public eye, thwarting informed discussions of it in the context of school education.

Success is more probable in the case of abilities whose structures readily lend themselves to study and description, thereby simplifying the identification of appropriate developmental tasks. However, such endeavors rarely transcended a few experiments of limited scope, with two factors having prevented the expected improvements from becoming fully fledged. On the one hand, no development is possible without some content, and neglecting the curricular, disciplinary content proved to be a dead end. On the other hand, the abilities that these projects aimed to develop are much more difficult to identify and are less understood than the widely known disciplinary contents or the knowledge gained through learning them.

The development of general abilities resulting from learning is more difficult to observe, and the process is more difficult to monitor. Therefore, approaches[9] that use restructured curricular materials to improve thinking processes that can be more easily identified are more successful and report more lasting effects. For example, Piaget offered a framework for describing the developing mind; furthermore, mathematics and science (and some other school subjects) provide well-structured (or restructurable) materials to practice reasoning skills (Adey 1999; Shayer/Adey 2002).

Mathematics and reading enjoy a special status among school subjects, given that learning them is so deeply embedded in the psychological apparatus of humans. Therefore, applying methods in teaching mathematics and reading that are based on the results of psychological research are the best tools for facilitating the development of students' minds (Nunes/Bryant 1996, 2009).

In general, there is a clear shift in cognitive training from direct methods using abstract materials towards embedded methods (see Csapó 1999) that use the content of

8 Costa (1991) presents a large number of programs from the U.S.A. aimed at teaching thinking, most of which assume the direct approach.

9 Such programs are discussed in the books edited by Hamers/Overtoom (1997) and Hamers/van Luit/Csapó (1999).

teaching to stimulate intellectual development.[10] This approach is more compatible with existing schooling practices, because it considers the goals mentioned in the previous section acceptable and contents of materials offered by the disciplines as more or less given. However, there are large differences between the two approaches. The first one regards transmitting disciplinary materials as a primary source for planning instruction and aims to produce expertise based on this knowledge before seeking psychological theories and scientific evidence that support this goal. The approach presented in this section considers development of human capabilities as a primary goal and looks for disciplinary content and methods of teaching that best serve this end (see also Kuhn 2005).

A new category of scientific knowledge on the psychological dimension of learning has proliferated in the last few decades. One of the most dynamic fields of modern sciences is brain research or cognitive neuroscience in general, which studies the biological apparatus of information processing. The heightened interest in this field gave momentum to several international projects and syntheses (OECD 2007; Geake 2004; Goswami 2004; Stern et al. 2005). Although cognitive neuroscience obviously cannot accomplish the universal task of laying the scientific groundwork for learning (Bruer 1997), its advances have had a direct impact on recent developments in formal education in several areas. The *How people learn* framework (Bransford/Brown/Cocking 2000) also considers brain research as a founding discipline for education in general; however, results of cognitive neuroscience in its present state are most helpful in learning settings that are characterized by the rapid development of the nervous system (in pre-school and early school years) or when development significantly diverges from the average. These findings can also be highly relevant in cases where the knowledge to be acquired is more closely tied to the biological apparatus or is determined by the other two (the disciplinary and the cultural) dimensions to a lesser degree: early reading and mathematics are good examples here.

The findings of brain research brought intelligence and the issue of general abilities to the foreground once again. If it is true that the human brain is plastic and can be transformed by appropriate stimuli and learning, then education cannot afford to ignore the implications. Therefore, opportunities of learning that develop plastic general abilities have to be assigned a more central role (Adey et al. 2007).

5. Social Needs and Application-Oriented Dimension of Learning Goals

The third main aspect of learning is the application dimension: Students are expected to acquire knowledge that is socially valid, which helps them to be successful in their private and professional life. Traditionally, schooling was expected to fulfill these aims.

10 This shift is clearly demonstrated by Klauer's work on training inductive reasoning. This model of inductive reasoning, first implemented in the form of three sets of training instruments using abstract materials, later served as a theoretical framework for several content-based training experiments in a number of school subjects (for an overview, see e.g., Klauer 2001; Klauer/Phye 2008).

Today, the fact that a great deal of learning takes place outside of the school walls makes it tempting to challenge this notion. However, learning that occurs outside the school usually takes place in the same context in which the knowledge needs to be utilized; consequently, application of the result of this kind of learning is natural. The burden is hence on formal education and educational researchers to find ways of teaching and learning when the context of the future application of the outcomes of learning is increasingly unknown.

Challenges and unsolved problems are most apparent in this dimension. Although it is obvious that schooling has to prepare students for life, there is little scientific knowledge of how this preparation should be best done. There are established methods to map disciplinary knowledge onto school curricula, and there are the experts as models of successful learners – models whom students may be expected to follow when aiming to master disciplinary knowledge. A growing body of psychological knowledge supports refining the goals of improving general abilities. However, there are no generally accepted scientific methods to identify social needs and expectations concerning useful, valid and applicable knowledge (Duschl 2008).

Educational systems face growing pressure to prepare students for life, but curriculum developers and assessment specialists find little research that indicates how this can be done. Several research paradigms did, however, examine the relationships between traditional schooling that focuses on subject matter knowledge and the requirements of the outside world, – in particular the discrepancy between learning that takes place at school and outside of it and between knowledge mastered at school and knowledge useful in life. The inconsistencies became most apparent in mathematics: Students were hardly able to utilize the de-contextualized, abstract knowledge in realistic contexts. The comparison of school mathematics and ‚street mathematics' revealed that transfer is not automatic in the other direction either: Children successful in practical numerical operations may fail at school (Nunes/Carraher/Schliemann 1993). Several approaches tried to bridge the gap; these include *realistic mathematical modeling*[11] and re-conceptualization of the role of real-world problems in mathematics education (Verschaffel/Greer/De Corte 2000). In some disciplines, economic pressure accelerates the identification of such skills and knowledge. The profound change in the curricula for foreign languages from grammatical and cultural studies to communication was a result of pressure from stakeholders.

Teaching abstract science contents in some modern areas of physics and chemistry has generated similar problems; in reaction it was often proposed that students be taught something ‚practical', meaning directly applicable in real life. A broad range of such practice-oriented approaches have appeared in the past decades, from ‚home-science', ‚kitchen science', and ‚hands-on science' to complex projects and the application of principles of problem-based learning. Such methods may have great motivating power and help form students' attitudes towards learning science. They are also great tools for integrating and structuring students' knowledge. But if they abandon the principles of

11 This approach is most prominently represented by the work of the Freudenthal Institute.

scientific reasoning and the resulting knowledge is bound to a narrow context lacking transferability, they are just as inert and ineffective as rote learning.

Exploring the ways in which schools can prepare their students for meeting the expectations of society and the economic environment has also become a central issue in contemporary large-scale assessment projects. The most influential analysis of this kind is taking place within the framework of the OECD PISA surveys. PISA broke with the practice of disciplinary, curriculum-based assessment and relies on the knowledge needs of modern society when defining the themes of its assessments. The theoretical framework for the surveys (OECD 2000, 2003, 2006) describes the body of knowledge fifteen-year-olds need in modern societies in order to be able to participate in social processes, to create a balanced way of living as well as to develop themselves. When this new concept of knowledge was outlined, literacy served as a point of departure. The earlier role played by literacy in the narrow sense of the word (i.e., reading and writing) was replaced by a body of broadly based knowledge applicable in various situations. The broadening of the term *literacy* generated concepts such as reading literacy, scientific literacy and mathematical literacy. In our interpretation, the literacy concept of the PISA frameworks points to this third dimension, and in this way, measures an important aspect of students‘ knowledge that had not received enough attention before.[12]

The findings of surveys[13] show that solving practical tasks different from the ones that are a given at school presents considerable difficulties for students, even if they possess the necessary skills. Studies have revealed that the transfer of knowledge does not come automatically and that further learning and development are necessary to facilitate the application and transfer of acquired knowledge to new contexts (Bransford/Schwartz 1999).

Obviously, one of the principal goals of schooling is to create knowledge applicable to practical real-life situations. In theory, there seem to be two paths to achieving this. One is to introduce radical changes in the content of education: Disciplinary knowledge has to be superseded by practical knowledge that is directly applicable in everyday life. Simple as it may seem, it is easy to see that doing so would not engender the desired results. First, the environment in which acquired knowledge is to be employed can be unpredictable and may change profoundly several times during the lifespan of the user of this knowledge. Second, social needs regarding applicable knowledge change very rapidly. Third, ‚common‘ everyday applied knowledge does not lend itself as readily as a scientific body of knowledge to being organized into a clear-cut system or into basic principles that are generally valid. This approach alone leads nowhere.

12 A similar, albeit less explicit three-dimensional thinking can be identified in the way Klieme/Hartig/Rauch (2008) introduce the essence of the PISA approach. „They neither restricted educational assessment to knowledge and skills within a few school subjects, nor referred to psychological theories. Instead, they took a functional view, asking whether young adults are prepared to cope with the demands and challenges of their future life“ (p. 8).

13 In addition to PISA, several in-depth Hungarian research programs have highlighted the difficulties of knowledge application in realistic contexts (Csapó 1998, 2002).

The other road to take is a more effective way of imparting scientific knowledge and reinforcing mechanisms that foster understanding and thus transfer more efficiently. It leads us to a comprehensive approach to the three dimensions of learning, that is, to the integration of the development of thinking and abilities and instruction in curriculum contents in order to create knowledge that is more deeply understood and can be more extensively applied. PISA has underpinned this trend by the inclusion of problem solving in addition to the three core domains in its 2003 survey (OECD 2004).

The OECD PISA project reaches beyond establishing a new concept of knowledge. It also sheds new light on learning itself. The first analysis of this kind, incorporated into the 2000 survey, formulated the question whether the learning methods that students in the participating countries adopt to prepare for the ‚real world' can meet the expectations raised by the modern age. Do they process through active reasoning what they learn and strive to understand it or do they aim only for rote learning? Have they acquired self-regulated learning, which enables them to organize their own learning processes effectively and to become high-achieving learners once they no longer have the external control of the school to rely on (Artelt et al. 2003)? Results showed significant differences between learners from the participating countries.[14]

6. Efforts Aiming to Connect Multiple Goals and the Concept of Competence

Although all three dimensions discussed in the previous sections can be traced back to ancient times as goals of education, combining them is a relatively new phenomenon. Expectations concerning education in the 21st century may be so radically different from those of the previous centuries that they call for an even closer integration of these dimensions. Two aspects underpin such a need: (1) Knowledge has never played such a decisive role in the lives of such a great proportion of people. (2) The pace of changes to the social and economic environment may be faster than the developmental changes in an individual's life; therefore, knowledge necessary throughout the lifespan has never been so difficult to foresee.

Several theoretical frameworks have been proposed and a number of empirical studies have been proving that learning that includes deep reasoning may be the best tool both for developing students' minds and for constructing and retaining a well-organized body of content knowledge. Complex methods integrating an increasing number of functions employ well-structured contents of learning in order to develop skills and abilities. Research and development projects and experiments have shown that well-structured instruction enriched with relevant practice fosters not only the acquisition of the subject matter but also develops intellectual abilities effectively. Ausubel (2000) proposes active, *meaningful* learning; others (e.g., Darling-Hammond et al. 2008) focus

14 Hungarian students ranked relatively low. The predominance of rote learning has been confirmed by a similar survey on several age groups (Németh/Habók 2006).

on understanding or even multiple understanding.[15] These approaches integrate two dimensions (content and psychological) of learning. I suggest going on to integrate the third dimension as well and contend that meaningful learning – that is, learning with understanding – is also the best way to increase the applicability and transferability of knowledge.

Thus, I have arrived at the concept of competence; recently, it has been one of the most frequently used and the most controversial constructs at the same time. Extensive theoretical conceptualizing efforts (e.g., Rychen/Salganik 2001, 2003; Hartig/Klieme 2007; Koeppen et al. 2008), political documents[16] and large-scale empirical projects (Klieme/Leutner 2006; Hartig/Klieme/Leutner 2008) have been using this concept.

Seeing this recent vast interest in competence, we may ask the question whether competence is a new, recently discovered psychological phenomenon or if it is the quintessence of „good knowledge", the form of knowledge that educators and educational researchers have been looking for. As several studies have pointed out (e.g., Weinert 2001; Klieme/Hartig/Rauch 2008), the term *competence* has been used in a great number of senses, quite often as a synonym for several other terms. Not surprisingly, interpretations may be easily found that point to one of the dimensions described earlier. For example, Chomsky's (1968) original concept of competence emphasizes its innate character, fitting into the psychological dimension. In the PISA terminology, *competence* and *literacy* are often used interchangeably, indicating that in the PISA interpretation, competence points to the application dimension as identical with applicable, socially valid and valuable knowledge. Other interpretations (see examples in Weinert 2001) relate competence to specific skills and knowledge of a profession, using competence as a synonym for expertise or expert knowledge (*Fachkompetenz*).

In educational contexts, competencies are often defined as complex ability constructs contextualized and usable in relevant situations (Klieme/Hartig 2007; Klieme/Hartig/Rauch 2008). In this approach, each dimension described earlier is present. Therefore, in the framework presented in this paper, I suggest considering competencies not as identical with one of the dimensions, but as a harmonious composition of all three.

Accepting the interpretation that competencies are complex constructs and regarding their development as the ultimate aim of instruction does not mean that each particular educational process always has to deal with competencies. Recent proliferation of the term may imply an interpretation that no other constructs play an important role in learning. I do not share such a view but propose a more differentiated approach, where several combinations of the described dimensions result in the desired outcomes. For example, in early childhood the psychological dimension may dominate: Education should stimulate the developing mind and this aspect should determine the selection of

15 Gardner (1991) distinguishes several kinds of understandings, while Bereiter (2002) elaborates an even broader range of the forms of understanding.

16 See the eight domains of key competencies defined by the European Reference Framework (European Commission 2004).

the content of learning. Later, especially when preparing for a profession, learning may be directed by the structure of knowledge organized by the logic of a given discipline or trade. Application plays a significant role in both cases, aiming for a broader transfer in the first case and a narrower one in the latter.

Authentic summative assessment cannot take place without considering competencies in their natural complexity. It cannot happen without (1) the application of knowledge in new contexts that require (2) highly developed general information processing skills and thinking abilities, and, of course, (3) well-structured disciplinary knowledge that is supposed to be applied. PISA took this global approach when introducing the concept of literacy and has been focusing on its assessment ever since. However, there are still unrealized potentials in this approach, given that in modern societies, knowledge learned earlier is often applied in other learning situations. For example, application of mathematics knowledge may happen in science. Therefore, not only real-life, meaning ‚everyday' situations can be considered authentic.

Formative or diagnostic assessment may require a different approach, focusing on one of the dimensions separately from the others. Just as a diagnosis in the medical practice assumes knowledge of the anatomy of the diagnosed body, diagnostic assessment assumes knowledge of the structure of the assessed competencies. Diagnostic assessments and student-level monitoring systems may focus on one of the dimensions in order to identify specific developmental abnormalities.

Competencies are considered dominantly cognitive constructs. At the same time, it has become increasingly clear that the processes determining the efficiency of formal learning cannot be understood by paying attention to cognitive factors alone, without considering the social environment where learning occurs and the non-cognitive psychological dimension. A few decades after the cognitive revolution, relying heavily on its advances and new research methods, research into non-cognitive factors took off, so much so that nowadays we are witnesses to an affective and sociocultural revolution. A deeper understanding of motivation, self-concept and attitudes to learning, various subjects and curriculum content has effected changes in instructional practice and contributed to improving pedagogical culture. Clarifying the interactions between competencies and the affective domain offers further potential for psychological and educational research.

7. Acknowledgements

This paper is based on work carried out in the Research Group on the Development of Competencies, Hungarian Academy of Sciences. Some of the issues discussed here were presented at the opening conference of the Priority Programme „Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes" and have been published in Hungarian. I am grateful to Paul Andrews, Jens Fleischer, Andrea Kárpáti, Martina Kenk, and Marianne Nikolov for their comments on an earlier draft of this paper.

References

Adey, P. (1999): Thinking science: Science as a gateway to general thinking ability. In: Hamers, J.H.M./van Luit, J.E.H./Csapó, B. (Eds.): Teaching and learning of thinking skills. Lisse: Swets and Zeitlinger, pp. 63–80.

Adey, P./Csapó, B./Demetriou, A./Hautamaki, J./Shayer, M. (2007): Can we be intelligent about intelligence? Why education needs the concept of plastic general ability. In: Educational Research Review 2, No. 2, pp. 75–97.

Artelt, C./Baumert, J./Julius-McElvany, N./Peschar, J. (2003): Learners for life. Students approaches to learning. Results from PISA 2000. Paris: OECD.

Ausubel, D.P. (2000): The Acquisition and Retention of Knowledge: A Cognitive View. Dordrecht: Springer.

Benavot, A./Braslavsky, C. (2006): School knowledge in comparative and historical perspective. Changing curricula in primary and secondary education. New York: Springer.

Bereiter, C. (2002): Education and mind in the knowledge age. London: Lawrence Erlbaum Associates.

Blagg, N. (1991): Can we teach intelligence? A comprehensive evaluation of Feuerstein's Instrumental Enrichment Program. Hillsdale, NJ.: Lawrence Erlbaum Associates.

Bransford, J.D./Brown, A.L./Cocking, R.R. (2000): How people learn. Brain, mind, experience and school. Washington, DC: National Academic Press.

Bransford, J.D./Schwartz, D.L. (1999): Rethinking transfer: A simple proposal with multiple implications. In: Review of Research in Education 24, pp. 61–100.

Bruer, J.T. (1997): Education and the brain: A bridge too far. In: Educational Researcher 26, No. 8, pp. 4–16.

Chomsky, N. (1968): Language and mind. New York: Hartcourt, Brace & World Inc.

Costa, A.L. (Ed.) (1991): Developing minds. Programs for teaching thinking. Alexandria, VA: Association for Supervision and Curriculum Development.

Csapó, B. (Ed.) (1998): Az iskolai tudás [School knowledge]. Budapest: Osiris Kiadó.

Csapó, B. (1999): Improving thinking through the content of teaching. In: Hamers, J.H.M./van Luit, J.E.H./Csapó, B. (Eds.): Teaching and learning thinking skills. Lisse: Swets and Zeitlinger, pp. 37–62.

Csapó, B. (Ed.) (2002): Az iskolai műveltség [School literacy]. Budapest: Osiris Kiadó.

Csapó, B. (2004): Knowledge and competencies. In: Letschert, J. (Ed.): The integrated person – How curriculum development relates to new competencies. Enschede: CIDREE, pp. 35–49.

Csapó, B. (2007): Research into learning to learn through the assessment of quality and organization of learning outcomes. In: The Curriculum Journal 18, No. 2, pp. 195–210.

Darling-Hammond, L./Pearson, P.D./Schoenfeld, A.H./Stage, E.K./Zimmerman, T.D./Cervetti, G.N./Tilson, J.L. (2008): Powerful learning. What we know about teaching for understanding. San Francisco: Jossey-Bass.

Duschl, R. (2008): Science education in three parts harmony: Balancing conceptual, epistemic and social learning goals. In: Kelly, G.J./Luke, A./Green, J.G. (Eds.): What counts as knowledge in educational settings: Disciplinary knowledge, assessment and curriculum. In: Review of Research in Education 32, pp. 268–291.

Ericsson, K.A./Smith, J. (Eds.) (1991): Toward a general theory of expertise. Prospects and limits. Cambridge: Cambridge University Press.

European Commission (2004). Implementation of „Education and training 2010" Work Programme. Working Group B. „Key competences". Key competences for Lifelong Learning. A European Reference Framework. http://ec.europa.eu/education/policies/2010/doc/basicframe.pdf [May 20, 2009].

Ford, M.J./Forman, E. (2006): Redefining disciplinary learning in classroom contexts. In: Green, J./Luke, A. (Eds.): Rethinking learning: What counts as learning and what learning counts. In: Review of Research in Education 30, pp. 1–32.

Gardner, H. (1991): The unschooled mind. How children think and how schools should teach. New York: Basic Books.

Geake, J. (2004): Cognitive neuroscience and education: two-way traffic or one-way street? Westminster Studies in Education 1, pp. 87–98.

Goswami, U. (2004): Neuroscience and education. British Journal of Educational Psychology 1, pp. 1–14.

Hamers, J.H.M./Overtoom, M.Th. (Eds.) (1997): Teaching thinking in Europe. Inventory of European Programmes. Utrecht: Sardes.

Hamers, J.H.M./van Luit, J.E.H./Csapó, B. (Eds.) (1999): Teaching and learning thinking skills. Lisse: Swets and Zeitlinger.

Hartig, J./Klieme, E. (Eds.) (2007): Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Bonn/Berlin: Bundesministerium für Bildung und Forschung.

Hartig, J./Klieme, E./Leutner, D. (Eds.) (2008): Assessment of competencies in educational contexts. Göttingen: Hogrefe & Huber.

Klauer, K.J. (2001): Training des induktiven Denkens. In: Klauer, K.J. (Ed.): Handbuch Kognitives Training. Göttingen: Hogrefe & Huber, pp. 165–209.

Klauer, K.J./Phye, G.D. (2008): Inductive reasoning: A training approach. Review of Educational Research 78, No. 1, pp. 85–123.

Klieme, E./Hartig, J. (2007): Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. In: Prenzel, M./Gogolin, I./Krüger, H.-H. (Eds.): Kompetenzdiagnostik. Zeitschrift für Erziehungswissenschaft, Sonderheft 8, pp. 11–29.

Klieme E./Hartig, J./Rauch, D. (2008): The concept of competence in educational context. In: Hartig, J./Klieme, E./Rauch, D. (Eds.): Assessment of competencies in educational context. Göttingen: Hogrefe, pp. 3–22.

Klieme, E./Leutner, D. (2006): Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen: Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. Zeitschrift für Pädagogik 52, pp. 876–899.

Koeppen, K./Hartig, J./Klieme, E./Leutner, D. (2008): Current issues in competence modeling and assessment. Zeitschrift für Psychologie/Journal of Psychology 216, No. 2, pp. 61–73.

Kuhn, D. (2005): Education for thinking. Cambridge, MA: Harvard University Press.

Németh, B.M./Habók, A. (2006): A 13 és 17 éves tanulók viszonya a tanuláshoz [13–17-years-olds' attitudes to learning]. In: Magyar Pedagógia 106, No. 2, pp. 83–105.

Nunes, T./Bryant, P. (1996): Children doing mathematics. Oxford: Blackwell Publishers.

Nunes, T./Bryant, P. (2009): Children's reading and spelling. Beyond the first steps. Oxford: Blackwell Publishers.

Nunes, T./Carraher, D.V./Schliemann, A.D. (1993): Street mathematics and school mathematics. Cambridge: Cambridge University Press.

OECD (2000): Measuring student knowledge and skills. The PISA 2000 assessment of reading, mathematical and scientific literacy. Paris: OECD.

OECD (2003): The PISA 20003 assessment framework. Paris: OECD.

OECD (2004): Problem solving for tomorrow's world. First measures of cross-curricular competencies from PISA 2003. Paris: OECD.

OECD (2006): Assessing scientific, reading and mathematical literacy. A framework for PISA 2006. Paris: OECD.

OECD (2007): Understanding the brain. The birth of a learning science. Paris: OECD.

Polanyi, M. (1958): Personal knowledge: Towards a post-critical philosophy. Chicago: University of Chicago Press.

Popper, K.R. (1972): Objective knowledge. An evolutionary approach. Oxford: Clarendon Press.

Rychen, D.S./Salganik, L.H. (Eds.) (2001): Defining and selecting key competencies. Seattle: Hogrefe & Huber.
Rychen, D.S./Salganik, L.H. (Eds.) (2003): Key competencies for a successful life and a well-functioning society. Seattle: Hogrefe & Huber.
Shayer, M./Adey, P. (Eds.) (2002): Learning intelligence. Cognitive acceleration across the curriculum. Buckingham: Open University Press.
Simon, H.A. (1979): Information processing models of cognition. American Review of Psychology 30, pp. 363–369.
Stern, E./Grabner, R./Schumacher, R./Neuper, C./Saalbach, H. (2005): Educational research and neurosciences – expectations, evidence and research prospects. Bonn/Berlin: Bundesministerium für Bildung und Forschung.
Verschaffel, L./Greer, B./De Corte, E. (2000): Making sense of word problems. Lisse: Swets & Zeitlinger.
Weinert, F.E. (2001): Concept of competence: A conceptual clarification. In: Rychen, D.S./Salganik, L.H. (Eds.): Defining and selecting key competencies. Seattle: Hogrefe & Huber, pp. 45–65.

Anschrift des Autors

Benő Csapó, University of Szeged, Institute of Education, HU-6722 Szeged,
Petőfi sgt. 30–34, Hungary

Marianne Bayrhuber/Timo Leuders/Regina Bruder/Markus Wirtz

Repräsentationswechsel beim Umgang mit Funktionen – Identifikation von Kompetenzprofilen auf der Basis eines Kompetenzstrukturmodells

Projekt HEUREKO[1]

1. Theoretischer Hintergrund

Ziel des Projektes HEUREKO ist die Konstruktion und Überprüfung eines Kompetenzstrukturmodells in dem für den Mathematikunterricht zentralen Bereich der Leitidee „Wachstum und Veränderung". Auf der Basis dieses Modells sollen typische Kompetenzprofile von Schülerinnen und Schülern beim Arbeiten mit mathematischen Funktionen und ihren Repräsentationsformen identifiziert werden.

Mathematische Repräsentationsformen sowie der Wechsel zwischen ihnen spielen in der mathematikdidaktischen Forschung eine zentrale Rolle (vgl. z.B. Swan 1985; Goldin 1998; Ainsworth/Bibby/Wood 2002). Im mathematischen Inhaltsbereich „Wachstum und Veränderung" sind Tabelle, Graf, Term und Situation (verbale Beschreibung oder bildliche Darstellung eines Sachverhalts) als typische externe Repräsentationsformen von Funktionen zentrale Konzepte (vgl. Malle 2000). Trotz der hohen Relevanz innerhalb der Mathematikdidaktik, steht eine systematische empirische Untersuchung des Wechsels zwischen „Tabelle" bzw. „Term" und anderen externen Repräsentationsformen von Funktionen noch weitgehend aus.

Empirische Untersuchungen zum Multimedia-Lernen zeigen, dass unterschiedliche Repräsentationsformen mathematischer Zusammenhänge und der Wechsel zwischen diesen entscheidend für individuelle Lernprozesse sind. Damit sich ein positiver Lerneffekt einstellt, müssen die Schülerinnen und Schüler in der Lage sein, zwischen den vorliegenden Repräsentationen zu wechseln, die verbindenden Elemente zu identifizieren und zu verknüpfen (vgl. Ainsworth/Bibby/Wood 2002; Seufert 2003). Während des Erlernens dieser Grundfertigkeiten zeigen sich typische Fehler (vgl. Bodemer u.a. 2004).

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennzeichen: FR 2552/2-1, FR 2552/2-2 und WI 3210/2-1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen" (SPP 1293).

Kozma und Russell (1997) berichten, dass Novizinnen und Novizen beispielsweise verschiedene Repräsentationen nur über deren Oberflächenmerkmale aufeinander beziehen, Expertinnen und Experten hingegen stellen Bezüge über dahinterliegende Konzepte her.

Eine Möglichkeit der Beschreibung der kognitiven Prozesse bei der Entnahme von Informationen aus Texten und Grafen wird im Modell zur integrativen Bild- und Textverarbeitung (vgl. Schnotz 2005) beschrieben. Hier werden drei hierarchische Verarbeitungsstufen der Bild-Text-Integration unterschieden: (a) Es werden Beziehungen zwischen Bild- und Textelementen hergestellt, Detailinformationen aus Bild und Text werden miteinander verknüpft. (b) Verknüpfung von einfachen semantischen Relationen zwischen Text und Bild: diese werden auf die jeweils andere Repräsentationsform bezogen. (c) Das Individuum verknüpft komplexe semantische Relationsgefüge zwischen Text und Bild. Solche Kompetenzen von Schülerinnen und Schülern hinsichtlich der Kombination von Graf bzw. Tabelle und Text sind z.B. für das Problemlösen mit Funktionen im Mathematikunterricht von Bedeutung. Unterschiedliche Schülerprofile beim Umgang mit Funktionen können auf verschiedene Verarbeitungstypen und -niveaus bei der Integration von Bild und Text zurückgeführt werden.

Eine Studie von Pesonen, Ehmke und Haapasalo (2005) hat gezeigt, dass beim Umgang mit verschiedenen Repräsentationsformen unterschiedliche Typen von Lernenden vorliegen. Anhand einer latenten Klassenanalyse ließen sich drei qualitativ verschiedene Verständnisstufen vom mathematischen Begriff der binären Operation (Rechenoperation) nachweisen. Studierende auf der höchsten Stufe können inhaltsgleiche Darstellungen zuordnen, binäre Operationen in verschiedenen Repräsentationsformen identifizieren, sowie symbolische, verbale oder grafische Beispiele produzieren. Auf der mittleren Verständnisstufe können Personen zwischen verschiedenen Repräsentationsformen wechseln, weisen aber Defizite in der Identifikationsphase auf. Studierende auf der untersten Verständnisstufe erreichen in allen drei untersuchten Phasen die geringsten Werte.

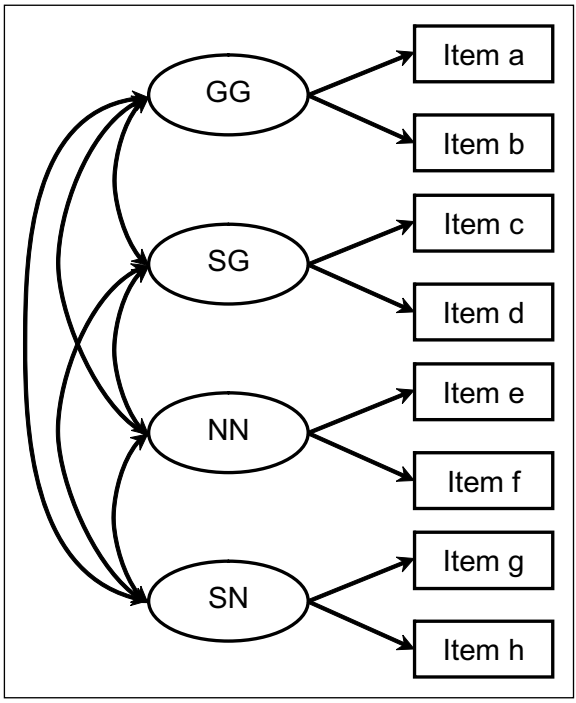
Ob und inwiefern bestimmte Kompetenzprofile mit kognitiven Leistungsaspekten zusammenhängen, wurde unseres Wissens bisher noch nicht untersucht. Allerdings zeigen Befunde zur Präferenz von medial präsentierten Lernumgebungen, dass schwächere Schülerinnen und Schüler von grafischer Repräsentation einer Aufgabe profitieren (vgl. Snow/Yalow 1982).

Ausgehend von dieser Literaturlage erscheint es plausibel, von differenzierten Kompetenzprofilen beim Umgang mit mathematischen Repräsentationsformen auszugehen. Als Grundlage für die vorliegende Untersuchung diente das Kompetenzstrukturmodell, das im Projekt HEUREKO entwickelt wurde (vgl. Bayrhuber u.a., in Vorb.). Die Repräsentationsformen „Tabelle“, „Graf“ und „Situation“ bildeten dabei die Basis für mögliche Dimensionen dieses Modells.

Unter einer Situation verstehen wir dabei die verbale oder ikonische Darstellung einer Realsituation, die noch keine mathematischen Symbole oder Strukturen verwendet. Die Kompetenz im Umgang mit der Repräsentationsform „Term“ wird erst mit Beginn der Klassenstufe 8 aufgebaut und wurde deshalb in der ersten Phase des Projekts, die sich auf die Klassen 7 und 8 konzentriert, nicht berücksichtigt.

Das Modell postuliert, dass die Kompetenz des Problemlösens mit funktionalen Repräsentationen durch die Fähigkeit der Übersetzung zwischen grafischer bzw. numerischer Darstellung und Situation sowie die Verarbeitung innerhalb der numerischen oder grafischen Repräsentationen bestimmt wird:

1. Dimension: Mathematisieren innerhalb der *grafischen* Darstellung (GG): Verarbeiten grafischer Daten ohne Situationsbezug, z.B. Ablesen einzelner Werte aus einem Funktionsgrafen.
2. Dimension: Wechsel zwischen *situativer und grafischer* Repräsentation (SG): Herstellen von Beziehungen zwischen Textelementen und grafischen Daten, meist vorliegend als Funktionsgrafen.
3. Dimension: Mathematisieren innerhalb der *numerischen* Darstellung (NN): Verarbeitung numerischer Daten ohne Situationsbezug.
4. Dimension: Wechsel zwischen *situativer* und *numerischer* Repräsentation (SN): Verknüpfung von Textelementen und numerischen Daten, meist vorliegend in Tabellenform.



Anmerkung: GG = grafisch-grafisch, SG = situativ-grafisch
NN = numerisch-numerisch, SN = situativ-numerisch

Abb. 1: Schematische Darstellung des Kompetenzstrukturmodells

Ziel der vorliegenden Studie ist es, auf Grundlage des beschriebenen Modells Kompetenzprofile von Schülerinnen und Schülern der 7. und 8. Klasse abzuleiten und den Zusammenhang dieser Profile mit der Fähigkeit im figuralen Denken – als einem zentralen Aspekt der kognitiven Leistungsfähigkeit – zu untersuchen. Man versteht darunter die Fähigkeit, figural-räumlich präsentierte Probleme zu lösen (vgl. Heller/Perleth 2000).

2. Methode

2.1 Konstruktion und Überprüfung eines mehrdimensionalen Kompetenzstrukturmodells

Ausgangspunkt der Konstruktion des Kompetenzstrukturmodells war die Identifikation der relevanten Dimensionen auf der Basis bestehender didaktischer Theorien und empirischer Befunde, wie sie in Abschnitt 1 beschrieben wurden. Zur Absicherung des theoretischen Konstrukts und der Vollständigkeit hinsichtlich didaktisch relevanter Aspekte wurden Interviews mit Expertinnen und Experten aus der Fachdidaktik geführt. Das Modell, das als Basis für die Itemkonstruktion diente, wurde als vollständig und relevant für den Inhaltsbereich eingeschätzt.

Die Items zur Operationalisierung der Dimensionen des Modells wurden auf Basis von Schulbuchaufgaben entwickelt, um eine möglichst hohe curriculare Validität zu gewährleisten. Zusätzlich wurden in Einzel- und Gruppeninterviews mit Schülerinnen und Schülern (N = 27) Informationen zur Optimierung des Itempools gewonnen.

Für die endgültige Datenerhebung standen insgesamt 80 Items zur Verfügung, von denen alle Schülerinnen und Schüler je 34 Items zur Bearbeitung vorgelegt wurden (Testheft-Design s. Tabelle 1).

Heft 1	Heft 2	Heft 3	Heft 4	Heft 5	Heft 6	Heft 7
2	3	4	5	6	7	1
1	2	3	4	5	6	7
3	4	5	6	7	1	2

Jedes Item wurde im Mittel von 372 Schülerinnen und Schülern bearbeitet (Bereich = 348–392; SD = 12,06).

Tab.1: Testheft-Design

Die Stichprobe umfasste 37 Gymnasialklassen, 20 (54,1%) Klassen stammten aus Baden-Württemberg und 17 (45,9%) Klassen kamen aus Hessen. 17 (45,9%) Klassen gehörten der siebten Klassenstufe an und 20 (54,1%) Klassen der achten Klassenstufe. Insgesamt nahmen 872 Schülerinnen und Schüler an der Studie teil, die Stichprobe setzte sich aus 471 (54,0%) Mädchen und 399 (46,0%) Jungen zusammen (2 fehlende Angaben). Es wurden ausschließlich Klassen aus Gymnasien berücksichtigt, um die

Schülerpopulation möglichst homogen hinsichtlich sprachlicher Fähigkeiten und schulformtypischer Bildungsanforderungen zu halten.

Das in Abschnitt 1 beschriebene vierdimensionale Kompetenzmodell hat sich auf Basis des informationstheoretischen Indizes „consistent akaike information criterion (CAIC) im Vergleich mit theoretisch plausiblen Alternativmodellen[2] als das am besten passende Modell herausgestellt (vgl. Bayrhuber u.a., in Vorb.). Zur Analyse der Modelle wurde das Multidimensional Random Coefficients Multinomial Logit Model (vgl. Adams/Wilson/Wang 1997) zugrunde gelegt. Dieses ist bereits in mehreren großen Schulleistungsstudien wie TIMSS und PISA zur Modellierung von Schülerkompetenzen zur Anwendung gekommen und ist in der Software ConQuest implementiert (vgl. Wu/Adams/Wilson 2001).

In Tabelle 2 sind die Reliabilitäten der einzelnen Dimensionen sowie die latenten Korrelationen zwischen den vier Dimensionen des Modells aufgeführt. Es wurde die EAP/PV-Reliabilität eingesetzt, die im Rahmen der Rasch-Analyse bestimmt wurde und mit Cronbachs Alpha vergleichbar ist (vgl. Rost 2004).

	SN	**NN**	**GG**	**SG**
SN	0.64			
NN	0.65	0.62		
GG	0.70	0.57	0.52	
SG	0.76	0.71	0.64	0.72

Anmerkung: Die Reliabilitätskoeffizienten sind in der Diagonale aufgeführt.

SN = situativ-numerisch, NN = numerisch-numerisch, GG = grafisch-grafisch, SG = situativ-grafisch.

Tab. 2: Latente Korrelationen und Reliabilität der Dimensionen des vierdimensionalen Modells

2.2 Statistische Analyseverfahren

Latente Klassenanalyse

Als geeignetes Verfahren zur Analyse von Kompetenzprofilen wurde eine latente Klassenanalyse (LCA) durchgeführt. Der LCA liegt ein psychometrisches Modell aus der probabilistischen Testtheorie zugrunde: Es wird angenommen, dass die Zugehörigkeit einer Schülerin oder eines Schülers zu einer latenten, qualitativen Merkmalsklasse ausreichend ist, um deren oder dessen manifeste Merkmalsausprägungen bis auf eine stochastische Komponente vorherzusagen. Das Modell postuliert, dass jede/r Schüler/in einer Klasse angehört, die sich durch ein spezifisches Fähigkeitsprofil auf den Modell-

2 Alternativ wurden ein eindimensionales Modell und ein zweidimensionales Modell, das zwischen numerischer und grafischer Repräsentation differenziert, getestet.

dimensionen auszeichnet. Für jede Klasse werden klassenspezifische Erwartungswerte auf den vier Analyseskalen angenommen. Ein solches Modell hat im Rahmen einer Kompetenzdiagnostik den Vorteil, dass Personen nach Fähigkeitsprofilen unterschieden werden können. Die Feststellung individueller Stärken oder Schwächen ist dann geeigneter Ausgangspunkt einer Fördermaßnahme (vgl. Hartig 2007).

Die LCA wurde mit der Software Latent Gold 4.5 durchgeführt. In einem iterativen Verfahren wurde dabei für post hoc vorgegebene Klassenzahlen nach dem Maximum-Likelihood-Prinzip die optimale Lösung erstellt. Die Entscheidung bei der Auswahl des Lösungsmodells wird anhand des CAIC getroffen (vgl. Vermunt 2004). Die durch die Erhebung in Schulklassen bedingte Mehrebenenstruktur wurde durch den in Latent Gold 4.5 (GClasses, vgl. Vermunt 2008) Expectation-Maximization-Algorithmus berücksichtigt.

2.3 Zusammenhang von Kompetenzprofilen und übergreifenden Leistungsaspekten

Um erste Hinweise auf Erklärungen für Fähigkeitsprofile zu erhalten, wurde das figurale Denken als Kovariate ausgewertet. Es wird vermutet, dass eine hohe Kompetenz im figuralen Denken mit einer hohen Leistung bei den grafisch repräsentierten Aufgaben einhergeht.

Das figurale Denken wurde mit dem Untertest „Figurenanalogien“ (N2) aus dem Kognitiven Fähigkeitstest (KFT 4–12 + R; vgl. Heller/Perleth 2000) erhoben und ist in T-Werten (M = 50, s = 10) angegeben. Die Subskala N 2 besteht aus 25 figuralen Items, auf deren Basis ein Fähigkeitswert bestimmt wurde (Cronbachs Alpha = 0,80). Um zu überprüfen, ob sich die identifizierten Cluster von Schülerinnen und Schülern mit bestimmten Kompetenzprofilen in der Ausprägung des figuralen Denkens unterscheiden, wurde eine univariate Varianzanalyse (ANOVA) durchgeführt.

3. Ergebnisse

3.1 Analyse von Kompetenzprofilen

Zur Überprüfung, ob es Schülerinnen und Schüler gibt, die bezüglich der postulierten vier Dimensionen unterschiedliche Fähigkeitsmuster aufweisen, wurde eine latente Klassenanalyse (LCA) unter Berücksichtigung der Mehrebenenstruktur durchgeführt.

In der 7. Klassenstufe ließen sich keine typischen Profile von Schülerinnen und Schülern nachweisen, welche unterschiedliche Stärken und Schwächen in den verschiedenen Dimensionen des Wechsels von Darstellungsart und Repräsentationsform aufweisen. In der 8. Klasse hingegen erweist sich eine Modellierung mit Fähigkeitsklassen als empirisch tragfähig, da sich eine bedeutsame Interaktion von Clusterzugehörigkeit und Ausprägungen auf den Merkmalsdimensionen zeigt (s. Tabelle 3). Das 6-Cluster-

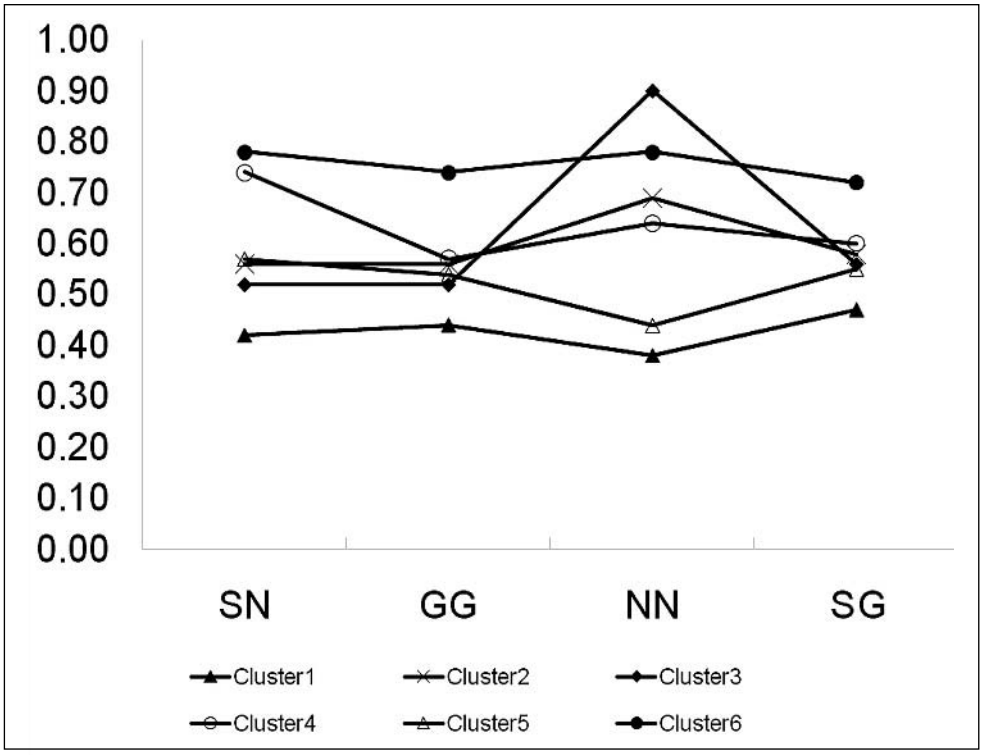
Modell	LL	CAIC	BIC	Class. Err.
1 Cluster	−3037.31	6131	6123	0.0000
2 Cluster	−2949.89	6021	6004	0.1418
3 Cluster	−2896.31	5978	5952	0.1622
4 Cluster	−2862.49	5975	5940	0.1828
5 Cluster	−2840.95	5996	5952	0.2138
6 Cluster	−2772.91	5924	5871	0.1189
7 Cluster	−2758.66	5960	5898	0.2083

Anmerkung: LL=Likelihood; CAIC= Consistent Akaike Information Criterion; BIC= Baysian Information Criterion; Class. Err.= Fehlklassifikationsrate

Tab. 3: Latente Klassenanalyse

Modell weist den geringsten CAIC auf. Auch die im Vergleich zu den übrigen Modellen niedrige Fehlklassifikationsrate weist dieses Modell als optimal aus.

Hier zeigen sich typische Schülerprofile, die grafische Darstellung der Ergebnisse für die 8. Klasse findet sich in der folgenden Abbildung 2.



Anmerkung: SN = situativ-numerisch, GG = grafisch-grafisch, NN = numerisch-numerisch, SG = situativ-grafisch

Abb. 2: Kompetenzprofile in der 8. Klassenstufe

Die Cluster 1 (29% der untersuchten Achtklässler/innen) und 6 (6%) unterscheiden sich lediglich hinsichtlich des Grundniveaus über alle Dimensionen hinweg in ähnlicher Weise. Hier bildet sich also ein generell unterschiedliches Fähigkeitsniveau in allen vier Dimensionen in ähnlicher Weise ab. Für die übrigen Cluster zeigen sich jedoch auffällige typologische Strukturen, die unterschiedliche diagnostisch relevante Kompetenzstrukturen widerspiegeln. So repräsentiert beispielsweise Cluster 4 (15%) Schülerinnen und Schüler, die bei ansonsten eher durchschnittlicher Leistung über eine hohe Kompetenz beim Repräsentationswechsel von der Situation ins Numerische (SN) sowie bei der Verarbeitung innerhalb der numerischen Repräsentation (NN) verfügen. Cluster 3 (15%) hingegen ist gekennzeichnet durch eine markante Stärke im Bereich der numerischen Verarbeitung (NN) bei sonst durchschnittlichen Leistungen. Schülerinnen und Schüler, die dem Cluster 5 (11%) angehören, zeigen einen ähnlichen Profilverlauf, jedoch ist die Fähigkeit zur Verarbeitung innerhalb der numerischen Repräsentation nicht so hoch ausgeprägt wie in Cluster 3. Cluster 2 (21%) ist gekennzeichnet durch eine relative Schwäche innerhalb der numerischen Verarbeitung. Es wird zu prüfen sein, ob diese Profile sich als stabil gegenüber der Wahl anderer Populationen oder gegenüber zeitlichen Entwicklungen erweisen.

3.2 Zusammenhang von Kompetenzprofilen und kognitiver Leistungsfähigkeit

Zum besseren Verständnis dieser Kompetenzprofile wurde der Zusammenhang zwischen der Clusterzugehörigkeit und der Fähigkeit im figuralen Denken analysiert. Es zeigte sich, dass sich die verschiedenen Cluster signifikant ($F_{5, 385} = 7{,}75$; $p < 0{,}001$) und mit hoher Effektstärke ($\eta^2 = 0{,}09$) in ihren Mittelwerten im Untertest N2 im KFT 4–12 + R unterscheiden. Cluster 1 (T = 52,1) und Cluster 3 (T = 51,8) weisen beide vergleichsweise niedrige Werte figuralen Denkens auf (vgl. Tabelle 4), Cluster 3 ist jedoch durch eine hohe Kompetenz im Lösen von numerischen Aufgaben gekennzeichnet. Bei den grafischen Aufgaben und bei Aufgaben, die einen Repräsentationswechsel von der Situation ins Numerische verlangen, zeigen sich in diesem Cluster Schwächen. Der höchste Mittelwert im figuralen Denken zeigt sich in Cluster 6 (T = 60,7), dieser überdurchschnittliche Wert spiegelt sich ebenfalls in den hohen Kompetenzen in den verschiedenen Dimensionen des Mathematiktests wieder. Obwohl Cluster 4 relative Schwächen beim Bearbeiten von grafischen Aufgaben hat, ist der Wert im figuralen Denken (T = 57,8) relativ hoch ausgeprägt.

Cluster	Mittelwert	Standardabweichung	N
1	52,1	7,6	90
2	56,8	9,1	108
3	51,8	9,2	64
4	57,8	6,3	73
5	57,2	9,6	40
6	60,7	5,0	10

Anmerkung: Die Mittelwerte sind in T-Werten angegeben

Tab. 4: Mittelwert der Cluster im Untertest N2 aus dem KFT 4−12 + R

Diese Befunde hinsichtlich des Zusammenhangs von Clusterzugehörigkeit und figuralem Denken geben erste Hinweise auf den Einfluss von Moderatorvariablen auf unterschiedliche Kompetenzprofile von Schülerinnen und Schülern.

4. Diskussion und Ausblick

4.1 Kompetenzprofile

Unsere Ergebnisse zeigen, dass der Repräsentationstyp für den mathematischen Inhaltsbereich „funktionale Veränderung" bedeutsam für die Kompetenzstruktur ist.

Auf der Basis eines vierdimensionalen Kompetenzstrukturmodells konnten in der 8. Klassenstufe Kompetenzprofile ermittelt werden, welche die grafische und numerische Modellierung von Situationen bzw. den Wechsel zwischen diesen Repräsentationen beinhalten. Es konnten Cluster von Schülerinnen und Schülern identifiziert werden, die typische Kompetenzprofile aufweisen. So zeigt beispielsweise eine Gruppe eine hohe Kompetenz beim Repräsentationswechsel von der situativen Darstellung in die numerische Repräsentation sowie innerhalb der numerischen Darstellung, während die Leistungen bei grafischen Aufgaben eher durchschnittlich sind. Bei einer anderen Gruppe mit einem niedrigen Wert im figuralen Denken wurde eine markante Stärke in der numerischen Verarbeitung festgestellt, jedoch Schwächen bei der Bearbeitung von grafischen Aufgaben.

Die Ergebnisse zum Zusammenhang zwischen den Kompetenzprofilen und figuralem Denken sind inkonsistent. Die Erwartung, dass hohe Kompetenzen bei der Bearbeitung von Aufgaben mit grafischem Inhalt mit einer hohen Leistung im figuralen Denken einhergehen, konnte nicht bestätigt werden. Möglicherweise sind die Befunde durch unterschiedliche Kompetenzniveaus der Schülerinnen und Schüler in der Bild-Text-Integration (vgl. Schnotz 2005) zu erklären.

Die Ergebnisse von Snow und Yalow (1982) hinsichtlich der Präferenz schwächerer Schülerinnen und Schüler von grafischen Aufgaben konnten nicht auf den Inhaltsbereich „funktionale Veränderung" übertragen werden. Schwächere Schülerinnen und Schüler weisen keine Stärken bei grafisch repräsentierten Aufgaben auf.

In der 7. Klasse konnten – entgegen unserer Annahme – keine typischen Profile von Schülerinnen und Schülern identifiziert werden, welche unterschiedliche Stärken und Schwächen hinsichtlich des Wechsels von Darstellungsart und Repräsentationsform aufweisen. Diese Schülerinnen und Schüler unterscheiden sich lediglich im Niveau ihrer Lösungshäufigkeit in den untersuchten Dimensionen.

Diese Befunde sind den Ergebnissen von Kleine (2005) ähnlich, der parallele Profilverläufe dargestellt hat. Möglicherweise handelt es sich hier um eine Polytomisierung eines eigentlich kontinuierlichen Fähigkeitsspektrums. Allerdings könnte dies auch darauf zurückzuführen sein, dass die Schülerinnen und Schüler noch am Anfang des Kompetenzerwerbs in diesem Bereich stehen und ihre Fähigkeiten noch nicht in dem Maß ausdifferenziert sind wie in der 8. Klasse. Es kann aber auch als Indiz für eine relative Homogenität des Curriculums zum Einstieg in diesen Bereich gewertet werden.

4.2 Grenzen der vorliegenden Studie und Ausblick

Die Analyse der Binnenstruktur des Kompetenzmodells zeigt noch unbefriedigende Reliabilitäten für drei der vier Dimensionen. Es besteht für diese durch eine zu geringe Anzahl von Items repräsentierten Konstrukte der Bedarf nach zusätzlichen und konstruktvalideren Items, um die psychometrischen Eigenschaften der Skalen zu optimieren. Auch für die Skala „situativ-grafisch" (SG) mit akzeptabler Reliabilität müssen die Items in Bezug auf fokussiertere Iteminhalte optimiert werden, um deren Trennschärfe sowie die Skalenvalidität und -reliabilität zu erhöhen.

Die identifizierten vier Dimensionen werden noch weiteren Strukturanalysen unterzogen, insbesondere wird eine Stufung der Ausprägungen auf den Dimensionen in Kompetenzniveaus angestrebt (vgl. Hartig 2007). Dadurch wird es möglich, zwischen Personen mit unterschiedlichen Teilkompetenzniveaus und Teilkompetenzprofilen zu unterscheiden und perspektivisch Kompetenzentwicklungen abzubilden.

Es haben sich noch nicht genügend Hinweise zu grundsätzlichen Zusammenhängen von spezifischen Kompetenzprofilen beim Repräsentationswechsel und einzelnen Moderatorvariablen ergeben. Die vorliegende Studie ist explorativ angelegt, sodass weitere diagnostisch relevante Zusammenhänge insbesondere unter Einbeziehung zusätzlicher Moderatorvariablen geprüft werden müssen. Dies sollen z.B. verbale Fähigkeiten (vgl. Heller/Perleth 2000) zur Überprüfung der Bedeutung des Aufgabenkontexts („Situation") und weitere grafisch-räumliche Fähigkeiten sein (vgl. Jäger/Süß/Beauducel 1997).

Ob sich bedeutsam unterschiedliche Verteilungen der Typen zwischen verschiedenen Schulklassen ergeben, soll im weiteren Projektverlauf insbesondere unter Einbeziehung von Längsschnittdaten differenziert untersucht werden.

4.3 Relevanz für die Bildungsforschung und für die Schulpraxis

Im Projekt HEUREKO wird auf fundamentale und langfristig auszubildende mathematische Kompetenzen im Inhaltsbereich Wachstum und Veränderung fokussiert, welche mit Hilfe psychometrischer Modelle empirisch zugänglich gemacht werden. Das postulierte Kompetenzstrukturmodell und die gefundenen Kompetenzprofile können die Basis für ein Diagnoseinstrument mathematischer Problemlösefähigkeit beim Umgang mit Funktionen bilden. Diese kann sowohl zur Analyse von Entwicklungsverläufen in Längsschnittstudien dienen als auch in der Schulpraxis eingesetzt werden. Ziel des Diagnostikums ist es, Erkenntnisse zu Förderbedarf und Förderungsmöglichkeiten auf Lerngruppen- und Individuenebene zu gewinnen.

Zusätzlich werden im Projekt grundlegende, über den Gegenstandsbereich des Projektes hinaus anwendbare methodische Vorgehensweisen zur curricular validen Erfassung von Kompetenzstrukturen anhand von multidimensionalen Modellen für die Bildungsforschung entwickelt. Damit konnte in einem curricular zentralen Kompetenzbereich der Grundstein für eine differenzierte Kompetenzmessung gelegt werden.

Literatur

Ainsworth, S.E./Bibby, P.A./Wood, D.J. (2002): Examining the effects of different multiple representational systems in learning primary mathematics. In: Journal of the Learning Sciences 11, H. 1, S. 25–62.

Adams, R.J./Wilson, M./Wang, W. (1997): The multidimensional random coefficient multinominial logit model. In: Applied Psychological Measurement 21, S. 1–13.

Bayrhuber, M./Bruder, R./Leuders, T./Wirtz, M. (in Vorb.): Unidimensional or Multidimensional? Assessing and Modelling Mathematical Competence Structure.

Bodemer, D./Ploetzner, R./Feuerlein, I./Spada, H. (2004): The active integration of information during learning with dynamic and interactive visualisations. In: Learning and Instruction 14, S. 325–341.

Goldin, G.A. (1998): Representational systems, learning and problem solving in mathematics. In: Journal of Mathematical Behavior 17, H. 2, S. 137–165.

Hartig, J. (2007): Skalierung und Definition von Kompetenzniveaus. In: Beck, B./Klieme, E. (Hrsg.): Sprachliche Kompetenzen. Konzepte und Messung. DESI-Ergebnisse Band 1. Weinheim: Beltz, S. 83–99.

Heller, K.A./Perleth, Ch. (2000): Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision (KFT 4–12+ R). Göttingen: Hogrefe.

Jäger, A.O./Süß, H.-M./Beauducel, A. (1997): Berliner Intelligenzstrukturtest Form 4 (BIS-T4). Göttingen: Hogrefe.

Kleine, M. (2005): Latent-Class-Analyse: Ein Bindeglied zwischen Empirie und Theorie zur quantitativen Erfassung mathematischer Leistungen. In: Journal für Mathematik-Didaktik 26, H. 2, S. 97–113.

Kozma, R.B./Russell, J. (1997): Multimedia and understanding: Expert and novice responses to different representations of chemical phenomena. In: Journal of Research in Science Teaching 34, S. 593–619.

Malle, G. (2000): Zwei Aspekte von Funktionen: Zuordnung und Kovariation. In: Mathematik lehren, H. 103, S. 4–7.

Pesonen M./Ehmke T./Haapasalo L. (2005): Solving mathematical problems with dynamical sketches: a study on binary operations. In: Problem Solving in Mathematics Education. Proceedings of the Promath Meeting June 30–July 2, in Lahti, S. 127–140.

Rost, J. (2004): Lehrbuch Testtheorie/Testkonstruktion. Bern: Hans Huber.

Schnotz, W. (2005): An integrated model of text and picture comprehension. In: Mayer, R.E. (Hrsg.): The Cambridge handbook of multimedia learning. Cambridge: Cambridge University Press, S. 49–69.

Seufert, T. (2003): Supporting coherence formation in learning from multiple representations. In: Learning and Instruction 13, S. 227–237.

Snow, R.E./Yalow, E. (1982): Education and intelligence. In: Sternberg, R.J. (Hrsg.): A handbook of human intelligence. Cambridge: Cambridge University Press, S. 493–586.

Swan, M. (1985): The Language of Functions and Graphs. Nottingham: Shell Centre for Mathematical Education.

Vermunt, J.K. (2004): Multilevel latent class models. In: Sociological Methodology 33, S. 213–239.

Vermunt, J.K. (2008): Latent class and finite mixture models for multilevel data sets. In: Methods in Medical Research 17, H. 1, S. 33–51.

Wu, M.L./Adams, R.J./Wilson, M. (2001): ACER ConQuest version 2.0: generalised item response modelling software. ACER Press.

Anschrift der Autor/innen

Dr. Marianne Bayrhuber, Pädagogische Hochschule Freiburg,
Kunzenweg 21, D-79117 Freiburg
E-Mail: bayrhuber@ph-freiburg.de

Prof. Dr. Timo Leuders, Pädagogische Hochschule Freiburg, Kunzenweg 21,
D-79117 Freiburg
E-Mail: leuders@ph-freiburg.de

Prof. Dr. habil. Regina Bruder, Technische Universität Darmstadt, Schlossgartenstraße 7,
D-64289 Darmstadt
E-Mail: bruder@mathematik.tu-darmstadt.de

Prof. Dr. Markus Wirtz, Pädagogische Hochschule Freiburg, Kunzenweg 21,
D-79117 Freiburg
E-Mail: markus.wirtz@ph-freiburg.de

Andreas Frey/Nicki-Nils Seitz

Multidimensionale adaptive Kompetenzdiagnostik: Ergebnisse zur Messeffizienz

Projekt MAT[1]

1. Theoretischer Ansatz und Fragestellungen

Infolge des mittelmäßigen Abschneidens von Schülerinnen und Schülern aus Deutschland bei internationalen Vergleichsstudien wie PISA, IGLU oder TIMSS beschäftigt sich die Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland verstärkt mit Möglichkeiten zur Verbesserung schulischer Ausbildung. Als ein Ergebnis wurden seit 2003 bundesweit geltende Bildungsstandards für verschiedene Fächer und Schulabschlüsse beschlossen. Inwieweit diese Standards von Schülerinnen und Schülern erreicht werden, wird seit 2009 durch Tests empirisch im Ländervergleich untersucht. Dabei entsteht ein großer Testaufwand, der erhebliche Kosten mit sich bringt. Um diese zu begrenzen sowie die Kooperationsbereitschaft seitens der Schulen und der Schülerinnen und Schüler langfristig zu sichern, ist nach Wegen zu suchen, die Testungen möglichst effizient zu gestalten.

Eine Möglichkeit zur Steigerung der Messeffizienz im Vergleich zu den bislang eingesetzten Testverfahren besteht im *computerisierten adaptiven Testen* (CAT; vgl. Frey 2007; van der Linden/Glas 2000; Wainer 2000). Bei CAT orientiert sich die Auswahl der Aufgaben, die einem Individuum vorgegeben werden, an dessen Kompetenzniveau. Personen mit hoher Kompetenz bekommen schwierigere Aufgaben vorgelegt als Personen mit niedriger Kompetenz. Durch diese optimierte Aufgabenauswahl müssen jedem Individuum im Vergleich zu konventionellen Testverfahren (FIT)[2] in der Regel nur ca. 50% der Aufgaben präsentiert werden, um eine vergleichbare Messpräzision zu erreichen (vgl. z.B. Frey/Ehmke 2007; Segall 2005).

Das Projekt „Multidimensionale adaptive Kompetenzdiagnostik“ im Rahmen des DFG-Schwerpunktprogramms 1293 beschäftigt sich mit Grundlagenforschung zur vor kurzem entwickelten mehrdimensionalen Erweiterung des ursprünglich eindimensionalen Konzepts adaptiven Testens. Nachfolgend werden zunächst die Grundlagen des multidimensionalen adaptiven Testens (MAT) skizziert und danach Fragestellungen,

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennzeichen: FR 2552/2-1, FR 2552/2-2) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

2 Um komplizierte Formulierungen zu vermeiden, wird in diesem Text einheitlich von „konventionellem Testen“ gesprochen oder die Abkürzung FIT (fixed item testing) verwendet, wenn eine vor der Testung festgelegte Menge von Aufgaben in fester Reihenfolge vorgegeben wird.

Methode und Befunde einer Simulationsstudie zur Steigerung der Messeffizienz durch MAT dargestellt. Der Beitrag schließt mit einer Zusammenstellung offener Forschungsfragen im Bereich MAT und diskutiert die praktische Anwendbarkeit dieser Art des Testens bei der Überprüfung von Bildungsstandards.

1.1 Multidimensionales adaptives Testen (MAT)

In diesem Abschnitt werden die zentralen Aspekte von MAT skizziert. Eine umfassende Darstellung des Forschungsstandes ist bei Frey/Seitz (2009) zu finden. Während beim eindimensionalen computerisierten adaptiven Testen (U-CAT) das Antwortverhalten auf eine latente Dimension zurückgeführt wird, werden bei MAT mehrere latente Dimensionen als ursächlich für die gegebenen Antworten angesehen. Der Zusammenhang zwischen Antwortverhalten und der Ausprägung eines Individuums auf diesen latenten Dimensionen wird durch psychometrische Modelle der Item-Response-Theorie (IRT; vgl. z.B. van der Linden/Hambleton 1997) beschrieben. Durch die Verwendung von IRT-Modellen können interindividuelle Vergleiche von Testergebnissen auch dann durchgeführt werden, wenn Proband/innen verschiedene Aufgaben bearbeitet haben. Bislang wurden bei MAT fast ausschließlich mehrdimensionale IRT-Modelle (MIRT-Modelle, vgl. z.B. Reckase 2009) mit geringer Komplexität verwendet. Bspw. beschreibt Segall (1996) in einer viel beachteten Arbeit die Verwendung des mehrdimensionalen dreiparametrigen logistischen Testmodells (M3PL) im Rahmen von MAT. Beim M3PL wird die Wahrscheinlichkeit einer korrekten Antwort U auf eine Aufgabe i ($U_i = 1$) als Funktion der Ausprägung des untersuchten Individuums auf p latenten Merkmalsdimensionen $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ und drei Itemparametern konzeptualisiert:

$$P(U_i = 1|\boldsymbol{\theta}) = c_i + \frac{1 - c_i}{1 + \exp[-D\mathbf{a}'_i(\theta - b_i\mathbf{1})]} \tag{1}$$

Hierbei beschreiben $\mathbf{a}'_i$ einen $(1 \times p)$-Vektor der dimensionsspezifischen Diskrimination, b_i die Schwierigkeit und c_i den Pseudo-Rateparameter einer Aufgabe i. Durch die Multiplikation der Aufgabenschwierigkeit mit dem mit Einsen gefüllten $(p \times 1)$-Vektor $\mathbf{1}$ wird die Aufgabenschwierigkeit auf alle untersuchten Dimensionen übertragen. Der Term im Exponenten wird mit der Konstanten $D = 1.7$ multipliziert, um das Modell dem Normal-Ogiven-Modell anzupassen. Über das Modell (1) hinaus können prinzipiell viele weitere MIRT-Modelle bei MAT zum Einsatz kommen.

Durch die so gegebene Flexibilität ist MAT gut für die Messung der komplexen, mehrdimensionalen theoretischen Kompetenzmodelle der Bildungsstandards geeignet. Bspw. unterscheidet das theoretische Kompetenzmodell der Bildungsstandards in Mathematik für den Mittleren Schulabschluss (vgl. z.B. Blum u.a. 2005; Ehmke u.a. 2006) sechs mathematische Teilkompetenzen (mathematisch argumentieren, Probleme mathematisch lösen, mathematisch modellieren, mathematische Darstellungen verwenden,

mit Mathematik symbolisch/technisch umgehen, mathematisch kommunizieren), die durch Anforderungen in fünf mathematischen Inhaltsbereichen, den Leitideen (Zahl, Messen, Raum und Form, funktionaler Zusammenhang, Daten und Zufall), angesprochen werden können. Die einer Aufgabe inhärente kognitive Komplexität wird zusätzlich durch drei Anforderungsbereiche beschrieben (reproduzieren, Zusammenhänge herstellen, verallgemeinern und reflektieren). Die Struktur derartiger mehrdimensionaler Kompetenzmodelle kann in psychometrischen MIRT-Modellen direkt abgebildet und durch MAT einer Messung zugeführt werden.

Neben dem psychometrischen Modell besteht ein wesentliches Element eines multidimensionalen adaptiven Tests im *Algorithmus*, der während der Testung für die Aufgabenauswahl eingesetzt wird. Die beiden einflussreichsten Ansätze zur Aufgabenauswahl sind der bayesianische Ansatz von Segall (1996) und der Maximum-Likelihood-Ansatz von van der Linden (1999). Der Ansatz von Segall erfuhr bislang etwas größere Resonanz. Bei diesem wird jeweils diejenige Aufgabe aus einem zuvor mit einem MIRT-Modell kalibrierten Itempool ausgewählt und zur Bearbeitung vorgegeben, welche die größte Reduktion im Volumen des Konfidenzellipsoids (mehrdimensionales Pendant eines Konfidenzintervalls) des geschätzten p-dimensionalen Merkmalsvektors $\hat{\boldsymbol{\theta}}$ bewirkt. Es wird also die Aufgabe ausgewählt, deren Vorgabe die größte Steigerung der Messpräzision liefert.

Die beim bayesianischen MAT-Ansatz von Segall (1996) verwirklichte Art der Aufgabenauswahl verspricht die ohnehin sehr hohe Messeffizienz von U-CAT weiter steigern zu können, da zusätzlich Erkenntnisse über korrelative Zusammenhänge zwischen den zu messenden Merkmalsdimensionen direkt bei der Messung berücksichtigt werden können. Werden mehrere korrelierte Dimensionen erhoben, dann geben die Antworten einer Testperson auf Aufgaben, die eine Dimension messen, nicht nur Hinweise über die Ausprägung der Testperson auf dieser Dimension, sondern auch über deren Ausprägung auf den anderen Dimensionen. Zeigt bspw. eine Schülerin bzw. ein Schüler eine hohe Kompetenz in der mathematischen Leitidee „Zahl“, dann ist es wahrscheinlich (obgleich nicht sicher), dass sie bzw. er auch eine hohe Kompetenz in den anderen vier mathematischen Leitideen der Bildungsstandards aufweist. Dies führt dazu, dass bei MAT ein hohes Maß diagnostischer Information pro Aufgabe gewonnen wird. Bei Simulationsstudien (vgl. Liu 2007; Segall 1996; Wang/Chen 2004), Simulationsstudien auf Basis empirischer Daten (vgl. Gardner/Kelleher/Pajer 2002; Haley u.a. 2006; Li/Schafer 2005; Petersen u.a. 2006) und einer empirischen Anwendung (vgl. Mulcahey u.a. 2008) zeigten sich entsprechend dieser Annahme Vorteile von MAT gegenüber U-CAT und FIT hinsichtlich der Messeffizienz. Bislang ist jedoch noch nicht bekannt, wie groß die zu erwartende Messeffizienzsteigerung genau ist, wenn typische Gegebenheiten groß angelegter Vergleichsstudien vorliegen. Entsprechende Ergebnisse wären aber nötig, um die Zweckmäßigkeit eines Einsatzes von MAT bei den Untersuchungen zu den Bildungsstandards einschätzen zu können.

1.2 Fragestellungen

Um das Ausmaß und die Bedingungen von Steigerungen der Messeffizienz durch MAT zu verstehen, wurden die folgenden vier Fragestellungen untersucht:

1. Wie unterscheiden sich die Testalgorithmen FIT, U-CAT und MAT hinsichtlich der Messeffizienz?
2. Wie unterscheiden sich die Testalgorithmen FIT, U-CAT und MAT in Abhängigkeit der Anzahl untersuchter Dimensionen hinsichtlich der Messeffizienz?
3. Wie unterscheiden sich die Testalgorithmen FIT, U-CAT und MAT in Abhängigkeit der Korrelation zwischen den untersuchten Dimensionen hinsichtlich der Messeffizienz?
4. Welche Unterschiede sind zwischen den Testalgorithmen FIT, U-CAT und MAT hinsichtlich der Messeffizienz bei Bedingungen zu beobachten, die typisch für Untersuchungen zu den Bildungsstandards sind?

2. Methode

Die Untersuchung der Fragestellungen erfolgte mit einer Simulationsstudie auf der Grundlage eines vollständig gekreuzten experimentellen Versuchsplans mit dem dreifach gestuften Faktor Testalgorithmus (FIT, U-CAT, MAT), dem vierfach gestuften Faktor Dimensionsanzahl (2, 3, 4, 5) und dem dreifach gestuften Faktor Korrelation zwischen den Dimensionen (.00, .50, .85).

Als Grundlage der Simulation wurden zunächst zufallsabhängige Aufgabenschwierigkeiten und Personenparameter erzeugt. Je untersuchter Dimension wurden 200 Aufgabenschwierigkeiten aus einer Gleichverteilung im Bereich von -4 bis 4 Logits gezogen, $b_{ip} \sim U(-4,4)$. Für jede Kombination der Faktoren Dimensionsanzahl und Korrelation wurden 1000 multivariat normalverteilte Personenparameter unter Setzung eines Mittelwertsvektors $\boldsymbol{\mu}$ und einer Matrix $\boldsymbol{\Phi}$ mit den Korrelationen zwischen den p Dimensionen der jeweiligen Versuchsbedingung erzeugt, $\boldsymbol{\theta} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Phi})$. Die Mittelwerte der Dimensionen wurden in allen Versuchsbedingungen auf 0 festgelegt. Die Korrelationen zwischen den Dimensionen wurden auf den Wert der jeweiligen Stufe des Faktors Dimension gesetzt. In der Versuchsbedingung mit drei Dimensionen, die mit .85 korreliert sind, fand bspw. die folgende Matrix $\boldsymbol{\Phi}$ Verwendung:

$$\boldsymbol{\Phi} = \begin{pmatrix} 1 & .85 & .85 \\ .85 & 1 & .85 \\ .85 & .85 & 1 \end{pmatrix}$$

Unter Verwendung der erzeugten Aufgabenschwierigkeiten und der Personenparameter wurden unter Annahme der Gültigkeit des mehrdimensionalen Raschmodells zufallsabhängige Antworten aller virtuellen Proband/innen auf alle virtuellen Aufgaben erzeugt.

Das mehrdimensionale Raschmodell ergibt sich aus dem Modell (1), wenn die Annahmen getroffen werden, dass die durch $\mathbf{a'}_i$ repräsentierten Ladungen der Aufgaben auf den latenten Merkmalsdimensionen alle den gleichen Wert haben und dass Raten keinen Einfluss auf die Lösungswahrscheinlichkeit hat ($c_i = 0$).

Die resultierenden Antwortmatrizen wurden daraufhin genutzt, um die Testung mit den drei Testalgorithmen FIT, U-CAT und MAT zu simulieren. Dabei erfolgte bei FIT die Aufgabenauswahl per Zufall. Bei U-CAT und MAT wurde nur die erste Aufgabe zufällig ausgewählt. Danach orientierte sich die Aufgabenauswahl am Kriterium maximaler Information. Bei MAT wurde im Rahmen des von Segall (1996) beschriebenen bayesianischen Ansatzes zusätzlich die als bekannt angenommene Kovarianzmatrix der mehrdimensionalen a-priori-Verteilung der Kompetenzen berücksichtigt. Die Personenparameterschätzung erfolgte für FIT, U-CAT und MAT mit MAPs (Modal a-posteriori estimates), wobei bei MAT wiederum die Kovarianzmatrix der mehrdimensionalen a-prioi-Verteilung der Kompetenzen verwendet wurde (vgl. Segall 1996). Um der statistischen Unsicherheit der Simulation gerecht zu werden, wurden in jeder Zelle des Versuchsplans 200 Replikationen realisiert. Alle Berechnungen erfolgten mit dem Statistikpaket SAS 9.2.

Als zentrale abhängige Variable wurde die Messeffizienz (*ME*) für alle Faktorstufenkombinationen berechnet. Die Messeffizienz mehrdimensionaler Tests lässt sich in Anlehnung an Frey (2007) und Segall (2005) als Quotient von Messpräzision und Testlänge bestimmen. Als Kennwert für die Messpräzision dient der Kehrwert der mittleren quadratischen Abweichung der wahren Merkmalsausprägung θ von der geschätzten Merkmalsausprägung $\hat{\theta}$ für die $k = 1$ bis n Personen. Die Testlänge wird durch die Anzahl vorgegebener Aufgaben m definiert. Unter Berücksichtigung aller $j = 1$ bis p Dimensionen und nach Umformung, berechnet sich die mehrdimensionale Messeffizienz (ME_{MD}) folgendermaßen:

$$ME_{MD} = \frac{1}{p}\sum_{j=1}^{p} ME_j = \frac{n}{p}\sum_{j=1}^{p} = \left[m_j \sum_{k=1}^{n} (\hat{\theta}_{kj} - \theta_{kj})^2 \right]^{-1} \tag{2}$$

3. Ergebnisse

Tabelle 1 zeigt die Messeffizienz in Abhängigkeit von Testalgorithmus, Dimensionsanzahl und der Höhe der Korrelationen zwischen den Dimensionen.

Im Hinblick auf die *Fragestellung 1* zeigt sich, dass durch FIT nur eine niedrige Messeffizienz erzielt wird. Durch U-CAT kann sie signifikant auf etwa das Dreifache gesteigert werden. Eine zusätzliche nominale Steigerung ist bei der Verwendung von MAT zu verzeichnen (Abbildung 1).

Bezüglich der *Fragestellung 2* zeigt sich, dass der nominale Messeffizienzvorteil von MAT gegenüber U-CAT nicht auf die Anzahl der gemessenen Dimensionen zurück-

		Testalgorithmus					
		FIT		U-CAT		MAT	
Dimensionsanzahl	**Korrelation**	***M***	***SE***	***M***	***SE***	***M***	***SE***
2	0.00	0.23	0.03	0.68	0.03	0.68	0.03
	0.50	0.23	0.03	0.68	0.03	0.71	0.03
	0.85	0.23	0.03	0.68	0.03	0.84	0.04
3	0.00	0.23	0.03	0.68	0.03	0.66	0.03
	0.50	0.23	0.03	0.68	0.03	0.69	0.04
	0.85	0.23	0.03	0.68	0.03	0.86	0.05
4	0.00	0.23	0.03	0.68	0.03	0.65	0.04
	0.50	0.23	0.03	0.68	0.03	0.70	0.04
	0.85	0.23	0.03	0.68	0.03	0.88	0.06
5	0.00	0.23	0.03	0.68	0.03	0.65	0.03
	0.50	0.23	0.03	0.68	0.03	0.69	0.04
	0.85	0.23	0.03	0.68	0.03	0.87	0.06

Anmerkung: FIT = Konventioneller Test mit fester Aufgabenmenge in fester Reihenfolge, U-CAT = eindimensionaler computerisierter adaptiver Test, MAT = multidimensionaler adaptiver Test.

Tab. 1: Messeffizienz als Funktion von Testalgorithmus, Dimensionsanzahl und Korrelation zwischen den Dimensionen

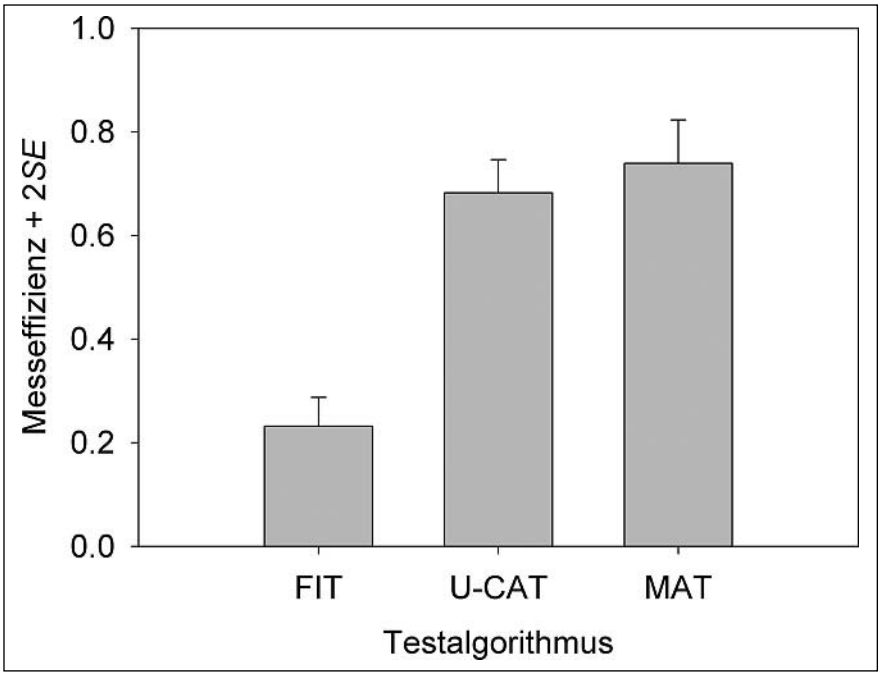


Abb. 1: Messeffizienz in Abhängigkeit des Testalgorithmus

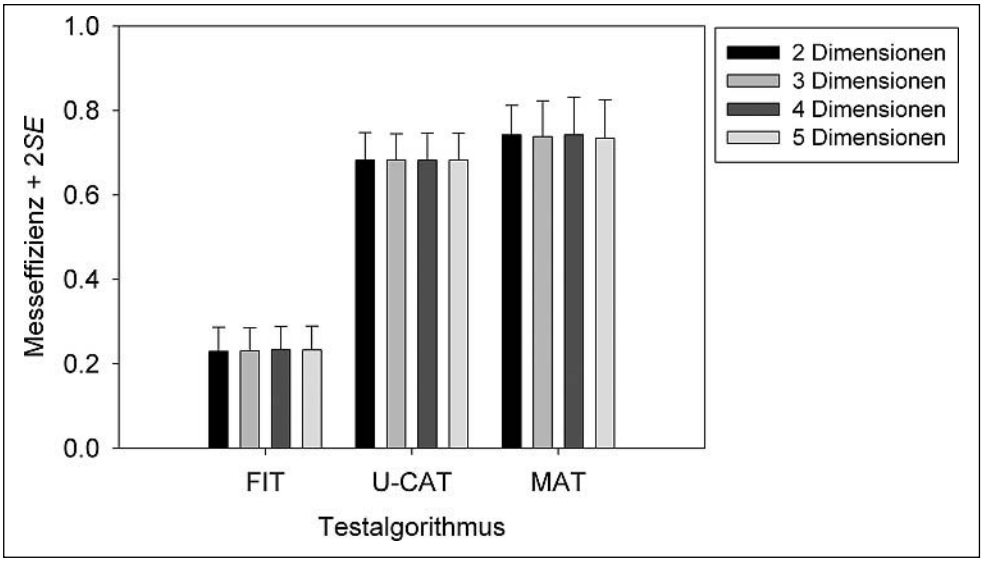


Abb. 2: Messeffizienz in Abhängigkeit von Testalgorithmus und Dimensionsanzahl

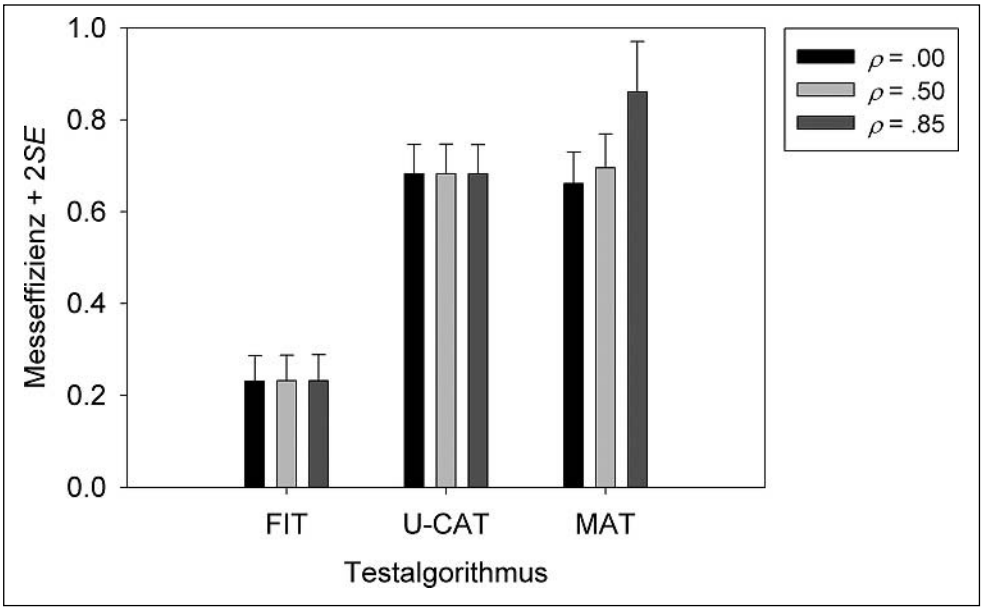


Abb. 3: Messeffizienz in Abhängigkeit von Testalgorithmus und Korrelation zwischen den Dimensionen

zuführen ist. Die Messeffizienz variiert bei MAT nicht signifikant in Abhängigkeit davon, ob 2, 3, 4 oder 5 Dimensionen gemessen werden (Abbildung 2).

Zur *Fragestellung 3* ergibt sich, dass die Messeffizienz bei MAT signifikant von der Höhe der Korrelation zwischen den untersuchten Dimensionen abhängt. Die Messeffizienz fällt bei MAT bei einer Korrelation von .85 signifikant höher aus als bei niedrigeren Korrelationen (Abbildung 3).

Die *Fragestellung 4* zielt direkt auf Bedingungen ab, die für Untersuchungen zu den Bildungsstandards charakteristisch sind (5 Dimensionen mit .85 korreliert; vgl. Prenzel/Blum 2007). Auch hier ergibt sich für FIT eine sehr niedrige Messeffizienz; bei U-CAT ist sie signifikant höher. Durch MAT resultiert gegenüber U-CAT eine zusätzliche signifikante Steigerung (Abbildung 4). Die Messeffizienz ist bei MAT rund 3.5-mal so hoch wie bei FIT. Für die Praxis bedeutet dies, dass ein konventioneller Test, bei dem jede Schülerin bzw. jeder Schüler bspw. 35 Aufgaben vorgegeben werden, durch MAT ohne Messpräzisionsverlust auf eine mittlere Länge von 10 Aufgaben verkürzt werden kann.

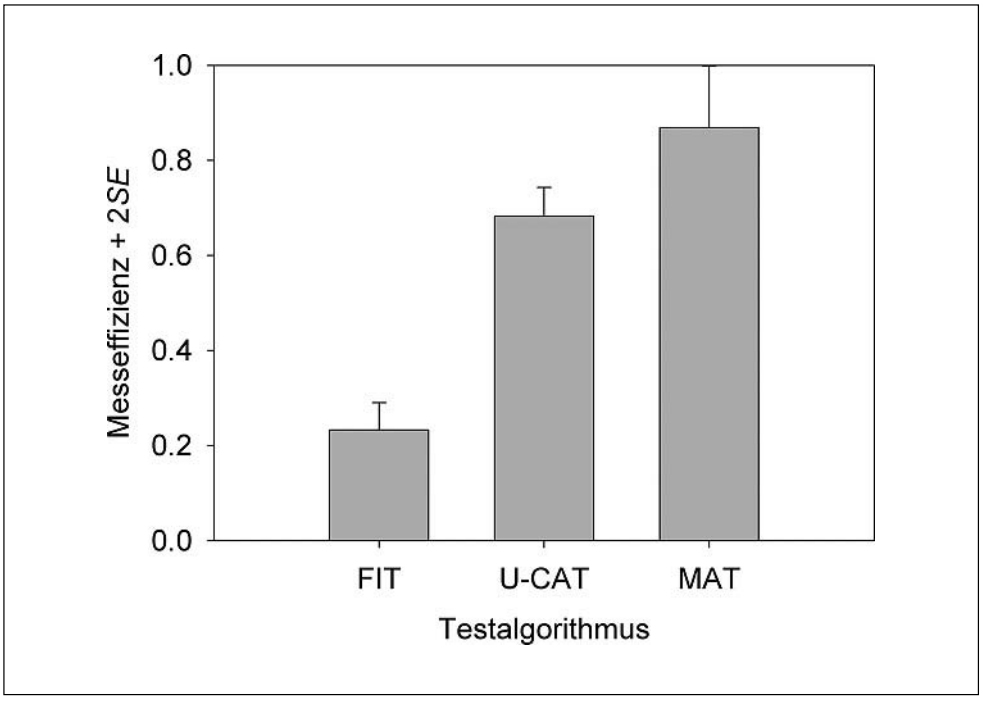


Abb. 4: Messeffizienz bei für Erhebungen zu den Bildungsstandards typischen Bedingungen nach Testalgorithmus

Es ist zusammenzufassen, dass durch MAT die Messeffizienz gegenüber FIT erheblich gesteigert werden kann. Das Ausmaß der Effizienzsteigerungen ist dabei stark von den Korrelationen zwischen den untersuchten Dimensionen abhängig. Bei Bedingungen,

wie sie für Untersuchungen zu den Bildungsstandards typisch sind, ist durch MAT eine Effizienzsteigerung auf mehr als das Dreieinhalbfache im Vergleich zu FIT zu erzielen. MAT kann somit helfen, den Testaufwand bei Erhebungen zu den Bildungsstandards erheblich zu senken.

4. Diskussion

Bei der Einordnung der Ergebnisse ist zu beachten, dass bei der vorliegenden Studie ein Itempool verwendet wurde, der optimale Eigenschaften für adaptives Testen aufweist. Die berichteten Effizienzsteigerungen sind deshalb als obere Grenze zu sehen, die maximal bei Verwendung von U-CAT und MAT erreicht werden können. Zukünftig ist zu klären, welche Effizienzsteigerungen bei typischen Itempools resultieren. Vorläufige Ergebnisse des Projekts MAT weisen erfreulicherweise darauf hin, dass auch bei nicht optimalen Itempools große Steigerungen der Messeffizienz resultieren. Die Vorteile von MAT gegenüber FIT und U-CAT scheinen vor allem durch hohe Korrelationen zwischen den untersuchten Dimensionen getrieben zu sein. Ist diese Voraussetzung erfüllt, kann schon mit einem relativ kleinen Itempool, einer für adaptives Testen nicht optimalen Verteilung der Aufgabenschwierigkeiten und wenigen Dimensionen eine sehr hohe Messeffizienz erreicht werden.

Die berichteten ersten Projektergebnisse zeichnen ein vorteilhaftes Bild für MAT. Bevor dieser neuen Art des computerisierten Testens jedoch bedenkenlos bei groß angelegten Vergleichsstudien wie den Erhebungen zu den Bildungsstandards eingesetzt werden kann, sind noch einige zentrale Fragen zu untersuchen und zu beantworten. Eine zentrale Herausforderung besteht nach unserem Dafürhalten in der Implementierung komplexer MIRT-Modelle in MAT. Komplexe MIRT-Modelle sind wünschenswert, um die Strukturen der zugrunde liegenden theoretischen Kompetenzmodelle direkt bei der Messung abzubilden. Hierdurch kann eine optimale Passung von theoretischem Modell, psychometrischem Modell und Messinstrument als Grundlage einer theoriebasierten Testwertinterpretation erzielt werden. Dies würde einen erheblichen Fortschritt bedeuten, da die theoretischen Modellannahmen direkt in empirischen Beobachtungsdaten abgebildet werden würden. Bislang sind allerdings nur wenige Modellklassen im Rahmen von MAT nutzbar gemacht worden. Neben dem oben beschriebenen M3PL wurde von Segall (2001) ein generalisierter Ansatz zur Verwendung hierarchischer MIRT-Modelle bei MAT eingeführt. Weitere Modelle wurden bei MAT noch nicht eingesetzt. Im Hinblick auf die theoretischen Modelle der Bildungsstandards sollten zwei weitere Modellklassen in MAT implementiert werden:

Erstens sind komplexe mehrdimensionale psychometrische Modelle nötig, bei denen Aufgaben nicht nur auf einer Dimension (*between-item multidimensionality*) sondern auf mehreren Dimensionen laden können (*within-item multidimensionality*). Bei den Bildungsstandards in Mathematik für den Mittleren Schulabschluss wird bspw. angenommen, dass für die Bewältigung vieler Aufgaben mehrere Kompetenzen benötigt werden. Es ist also ein psychometrisches Modell zu formulieren, das als latente Dimen-

sionen nicht nur die oben genannten fünf Leitideen, die sechs Kompetenzen und die drei Anforderungsbereiche enthält sondern auch erlaubt, dass einzelne Aufgaben auf mehreren Kompetenzdimensionen laden. Die Schätzung solch komplexer Modelle ist vor allem aufgrund ihrer hohen Parameteranzahl auch mit den in der empirischen Bildungsforschung verfügbaren großen Stichproben bei Verwendung konventionellen Testens problematisch (vgl. Carstensen/Frey 2007). Die resultierenden Personenparameterverteilungen sind aufgrund ihrer hohen statistischen Unsicherheit meistens nicht zur Ergebnisrückmeldung geeignet. Zur Verbesserung der Schätzung von Personenparameterverteilungen bei Verwendung komplexer MIRT-Modelle kann aber die hohe Messeffizienz von MAT genutzt werden. Der Messeffizienzvorteil von MAT gegenüber FIT lässt sich also nicht nur zur Reduzierung des Testaufwands und der damit verbundenen Kosten einsetzen, sondern auch dazu, differenziertere und besser auf zugrundeliegende theoretische Annahmen abgestimmte Ergebnisse bereitzustellen. Hierfür müsste jedoch eine Kalibrierungsstudie mit vermutlich sehr großer Stichprobe vorgeschaltet werden, um die für MAT benötigten Itemparameter und Korrelationen zwischen den zu messenden Merkmalsdimensionen erwartungstreu und konsistent zu schätzen. Dieser zusätzliche Initialaufwand kann durch die hohe Messeffizienz von MAT bei wiederholten Testungen vermutlich kompensiert werden. Bei einmaligem Einsatz würde sich der Aufwand jedoch nicht lohnen.

Zweitens werden bei den Erhebungen zu den Bildungsstandards Aufgaben in der Regel nicht einzeln, sondern gruppiert zu sogenannten *Testlets* vorgegeben. Ein Testlet besteht aus einem Stimulus und einer Anzahl von Einzelaufgaben, die sich auf diesen Stimulus beziehen. Hierdurch kann die bei herkömmlichen IRT-Modellen getroffene Annahme lokaler stochastischer Unabhängigkeit verletzt werden. Diese Annahme drückt aus, dass die Antwort eines Individuums auf eine Aufgabe eines Tests unabhängig davon ist, wie das Individuum andere Aufgaben des gleichen Tests beantwortet hat. Lokale Abhängigkeiten führen in der Regel zu einer Unterschätzung der Standardfehler der geschätzten Personenparameter und somit zu einer Überschätzung der Messpräzision des Tests (vgl. Pommerich/Segall 2008; Sireci/Thissen/Wainer 1991; Wainer/Bradlow/Wang 2007). Um diesem Problem zu begegnen, können lokale Abhängigkeiten explizit durch IRT-Modelle modelliert werden (vgl. z.B. Wang/Wilson 2005). Für U-CAT liegen mehrere Ansätze zur Berücksichtigung von Testlets vor (vgl. z.B. Scalise/Wilson 2007; Vos/Glas 2000; Wainer/Bradlow/Du 2000). Bislang fehlt aber noch die Generalisierung eines Ansatzes auf MAT.

Als spezielle Art des Testens misst sich der Erfolg von MAT letztendlich aber daran, ob es sich im praktischen Einsatz bewährt. Bislang liegt erst eine Studie zur praktischen Verwendung von MAT mit realen Proband/innen vor (vgl. Mulcahey u.a. 2008). Aufgrund der Möglichkeit, durch MAT den Testaufwand und die damit verbundenen Kosten erheblich zu senken, ist für die Zukunft mit weiteren praktischen Anwendungen zu rechnen.

Literatur

Blum, W./Drüke-Noe, C./Leiss, D./Wiegand, B./Jordan, A. (2005): Zur Rolle von Bildungsstandards für die Qualitätsentwicklung im Mathematikunterricht. In: Zentralblatt für Didaktik der Mathematik 37, S. 267–274.

Carstensen, C.H./Frey, A. (2007, August). Competency profiles from standard assessments. Paper presented at the 12th Biennial Conference for Research on Learning and Instruction (EARLI), Budapest, Ungarn.

Ehmke, T./Leiss, D./Blum, W./Prenzel, M. (2006): Entwicklung von Testverfahren für die Bildungsstandards Mathematik. Rahmenkonzeption, Aufgabenentwicklung, Feld- und Haupttest. In: Unterrichtswissenschaft 34, S. 220–238.

Frey, A. (2007): Adaptives Testen. In: Moosbrugger, H./Kelava, A. (Hrsg.): Testtheorie und Fragebogenkonstruktion. Berlin, Heidelberg: Springer, S. 261–278.

Frey, A./Ehmke, T. (2007): Hypothetischer Einsatz adaptiven Testens bei der Messung von Bildungsstandards in Mathematik. In: Prenzel, M./Gogolin, I./Krüger, H.-H. (Hrsg.): Kompetenzdiagnostik. 8. Sonderheft der Zeitschrift für Erziehungswissenschaft. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 169–184.

Frey, A./Seitz, N.N. (2009): Multidimensional Adaptive Testing in Educational and Psychological Measurement: Current State and Future Challenges. In: Studies in Educational Evaluation 35, S. 89–94.

Gardner, W./Kelleher, K.J./Pajer, K.A. (2002): Multidimensional adaptive testing for mental health problems in primary care. In: Medical Care 40, H. 9, S. 812–823.

Haley, S.M./Pengsheng, N./Ludlow, L.H./Fragala-Pinkham, M.A. (2006): Measurement precision and efficiency of multidimensional computer adaptive testing in physical functioning using the pediatric evaluation of disability inventory. In: Archives of Physical Medicine and Rehabilitation 87, S. 1223–1229.

Li, Y.H./Schafer, W.D. (2005): Trait Parameter Recovery Using Multidimensional Computerized Adaptive Testing in Reading and Mathematics. In: Applied Psychological Measurement 29, S. 3–25.

Liu, J. (2007): Comparing multi-dimensional and uni-dimensional computer adaptive strategies in psychological and health assessment. Unveröffentlichte Dissertation, Columbia University.

Mulcahey, M.J./Haley, S.M./Duffy, T./Pengsheng, N./Betz, R.R. (2008): Measuring physical functioning in children with spinal impairments with computerized adaptive testing. In: Journal of Pediatric Orthopaedics 28, S. 330–335.

Petersen, M.A./Groenvold, M./Aaronson, N./Fayers, P./Sprangers, M./Bjorner, J.B. (2006): Multidimensional Computerized Adaptive Testing of the EORTC QLQ-C30: Basic Developments and Evaluations. In: Quality of Life Research 15, S. 315–329.

Pommerich, M./Segall, D.O. (2008): Local dependence in an operational CAT: diagnosis and implications. In: Journal of Educational Measurement 45, S. 201–223.

Prenzel, M./Blum, W. (Hrsg.) (2007): Erprobung von Aufgaben zur Überprüfung der Anforderungen der Bildungsstandards in Mathematik: Technischer Bericht. Kiel: IPN.

Reckase, M.D. (2009): Multidimensional Item Response Theory. Dordrecht: Springer.

Scalise, K./Wilson, M. (2007): Bundle models for computerized adaptive testing in e-learning assessment. In: Weiss, D.J. (Hrsg.): Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing. http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/cat07scalise.pdf [20.11.2008].

Segall, D.O. (1996): Multidimensional adaptive testing. In: Psychometrika 61, S. 331–354.

Segall, D.O. (2001): General ability measurement: An application of multidimensional item response theory. In: Psychometrika 66, S. 79–97.

Segall, D.O. (2005): Computerized Adaptive Testing. In: Kempf-Leonard, K. (Hrsg.): Encyclopedia of Social Measurement. New York: Academic Press, S. 429–438.

Sireci, S.G./Thissen, D./Wainer, H. (1991): On the reliability of testlet-based tests. In: Journal of Educational Measurement 28, S. 237–247.
van der Linden, W.J. (1999): Multidimensional adaptive testing with a minimum error-variance criterion. In: Journal of Educational and Behavioral Statistics 28, S. 398–412.
van der Linden, W.J./Glas, C.A.W. (Hrsg.) (2000): Computerized Adaptive Testing: Theory and Practice. Dordrecht/Boston/London: Kluwer Academic Publishers.
van der Linden, W.J./Hambleton, R.K. (Hrsg.) (1997): Handbook of modern item response theory. New York: Springer.
Vos, H.J./Glas, C.A.W. (2000): Testlet-based adaptive mastery testing. In: van der Linden, W.J./Glas, C.A.W. (Hrsg.): Computerized adaptive testing: Theory and practice. Boston: Kluwer, S. 289–310.
Wainer, H. (2000): Computerized adaptive testing: A primer. Mahwah: Lawrence Erlbaum Associates.
Wainer, H./Bradlow, E.T./Du, Z. (2000): Testlet response theory: An analog for the 3-PL useful in testlet-based adaptive testing. In: van der Linden, W.J./Glas, C.A.W. (Hrsg.): Computerized adaptive testing: Theory and practice. Boston: Kluwer, S. 245–270.
Wainer, H./Bradlow, E.T./Wang, X. (Hrsg.) (2007): Testlet Response Theory and Its Applications. New York: Cambridge University Press.
Wang, W.C./Chen, P.H. (2004): Implementation and measurement efficiency of multidimensional computerized adaptive testing. In: Applied Psychological Measurement 28, S. 450–480.
Wang, W.C./Wilson, M. (2005): The Rasch testlet model. In: Applied Psychological Measurement 29, S. 126–149.

Anschrift der Autoren

Dr. Andreas Frey, Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN) an der Universität Kiel, Olshausenstraße 62, D-24098 Kiel
E-Mail: frey@ipn.uni-kiel.de

Dipl.-Stat. Nicki-Nils Seitz, Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN) an der Universität Kiel, Olshausenstraße 62, D-24098 Kiel
E-Mail: seitz@ipn.uni-kiel.de

Nina Zeuch/Hanneke Geerlings/Heinz Holling/Wim J. van der Linden/ Jonas P. Bertling

Regelgeleitete Konstruktion von statistischen Textaufgaben

Anwendung von linear logistischen Testmodellen und Aufgabencloning

Projekt Regelgeleitete Itementwicklung[1]

1. Einleitung

Neuere Trends in der Bildungsforschung erfordern effektive Kompetenztestung für Diagnose- und Testentwicklungszwecke. Groß angelegte Untersuchungen im Rahmen der Bildungsforschung (wie das Programme for International Student Assessment, PISA oder die Third International Mathematics and Science Study, TIMSS) liefern Ergebnisse und Hinweise für den nationalen und internationalen Vergleich der Lernergebnisse und auch für die Sicherstellung und Verbesserung der Qualität der Lehre. Mit dem steigenden, breit angelegten Gebrauch von bildungsorientierten Tests wächst auch die Notwendigkeit einer effizienteren Testentwicklung und -durchführung. Die Tests sollen möglichst kurz sein und ein Maximum an Informationen über die Kompetenzen der Testperson liefern.

Neuere Entwicklungen im Bereich der Testtechnologie umfassen Versuche der Automatisierung (auf Grundlage theoretischer kognitiver Anforderungen und technischer Formatvorlagen) der Konstruktion von Testaufgaben unter Einbeziehung der Item Response Theorie (IRT) zur Beschreibung der kognitiven Anforderungen der Testaufgaben, der Anwendung computergestützter adaptiver Testung und der Optimierung von Stichproben-Designs. Die Automatisierung der Testaufgabenkonstruktion kann dabei über theoriebasierte Anforderungsprofile und flexible Formatvorlagen erfolgen, die eine Erstellung der Aufgaben durch Computerprogramme ohne Einzelkalibrierung der so konstruierten Aufgaben ermöglichen. Das Grundprinzip des adaptiven Testens ist eine Zuschneidung auf die individuellen Bedürfnisse und Fähigkeiten einer Testperson. Dies erfordert eine fortlaufend aktualisierte Fähigkeitsschätzung während der Testdurchführung sowie große Aufgabenpools oder automatische Aufgabengenerierung (automatic item generation, AIG) und wird im eigentlichen Sinne erst durch computergestütztes Testen ermöglicht (computergestütztes adaptives Testen, CAT, vgl. van der Linden 2003).

Das hier beschriebene Projekt soll diese Entwicklungen aufgreifen, vertiefen und die Ergebnisse integrieren, um ihre Nützlichkeit für die Entwicklung und den Einsatz von

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: HO 1286/5-1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

Statistik-Textaufgaben bei OberstufenschülerInnen und UniversitätsstudentInnen zu demonstrieren. Dabei bieten Textaufgaben hervorragende Eigenschaften durch die Verbindung mathematischer Formalismen mit alltäglichen Erfahrungen und die Bedeutung für nahezu jeden Bereich wissenschaftlicher Arbeit sowie für Eingangstests z.B. an Universitäten.

Das übergeordnete Ziel des Projektes ist die Konstruktion eines Softwaresystems, das automatisch eine einzigartige Menge an Aufgaben für alle Proband/innen produzieren und präsentieren kann. Jede folgende Aufgabe sollte an die momentane Fähigkeitsschätzung der Probandin/des Probanden angepasst sein (adaptive Vorgehensweise), weshalb das System in der Lage sein sollte, die Antworten der Probandin/des Probanden zu bewerten und ihre/seine Fähigkeitsschätzung in Echtzeit zu berechnen. Um einzigartige Aufgaben für jede Probandin/jeden Probanden zu generieren, wurden Generierungsregeln für verschiedene Arten von Statistik-Textaufgaben (z.B. Aufgaben zur Wahrscheinlichkeitsrechnung oder zu Konfidenzintervallen) definiert.

In diesem Beitrag wird zunächst ein Überblick über die theoretischen Hintergründe des Projektes in Abschnitt 2 gegeben. Anschließend werden verschiedene IRT-Modelle beschrieben, die zur Bestimmung der Effekte der Generierungsregeln auf die Aufgabenschwierigkeiten herangezogen werden können. Eines dieser Modelle, das linear logistische Testmodell (LLTM), wurde in einer Reihe empirischer Studien eingesetzt, von denen jede zu einer Verfeinerung der aufgabengenerierenden Regeln führte. Beispielhaft werden dazu die Ergebnisse einer Studie zu Konfidenzintervall-Aufgaben in Abschnitt 4.2 beschrieben.

Eine Auswahl der Generierungsregeln für die Wahrscheinlichkeitsaufgaben basiert auf vorherigen Untersuchungen (vgl. Holling/Bertling/Zeuch 2009) und wurde für die Entwicklung eines ersten Prototyps eines automatischen Aufgabengenerators für diesen Aufgabentyp verwendet. Der Artikel schließt mit der Diskussion und einem Ausblick auf künftige Erweiterungen und Anwendungen ab.

2. Theoretischer Hintergrund

Im Folgenden wird die theoretische Grundlage für das vorliegende Projekt dargestellt. Zuerst werden Statistik-Textaufgaben mit wichtigen Merkmalen und Anwendungsbeispielen behandelt. Danach werden automatische Aufgabengenerierung, regelgeleitete Aufgabenkonstruktion und Aufgabencloning erläutert.

2.1 Statistik-Textaufgaben

Statistische Kompetenzen sind ein wichtiger Teil allgemeiner mathematischer Kompetenzen und spielen in nahezu allen wissenschaftlichen Bereichen eine große Rolle. Fundierte Statistikkenntnisse bereiten die Grundlage für wissenschaftliches Arbeiten an der Universität und auch für den Umgang mit Statistik im alltäglichen Leben.

Statistische Kompetenzen können mit diversen Aufgabentypen gemessen werden. Allerdings bieten Textaufgaben mit ihrer Informationsfülle über ein tieferes Verständnis entscheidender statistischer Konzepte über die Beherrschung und den Transfer dieser Kompetenzen über Formeln und Gleichungen hinaus einen idealen Aufgabentyp. Außerdem zeigen Textaufgaben im mathematischen Bereich eine hohe externe Validität, da sie gleichzeitig kreative, logische und mathematische Kompetenzen messen (vgl. z.B. Jonassen 2003). Dies stellt einen wichtigen Anlass dar, sich intensiv wissenschaftlich mit Statistik-Textaufgaben auseinanderzusetzen und damit Statistik einen höheren Stellenwert einzuräumen, ganz im Einklang mit einer der Hauptforderungen der OECD, „Unsicherheit" (eine der vier Subdimensionen mathematischer Kompetenzen bei PISA; vgl. OECD 2003) eine wichtigere Rolle im Bildungsbereich zu verschaffen.

Statistik-Textaufgaben sind eine Unterklasse von mathematischen Textaufgaben. Viele Erkenntnisse über mathematische und vor allem Algebra-Textaufgaben lassen sich auf Statistik-Textaufgaben übertragen, aber letztere sollten als eigenständiger Aufgabentyp betrachtet werden, da verschiedene Unterklassen von mathematischen Textaufgaben qualitativ unterschiedlich sein und deshalb nicht auf einer gemeinsamen konzeptuellen Dimension beschrieben werden können (Arendasy u.a. 2006).

Die Rückführung der Aufgabenschwierigkeit auf bestimmte Konstruktionsregeln und weitere Aufgaben- und Testcharakteristika (Schwierigkeitsmodellierung) für mathematische Textaufgaben wird zunehmend aufwändiger und ambitionierter und umfasst auch Konstruktvalidierung, Aufgabengenerierung und -klassifikation (vgl. Enright/Sheehan 2002). Arendasy u.a. (2006) betrachteten verschiedene Typen von Textaufgaben und entwickelten den Aufgabengenerator Agen, der Vorlagen für die Generierung von Isomorphen (Aufgabenvariationen mit gleicher grundlegender Struktur) nutzt. Dieses Vorgehen ist angelehnt an die Aufgabenproduktion auf der generelleren Basis von Radicals (systematischer Einfluss auf die Aufgabenschwierigkeit) und Incidentals (Oberflächenmerkmale ohne Einfluss auf Aufgabenschwierigkeit). So ist bei einer mathematischen Textaufgabe die zur Aufgabenlösung erforderliche Formel und deren Berechnung als Radical zu betrachten, da sie die Schwierigkeit der Aufgabe beeinflusst; die Rahmengeschichte, in die die Aufgabe eingebettet ist, ist jedoch als Incidental anzusehen, da die Aufgabenschwierigkeit hiervon nicht abhängen dürfte.

Nur sehr wenige Forschungsgruppen beschäftigen sich mit Statistik-Textaufgaben (vgl. z.B. Arendasy u.a. 2006). Regelgeleitete Konstruktion von Statistik-Textaufgaben kann ihren Einsatz in Lehre und Kompetenzmessung unter anderem durch die Möglichkeit von AIG und CAT erleichtern und flexibler gestalten. Hierfür sind vor allem IRT-Modelle wie das LLTM relevant. Derzeit sind uns allerdings keine Ansätze von regelgeleiteter Aufgabenkonstruktion oder Aufgabencloning im Bereich von Statistik-Textaufgaben bekannt.

2.2 Automatische Aufgabengenerierung, regelgeleitete Aufgabenkonstruktion und Aufgabencloning

AIG auf der Basis von theoretisch und empirisch validierten Qualitätskontrollmechanismen dient der Qualitätsverbesserung von Testungen und ermöglicht Aufgabenproduktion unter Minimierung von Fehlerquellen (z.B. uneinheitliche Gestaltung, Tippfehler) und Maximierung der Effizienz, da theoretisch unendliche Aufgabenmengen generiert werden können, sobald das System fertiggestellt ist. Auch wird die Interpretation von Testergebnissen vereinfacht (vgl. Arendasy u.a. 2006). Wenn die bestimmenden Merkmale und Konstruktionsregeln bekannt sind, kann AIG für die Produktion einer großen Anzahl qualitativ hochwertiger Aufgaben genutzt werden. Derzeitige Bemühungen zur AIG lassen sich in die beiden Ansätze der regelgeleiteten Aufgabenkonstruktion und des Aufgabencloning einteilen.

Bei der regelgeleiteten Aufgabenkonstruktion werden die Aufgaben eines Inhaltsbereiches hinsichtlich ihrer kognitiven Anforderungen und schwierigkeitsbestimmenden Merkmale untersucht. Daraus werden Regeln abgeleitet, die diese Strukturen bestimmen (Radicals). Diese Regeln werden in Computeralgorithmen implementiert, welche große Mengen an neuen Aufgaben auf dieser Basis generieren können.

Beim Aufgabencloning wird eine Menge von typischen Aufgaben des Inhaltsbereiches betrachtet, die dann die „Elternaufgaben“ darstellen, aus denen große Familien von „Geschwisteraufgaben“ geklont werden. Normalerweise besteht das Klonen aus der Anwendung von Computeralgorithmen, die unwesentliche Merkmale der Aufgaben (Incidentals) verändern. Die beiden Ansätze sind im Überblick bei Bejar (1993), sowie Irvine und Kyllonen (2002) dargestellt.

3. Statistische Modellierung

Für die Analyse von Daten, die regelgeleitet konstruiert oder durch Aufgabencloning erstellt wurden, wurden verschiedene Modelle entwickelt. In Abschnitt 3.1 werden Modelle beschrieben, die die Generierungsregeln als erklärende Faktoren für die Aufgabenschwierigkeit einbeziehen; in Abschnitt 3.2 wird dargestellt, wie die hierarchische Struktur von Aufgabenpools aus Aufgabencloning in einem Modell berücksichtigt werden kann.

3.1 IRT-Modellierung und das LLTM

Das LLTM gehört zu den IRT-Testmodellen und basiert auf dem Rasch-Modell (RM; vgl. Rasch 1960). Die Grundstruktur dieser Modelle besteht aus einer parametrischen bi- oder multinomialen Verteilung von Antworten von Testpersonen mit Parametern für die Effekte der Testpersonen und der Aufgaben auf die Antwortwahrscheinlichkeiten. Unter anderem können IRT-Modelle auch zur Testung von Hypothesen über mögliche

Problemstrukturen in den Aufgaben und zur Analyse von Antwortdaten von komplexeren Testformen wie adaptive Tests herangezogen werden. Für eine Testperson *j* mit der Fähigkeit θ_j und eine Aufgabe *i* mit der Schwierigkeit σ_j definiert das RM die Wahrscheinlichkeit einer korrekten Antwort ($X_{ij} = 1$) durch:

$$P\left(X_{ij}=1\middle|\theta_j,\sigma_i\right)=\frac{\exp\left(\theta_j-\sigma_i\right)}{1+\exp\left(\theta_j-\sigma_i\right)} \tag{1}$$

Das LLTM zerlegt die Aufgabenschwierigkeit σ_i in $k = 1, \ldots, K$ Basisparameter η_k mit den Gewichten q_{ik} (vgl. Fischer/Molenaar 1995):

$$P\left(X_{ij}=1\middle|\theta_j,q_i,\eta\right)=\frac{\exp\left(\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k\right)}{1+\exp\left(\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k\right)} \tag{2}$$

Die Basisparameter (also die Effekte der Radicals auf die Aufgabenschwierigkeit) werden meistens aufgrund von theoretischen Vorüberlegungen spezifiziert und in einer sogenannten *Q*-Matrix festgehalten, in deren Zeilen die einzelnen Aufgaben und in deren Spalten die Basisparameter stehen. Jeder Aufgabe wird so eine entsprechende Anzahl und Kombination an Basisparametern zugewiesen. Eine Eins in einer Zelle zeigt an, dass in der entsprechenden Aufgabe ein Basisparameter enthalten ist, eine Null, dass der entsprechende Basisparameter nicht enthalten ist. Entweder können vorhandene Aufgaben auf diese Weise klassifiziert werden oder die Aufgaben können nach einer a priori definierten *Q*-Matrix konstruiert werden. Dieses Vorgehen ist auch für die regelgeleitete Aufgabenkonstruktion unerlässlich, bei der Aufgaben entsprechend vorbestimmter kognitiver Strukturen erstellt werden.

Da das LLTM die sehr strikte und oft in der Realität nicht zutreffende Annahme beinhaltet, dass die *Q*-Matrix eine erschöpfende Erklärung für die Aufgabenschwierigkeiten liefert, kann ein zufälliger Fehlerterm in das LLTM einbezogen werden, mit dem Varianzanteile modelliert werden, die nicht durch die spezifizierten Basisparameter erklärt werden (vgl. Janssen/Schepers/Peres 2004). Dieses Modell wird auch Random-Effects LLTM (RE-LLTM) genannt:

$$P\left(X_{ij}=1\middle|\theta_j,q_i,\eta\right)=\frac{\exp\left(\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k+\varepsilon_i\right)}{1+\exp\left(\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k+\varepsilon_i\right)} \tag{3}$$

LLTMs liefern Schätzungen für die vordefinierten Basisparameter, die anzeigen, ob und wenn ja, welche Parameter in welchem Ausmaß einen signifikanten Einfluss auf die Aufgabenschwierigkeiten haben.

Aufgrund der dargestellten Modelleigenschaften eignen sich die LLTMs hervorragend zur Analyse regelgeleitet konstruierter Aufgaben sowie zur Hypothesentestung bezüglich der vermuteten und durch die *Q*-Matrix definierten kognitiven Basisparameter.

3.2 Aufgabencloning-Modelle

Beim herkömmlichen Vorgehen werden die Aufgabenparameter, basierend auf einer Kalibrierung anhand einer Stichprobe von ausreichender Größe, damit der Schätzfehler ignoriert werden kann, auf ihre geschätzten Werte fixiert. Da beim Aufgabencloning die Aufgaben normalerweise in Familien gruppiert sind, die aus der gleichen zugrundeliegenden Struktur abgeleitet sind, scheint ein zufälligkeitsbasierter Ansatz für die Aufgabenparameter eher angemessen. Glas und van der Linden (2003) schlagen ein hierarchisches IRT-Modell für dichotome Aufgaben vor, das diese Gruppierung berücksichtigt. Das Modell behandelt alle Aufgabenparameter im 3-Parameter-Logistischen (3PL)-Modell als zufällig unter Voraussetzung ihrer Familienstruktur. Seien $j = 1, \ldots, J$ die Personen, $f = 1, \ldots, F$ die Aufgabenfamilien, und $i_f = 1, \ldots, I_f$ die Aufgaben der Familie f, so kann das Modell der ersten Ebene als

$$p(X_{i_f j} = 1 \mid \theta_j, a_{i_f}, \sigma_{i_f}, c_{i_f}) = c_{i_f} + (1 - c_{i_f}) \frac{\exp[a_{i_f}(\theta_j - \sigma_{i_f})]}{1 + \exp[a_{i_f}(\theta_j - \sigma_{i_f})]} \tag{4}$$

definiert werden, wobei θ_j, α_{i_f}, σ_{i_f}, c_{i_f} die Fähigkeits-, Diskriminations-, Schwierigkeits-, und Rateparameter sind. Die Aufgabenparameter einer Familie f, als Einheit mit ξ_{i_f} bezeichnet, werden transformiert, sodass ihre Verteilung ausreichend nah an der multivariaten Normalverteilung liegt:

$$\xi_{i_f} \sim MVN(\mu_f, \Sigma_f), \tag{5}$$

mit μ_f und Σ_f als Familienparameter (zweite Ebene).

Als Teil dieses Projektes wurde das Modell erweitert, um eine Erklärung der Aufgabenschwierigkeiten durch die Effekte der angewandten Generierungsregeln zu ermöglichen. Das neue Modell kann somit als Kombination des LLTM mit dem Modell von Glas und van der Linden (2003) angesehen werden. Als Modell der ersten Ebene wurde ein 3-Parameter-Normal-Ogiven (3PNO)-Modell (das bis auf eine Skalierungskonstante annähernd gleich dem 3PL-Modell bei Glas und van der Linden (ebd.) ist) verwendet.

Im neuen Modell wird der Familienschwierigkeitsparameter als eine Kombination aus den Effekten der Radicals η_k angenommen:

$$\mu_{b_f} = \sum_{k=1}^{K} q_{fk} \eta_k \tag{6}$$

Die Variable q_{fk} gibt an, ob Radical k für Familie f benötigt wird. Die Parameter des Modells können in einem Bayesischen Rahmen durch einen Gibbs Sampler geschätzt werden (vgl. Geerlings/van der Linden/Glas eingereicht).

Wenn die Familienparameter mit einem der Aufgabencloning-Modelle geschätzt worden sind, muss prinzipiell eine neu generierte Aufgabe mit bekannter Familienzugehörigkeit nicht mehr kalibriert werden, sondern die bekannten Familienparameter können für die Berechnung der Probandenfähigkeit verwendet werden. Die Genauigkeit der resultierenden Fähigkeitsschätzungen hängt von der Varianz der Aufgabenparameter innerhalb der Familien (Σ_f) ab. Eine erfolgreiche Anwendung des Modells sollte idealerweise in einer großen Varianz der Aufgabenparameter zwischen den und einer geringen Varianz innerhalb der Familien resultieren. Für jeden Probanden/jede Probandin kann dann eine zufällige Aufgabe aus einer Familie (also einer Kombination von Radicals) generiert werden, die optimal bezüglich der aktuellen Fähigkeitsschätzung ist. Gleichzeitig wird der so konstruierte Test durch die Variation der Incidentals aber jedes Mal anders aussehen, was Wiedererkennungseffekte verhindert.

4. Erste Ergebnisse

Zunächst wird die Konstruktion der statistischen Textaufgaben im Rahmen des Projektes dargestellt und es werden erste empirische Ergebnisse beispielhaft aufgezeigt. Daran schließt sich die Darstellung eines Prototyps für einen automatischen Aufgabengenerator an.

4.1 Aufgabentypen und Designprinzipien

Anhand der Lehrpläne für Statistikinhalte im Unterricht der gymnasialen Oberstufe sowie in den Lehrveranstaltungen an Universitäten wurden wichtige basale Operationen der Statistik, die typischerweise Einfluss auf die Lösungswahrscheinlichkeit von Statistik-Textaufgaben haben sollten, identifiziert und als Basisparameter in mehreren Q-Matrizen zur Konstruktion mehrerer Itemmengen definiert. Die Aufgaben wurden halbautomatisch mit Hilfe von LaTex2e-Vorlagen generiert, die durch eine weitestgehend identische Wortwahl und Satzstruktur Missinterpretationen und zusätzliche Fehlervarianz durch unterschiedliches Textverständnis und Satzbaueffekte vermeiden.

4.2 Empirische Studien

Teilmengen der wie oben beschrieben konstruierten Aufgaben wurden in mehreren empirischen Studien mit insgesamt 1274 deutschen OberstufenschülerInnen und PsychologiestudentInnen getestet. Die ersten eingesetzten Aufgaben berücksichtigten vor allem Operationen der grundlegenden Wahrscheinlichkeitstheorie, z.B. den Umgang mit ab-

hängigen oder unabhängigen Wahrscheinlichkeiten. Diese Ergebnisse wurden bereits teilweise veröffentlicht und können im Detail z.B. bei Holling, Bertling und Zeuch (2009) eingesehen werden. Die daraufhin entwickelten Aufgaben widmen sich einem etwas breiteren Inhaltsspektrum (u.a. Varianzanalyse und Konfidenzintervalle). Dabei wurden auch mehrere Testaufgaben zur grundlegenden Wahrscheinlichkeitstheorie nach einem Aufgabencloning-Modell erstellt, diese befinden sich allerdings noch in der Kalibrierungsphase. Beispielhaft soll hier ein Subset von regelgeleitet konstruierten Aufgaben zu Konfidenzintervallen (KI) dargestellt werden, die sich gerade in der Pilotierung befinden. Die einzelnen Aufgaben bestehen aus einer kurzen Rahmengeschichte mit einer daran anschließenden Aufforderung, ein KI oder einen für ein KI benötigten Wert aus einem gegebenen KI zu berechnen.

Der Test besteht aus acht Aufgaben, die unterschiedliche Kombinationen der berücksichtigten Basisparameter VAR (KI für eine Varianz), ANT (KI für einen Anteil), EIN/ZWEI (einseitiges oder zweiseitiges KI), und INV (Inversion der Formel) beinhalten. Wenn die Q-Matrix-Einträge für VAR und ANT Null sind, handelt es sich um ein KI für einen Mittelwert. Abbildung 1 zeigt eine Beispielaufgabe, die die Berechnung eines KIs für einen Anteilswert beinhaltet.

In einer Therapiestudie wird der Frage nachgegangen, ob ein neuartiges Therapieprogramm effektiv im Sinne der Befindlichkeitsverbesserung der Patienten ist. 72 der insgesamt 120 Teilnehmer berichten eine deutliche Besserung. Die Klinik verspricht aber, dass mit diesem Programm mindestens 50 Prozent der Patienten eine Verbesserung der Befindlichkeit erfahren.
Berechnen Sie ein Konfidenzintervall für die Ergebnisse der Studie, das die Klinikleitung zur Überprüfung ihres Versprechens heranziehen könnte, wenn eine Sicherheit von 90 Prozent berücksichtigt werden soll.

Abb. 1: Beispielaufgabe Konfidenzintervall-Test

Aufgabe	VAR	ANT	EIN/ZWEI	INV
1	0	0	0	1
2	0	1	0	0
3	1	0	0	1
4	0	0	1	0
5	0	1	1	0
6	1	0	1	0
7	0	0	1	1
8	0	1	1	1

Anmerkungen: VAR = „Varianz“, ANT = „Anteilswert“, EIN/ZWEI = „ein- oder zweiseitig“, INV = „Inversion“.

Tab. 1: Q-Matrix für den Konfidenzintervall-Test

Tabelle 1 zeigt die *Q*-Matrix für die acht Aufgaben (die Beispielaufgabe aus Abbildung 1 entspricht Aufgabe 2 in der Designmatrix).

Die Aufgaben wurden 86 PsychologiestudentInnen der ersten beiden Fachsemester an der Westfälischen Wilhelms-Universität Münster vorgelegt. Durchschnittlich wurden 3,59 (45 Prozent) der acht Aufgaben korrekt beantwortet. Die Aufgabenschwierigkeiten bewegen sich zwischen 0,28 für Aufgabe 8 und 0,69 für Aufgabe 4. Die interne Konsistenz ist mit einem Cronbachs Alpha von 0,48 sehr gering. Der *Q*-Index weist mit Werten zwischen 0,14 für Aufgabe 7 und 0,21 für Aufgabe 5 einen guten Rasch-Modellfit für alle Aufgaben auf (vgl. Rost/von Davier 1994). Tabelle 2 zeigt die sehr ähnlich ausfallenden LLTM- und RE-LLTM-Schätzungen.

		LLTM		**RE-LLTM**	
	Parameter	**Schätzung**	**SE**	**Schätzung**	**SE**
	Konstante	0.36	0.25	0.37	0.35
Feste Effekte	VAR	−0.20	0.22	−0.20	0.31
	ANT	−0.63**	0.20	−0.64*	0.29
	EIN/ZWEI	0.11	0.18	0.11	0.26
	INV	−0.76**	0.18	−0.77**	0.26
Zufällige Effekte	θ_j	0.51	0.19	0.49	0.19
	ε_i	–	–	0.06	0.06

Anmerkungen: $*p < .05$. $**p < .01$. SE = Standardfehler. VAR = „Varianz“, ANT = „Anteilswert“, EIN/ZWEI = „ein- oder zweiseitig“, INV = „Inversion“.

Tab. 2: Parameterschätzungen für LLTM und RE-LLTM im Konfidenzintervall-Test

Zwei der vier Basisparameter (ANT und INV) sind sowohl im LLTM als auch im RE-LLTM statistisch signifikant und haben somit einen inkrementellen Einfluss auf die globale Aufgabenschwierigkeit. Ein Likelihood-Ratio-Test konnte keinen Vorteil des RE-LLTM gegenüber dem LLTM nachweisen.

Die Korrelation zwischen LLTM- und Rasch-Aufgabenparametern beträgt 0,79. Daraus ergibt sich eine gute Varianzaufklärung von $R^2 = 0{,}63$.

Diese Resultate des KI-Tests sind, wahrscheinlich aufgrund der geringen Testlänge und Stichprobengröße, nicht einwandfrei. Die Tendenz ist dennoch vielversprechend, konnten doch zwei statistisch signifikante Basisparameter identifiziert werden. Diese Aufgabenform soll weiterentwickelt und an größeren Stichproben getestet werden.

4.3 Automatischer Aufgabengenerator

Ein automatischer Aufgabengenerator wird für Statistik-Textaufgaben entwickelt, in dem Operationen der grundlegenden Wahrscheinlichkeitstheorie geprüft werden. Die Aufgaben ähneln denen, die in Holling, Bertling und Zeuch (2009) dargestellt werden.

Jede Statistik-Textaufgabe besteht aus einer Rahmengeschichte, die die relevanten numerischen Informationen für die Antwortberechnung enthält und einer Frage, die die Berechnung einer bedingten Wahrscheinlichkeit („ua“), eines Komplementärereignisses („nicht“), einer Wahrscheinlichkeit für eine Schnittmenge („uu“) oder einer Wahrscheinlichkeit für eine Verbundmenge („oder“) erfordert. Die Struktur der Kontextgeschichten ist für jede Frage gleich. Die einzige Kontextvariation wird durch die Incidentals verursacht, die Informationen über Subjekt und Objekt der Geschichte und die Interpretation der verwendeten Variablen liefern. Eine Frage, die beispielsweise die Berechnung einer Gegen-, bedingten und Verbundmengen-Wahrscheinlichkeit erfordert, kann durch die Anwendung einiger weniger Aussagen (siehe Abbildung 2) generiert werden. Eine Sammlung von Subjekten, Objekten und Variablen wird zur Erzeugung der Oberflächenunterschiede zwischen den Aufgaben verwendet.

“Wie groß ist die Wahrscheinlichkeit, dass”	Wie groß ist die Wahrscheinlichkeit, dass
<*Subjekt-Artikel*> <*Subjekt*> „ein/eine/einen“	der Buchhändler ein
<*Objekt Singular*> „hat“, <*Relativpronomen*>	Buch hat, das
if(not=1) “nicht”	nicht
if(uu=1 & oder!=1) “sowohl”	
if(oder=1) „entweder“	entweder
„ein/eine/ein“ <*Merkmalsausprägung 1*> <*Variable 1*> <*Verb 1*>	einen gelben Umschlag hat
if(oder=1) „oder“	oder
if(oder=1 & uu=1) „sowohl“	
„ein/eine/ein“ <*Merkmalsausprägung 1*> <*Variable 1*> <*Verb 1*>	einen grünen Umschlag hat,
if(ua=1) „ , vorausgesetzt, ‘‘ <*Artikel*>	vorausgesetzt, das
<*Objekt Singular*> <*Verb 3*> „ein/eine“einen“	Buch hat eine
<*Merkmalsausprägung 3*> <*Variable 3*>	männliche Hauptperson
if(uu=1) „als auch ein/eine/ein“	
<*Merkmalsausprägung 2*> <*Variable 2*>	
<*Verb 2*>	

Abb. 2: Vereinfachte Struktur der Fragen und eine Beispielfrage

5. Diskussion und Ausblick

Im vorliegenden Projekt werden Inhalte der kognitiven Psychologie, Psychometrie und Computerwissenschaften kombiniert, um ein Testsystem für Statistik-Textaufgaben zu entwickeln, das einen adaptiven, einzigartigen Test für jeden Probanden/jede Probandin erschaffen kann.

Die ersten Projektergebnisse sind durchweg vielversprechend. So konnten verschiedenste kognitive Komponenten identifiziert und in halbautomatischer Aufgabenkonstruktion als Vorstufe zur vollautomatischen Generierung mit Hilfe von Textbausteinen umgesetzt werden. Beispielhaft wurde ein Test zu Konfidenzintervallen dargestellt. Diese Ergebnisse und die aller weiteren empirischen Untersuchungen zeigen die Anwendbarkeit von regelgeleiteter und automatischer Aufgabengenerierung auf Textaufgaben mit Statistikinhalten. Die regelgeleitete Konstruktion wurde in mehreren Aufgabenmengen umgesetzt und empirisch überprüft. Die Ergebnisse dienen nun zur Verfeinerung des automatischen Aufgabengenerators. Es konnten in LLTM-Analysen verschiedene signifikante schwierigkeitsgenerierende Merkmale identifiziert werden. Außerdem zeigen die konstruierten Aufgaben einen guten RM-Fit und stellen den Ausgangspunkt für verfeinerte inhaltliche und konstruktionstechnische Weiterentwicklungen und adaptive Implementierungen dar. Verschiedene mögliche Modelle für die Kalibrierung der automatisch generierten Aufgaben wurden aufgezeigt. Eine wichtige Frage ist, ob bessere Modellpassung die höhere Parameteranzahl im komplexeren hierarchischen IRT-Modell rechtfertigt. Es ist eine Studie geplant, in der die Passung aller hier erwähnten Modelle verglichen wird: Das LLTM, das Aufgabencloning-Modell von Glas und van der Linden (2003) sowie die erweiterte Version dieses Modells (beschrieben in Abschnitt 3.2). Die Modelle sollen mit Hilfe der gewonnenen Daten aus den empirischen Erhebungen zu Statistik-Textaufgaben aus 4.3 verglichen werden.

Literatur

Arendasy, M./Sommer, M./Gittler, G./Hergovich, A. (2006): Automatic generation of quantitative reasoning items: A pilot study. In: Journal of Individual Differences 27, S. 2–14.

Bejar, I. (1993): A generative approach to psychological and educational measurement. In: Frederiksen, N./Mislevy, R.J./Bejar, I.I. (Hrsg.): Testtheory for a new generation of tests. Hillsdale, NJ: Lawrence Erlbaum Associates, S. 323–357.

Enright, M.K./Sheehan, K.M. (2002): Modeling the difficulty of quantitative reasoning items: Implications for item generation. In: Irvine, S.H./Kyllonen, P.C. (Hrsg.): Item generation for test development. Mahwah, NJ: Erlbaum, S. 129–157.

Fischer, G.H./Molenaar, I.W. (Hrsg.) (1995): Rasch Models. Foundations, Recent Developments, and Applications. New York: Springer.

Geerlings, H./van der Linden, W.J./Glas, C.A.W. (eingereicht): Modeling rule-based item generation.

Glas, C.A.W./van der Linden, W.J. (2003): Computerized adaptive testing with item cloning. In: Applied Psychological Measurement 27, S. 247–261.

Holling, H./Bertling, J.P./Zeuch, N. (2009): Probability word problems: Automatic item generation and LLTM modelling. In: Studies in Educational Evaluation 35, S. 71–76.

Irvine, S.H./Kyllonen, P.C. (2002): Item generation for test development. Mahwah, NJ: Erlbaum.

Janssen, R./Schepers, J./Peres, D. (2004): Models with item and item group predictors. In: De Boeck, P./Wilson, M. (Hrsg.): Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. New York: Springer, S. 189–210.

Jonassen, D.H. (2003): Designing research-based instruction for story problems. In: Educational Psychology Review 15, S. 267–296.

OECD (Hrsg.) (2003): The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills. Paris: OECD.
Rasch, G. (1960): Probabilistic Models for some Intelligence and Attainment tests. Copenhagen: Pædagogiske Institut.
Rost, J./von Davier, M. (1994): A conditional item fit index for Rasch models. In: Applied Psychological Measurement 18, S. 171–182.
Van der Linden, W. (2003): Some new developments in adaptive testing technology. In: Journal of Psychology 216, S. 3–11.

Anschrift der Autor/innen

Dipl.-Psych. Nina Zeuch, Westfälische Wilhelms-Universität Münster,
Lehrstuhl für Statistik und Methoden, Fliednerstraße 21, D-48149 Münster
E-Mail: n_hoff01@uni-muenster.de

Prof. Dr. Heinz Holling, Westfälische Wilhelms-Universität Münster,
Lehrstuhl für Statistik und Methoden, Fliednerstraße 21, D-48149 Münster
E-Mail: holling@uni-muenster.de

Dipl.-Psych. Jonas P. Bertling, Westfälische Wilhelms-Universität Münster,
Lehrstuhl für Statistik und Methoden, Fliednerstraße 21, D-48149 Münster
E-Mail: jonas.bertling@uni-muenster.de

MSSc. Hanneke Geerlings, University of Twente, Department of Research Methodology,
Measurement and Data Analysis, Faculty of Behavioral Sciences, P.O. Box 217,
7500 AE Enschede, The Netherlands
E-Mail: h.geerlings@gw.utwente.nl

Prof. Dr. Wim J. van der Linden, University of Twente, Department of Research Methodology,
Measurement and Data Analysis, Faculty of Behavioral Sciences, P.O. Box 217,
7500 AE Enschede, The Netherlands
E-Mail: w.j.vanderlinden@utwente.nl

Eckhard Klieme/Anika Bürgermeister/Birgit Harks/Werner Blum/Dominik Leiß/ Katrin Rakoczy

Leistungsbeurteilung und Kompetenzmodellierung im Mathematikunterricht

Projekt Co²CA[1]

Einleitung

Das Forschungsprojekt „Conditions and Consequences of Classroom Assessment (Co²CA)“[2] untersucht, in welchem Verhältnis Leistungsmessung, -bewertung und -beurteilung[3] zum Unterricht und zum Lernprozess der Schüler[4] stehen. Die alltägliche Praxis der Leistungsbeurteilung – von der informellen Fehlerdiagnose im Unterrichtsgespräch bis zu Kriterien und Verfahrensweisen der Notengebung – spiegelt einerseits die Ziele und Prozessqualitäten des Unterrichts, andererseits wirkt sie sich ihrerseits auf das Unterrichtsgeschehen sowie die kognitive und motivationale Entwicklung der Lernenden aus. In diese Praxis greifen neuerdings extern entwickelte Tests und Vergleichsarbeiten ein, die sich auf Bildungsstandards beziehen. In diesem Kontext untersucht das Projekt, wie verschiedene Formen des Assessments[5] genutzt werden und welche Wirkung sie entfalten. Es konzentriert sich dabei auf den Mathematikunterricht in mittleren Bildungsgängen der neunten Jahrgangsstufe.

1. Aktuelle Forschungen zur Leistungsbeurteilung im Unterricht

1.1 Formative Leistungsbeurteilung

In der angelsächsischen Fachsprache sind mit „formative assessment“ alle Formen von Leistungsbeurteilung gemeint, die Informationen über die Diskrepanz zwischen Lernzie-

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: KL 1057/10) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

2 2007–2009 gefördert unter KL 1057/10; seit Herbst 2009 gemeinsam geleitet von E. Klieme, K. Rakoczy, W. Blum und D. Leiß.

3 Wir betrachten hier „Leistungsbeurteilung“ (englisch: Assessment) als einen Vorgang, der sowohl die Feststellung von Leistungen – bei quantitativen Verfahren mit nachweisbarer Güte als Messung bezeichnet – als auch deren normative Bewertung umschließt.

4 Aus Gründen der Lesbarkeit wird in diesem Beitrag nur die männliche Form verwendet, auch wenn beide Geschlechter gemeint sind.

5 Wir bezeichnen sowohl lernprozessbegleitende (formative) als auch bilanzierende (summative) Leistungsbeurteilung als Assessment – in Übereinstimmung mit der aktuellen angloamerikanischen Literatur, aber anders als im Rahmenantrag des DFG-Schwerpunktprogramms von 2005, wo dieser Begriff auf summative Erhebungen eingeschränkt wurde.

len und aktuellem Lernstand liefern und dadurch den Lehrenden und/oder den Lernenden selbst helfen, den weiteren Lernprozess zu gestalten (vgl. Sadler 1989; Black/William 1998). Was dies genau bedeutet, wird durchaus kontrovers diskutiert. Während einige Autoren (vgl. z.B. Hattie/Timperley 2007) betonen, dass jedes Testverfahren – auch ein breit angelegter standardisierter Multiple Choice-Test – formativ genutzt werden kann, bringen andere Autoren das Konzept in enge Verbindung mit spezifischen Unterrichtssituationen: diskursiven Sequenzen und Aufgabenstellungen, mit denen Wissen und Verständnis von Schülern offen gelegt werden können (vgl. Heritage 2007; Shavelson u.a. 2008).

Heritage und Shavelson unterscheiden dementsprechend verschiedene Arten von formativem Assessment: spontane Sequenzen („on the fly assessment"), geplante Frage-Antwort-Sequenzen, die diagnostische Informationen liefern („planned for interaction"), sowie stärker formalisiertes „curriculum-embedded assessment". Die beiden erstgenannten Varianten stellen diagnostische Situationen dar, in denen Leistungen von Schülern evoziert, beobachtet, interpretiert und durch die Lehrkraft kommentiert werden. Hier wird also adaptiv unterrichtet; eine explizite Messung und Bewertung von Schülerleistungen findet jedoch nicht statt.[6] Im Rahmen des DFG-Schwerpunktprogramms interessiert daher eher die stärker formalisierte Art der Leistungsmessung und -bewertung, also das „curriculum-embedded assessment". Auch hierfür finden sich in der einschlägigen Literatur und der Testpraxis in den USA ganz unterschiedliche Beispiele: Kommerzielle Testanbieter bieten Varianten ihrer summativen, standardbezogenen Verfahren als „benchmark tests" an, die beispielsweise quartalsweise den Lernfortschritt festhalten sollen. Die Tests der Projekte des Berkeley Evaluation and Assessment Research (BEAR) Center (vgl. Wilson 2008) und „Power Source" (vgl. Baker 2007) sind hingegen sehr unterrichtsnah angelegt. Während Wilson komplexe offene Aufgaben stellt, aus deren Beantwortung auf die Entwicklung des Verständnisses geschlossen werden soll, arbeiten Baker und Mitarbeiter mit Serien von Tests, die jeweils spezifische Zielbereiche prüfen. Shavelson und Mitautoren (2008) haben zunächst mit ähnlichen formalisierten Tests gearbeitet, mussten aber erfahren, dass die Lehrkräfte aufwändige Testserien nicht in ihren Unterricht integrieren konnten. Sie vermeiden inzwischen den „Assessment"-Begriff und sprechen stattdessen von „reflective lessons".

Einigkeit besteht aber darin, dass Durchführung, Interpretation und weiterführende Nutzung unterrichtsbezogener (formativer) Leistungsbeurteilung spezifische Lehrerkompetenzen erfordert und daher nicht ohne ein Lehrertraining eingeführt werden kann, das neben diagnostischen und instruktionalen Techniken auch ein vertieftes Verständnis fachlicher Lerninhalte und -prozesse erfordert. Dabei müsste auch die Notengebung (vgl. Brookhart 1993) als – zumindest im deutschen Schulsystem – wichtigste Art der formativen Leistungsbeurteilung berücksichtigen werden, die vermutlich von Wissen und Einstellungen der Lehrkräfte (vgl. Rakoczy u.a. 2008) sowie deren diagnostischer Kompetenz (vgl. Schrader 2006) beeinflusst ist und bestimmte Schülergruppen leicht verzerrt bewertet (vgl. Klieme 2003).

---

6 Ob es sich hier überhaupt um Assessment handelt oder „nur" um diskursiven Unterricht, wäre im Einzelfall daran zu prüfen, ob zumindest eine implizite Bewertung von Leistungen stattfindet.

1.2 Abgrenzung zur summativen Leistungsbeurteilung

In Abgrenzung zum formativen Assessment orientieren sich summative Leistungsbeurteilungen, die einen Bildungsabschnitt bilanzieren sollen, in den USA wie in Deutschland heutzutage an (Bildungs-)Standards. Standards geben landesweit die Kompetenzbereiche (Dimensionen), die Messeinheiten (Skalen und Stufen) und die „Sollwerte" (Minimal- oder Regelstandards) an, nach denen Schüler beurteilt werden.

Summatives und formatives Assessment können sich durchaus auf dieselben Kompetenzmodelle stützen und sollten es sogar, um die oft beschworene Passung („Alignment") zwischen Unterrichtsprozessen und Standards zu erreichen (vgl. Pellegrino/Chudowsky/Glaser 2001). Dennoch bleibt ein Spannungsverhältnis bestehen. Zum einen sind die Standards und die darauf bezogenen Tests wesentlich breiter angelegt als unterrichtsbegleitende Messungen. Zum anderen signalisieren Standards, dass Lehrende wie auch Lernende für die Erreichung der betreffenden Sollwerte verantwortlich gemacht werden, was zur Einengung von Unterrichtsprozessen führen und die Motivation der Schüler beeinträchtigen kann (vgl. Deci u.a. 1981; Koretz 2008). Zudem haben Rückmeldungen aus summativen Tests – sofern sie überhaupt gegeben werden – häufig eine weniger unterstützende Form.

1.3 Feedback als zentrales Element von formativem Assessment

Schon Sadler (1989) sah die Information über festgestellte Leistungen, die Beteiligten rückgemeldet wird, als Kernelement des formativen Assessments. Hattie und Timperley (2007) kritisieren an gängigen Assessmentsystemen, dass diese nur Momentaufnahmen des Lernstands abbilden, anstatt Informationen bereit zu stellen, die von den Schülern für den weiteren Lernprozess und von Lehrkräften für die Unterrichtsgestaltung genutzt werden können. Um ein tieferes Verständnis des Lerngegenstands zu erreichen, sollte Feedback konkrete Aussagen darüber machen, wie man dem Lernziel noch näher kommen kann, indem Bearbeitungsprozesse nachvollzogen, Fehler und Lücken identifiziert und Strategien benannt werden. Wir sprechen hier im Folgenden von prozessbezogenem Feedback.

Im Rahmen der Cognitive Evaluation Theory (vgl. Deci/Koestner/Ryan 1999) wird zwischen „informierendem" und „kontrollierendem" Feedback unterschieden. Von informierendem Feedback wird eine positive Wirkung auf Motivation und Leistung erwartet, da es durch die Information über die individuellen Kompetenzen der Lernenden das grundlegende Bedürfnis nach Kompetenz unterstützt. Operationalisiert wird informierendes Feedback häufig durch Formulierungen, die die individuelle Leistung der Lernenden mit der durchschnittlichen Leistung in der jeweiligen Lerngruppe (soziale Bezugsnorm), dem vorherigen Leistungsstand (individuelle Bezugsnorm) oder dem Leistungsziel (kriteriale Bezugsnorm, z.B. angestrebte Kompetenzstufe) in Beziehung setzen. Kontrollierendes Feedback hingegen betont, wie sich Lernende (hätten) verhalten sollen. Es kann damit als Bedrohung des Bedürfnisses nach Autonomie wahrgenommen werden und die Motivations- und Leistungsentwicklung beeinträchtigen.

Aus der Forschungsliteratur lassen sich auch Hypothesen darüber ableiten, welche Faktoren die Verarbeitung des Feedbacks bestimmen. Hierzu zählen u.a. die Attribution des Erfolgs bzw. Misserfolgs und die Anstrengungsbereitschaft der Lernenden.

2. Zielsetzungen und Vorgehensweise des Forschungsprojekts Co²CA

2.1 Phasen des Projekts

In dem auf insgesamt sechs Jahre angelegten Projekt sind unterschiedliche empirische Zugänge zum Thema „Leistungsbeurteilung und Feedback im Mathematikunterricht" vorgesehen:

1. Die *sekundäranalytische Auswertung* einer *videobasierten Unterrichtsstudie* (vgl. Klieme/Pauli/Reusser 2009) im Hinblick auf verbales Feedback im Unterricht sowie Notengebung (vgl. Rakoczy u.a. 2008).
2. Eine breite Test- und Befragungsstudie (*Survey*), in der zum einen Tests und Kompetenzmodelle für spezifische mathematische Unterrichtseinheiten entwickelt werden und zum anderen die Praxis der Leistungsbeurteilung im Unterricht mittels Lehrer- und Schülerbefragungen untersucht wird.
3. Eine mit dem Survey verbundene experimentelle Vorstudie *(Rückmeldestudie)*, bei der Schülern individuelle schriftliche Rückmeldungen gegeben werden, deren Akzeptanz und Wirkung auf die Attribution von Erfolg bzw. Misserfolg geprüft wird.
4. Ein *„Labor"-Experiment,* bei dem untersucht wird, wie sich die Breite des Testinhalts („formativer" Test, bezogen auf eine Unterrichtseinheit vs. „summativer" Test, bezogen auf Bildungsstandards allgemein) und die Art der Rückmeldung (kriterial, sozial vergleichend oder prozessbezogen) auf Motivation und Leistung von Schülern auswirken.
5. Eine *Interventionsstudie (Feldexperiment)*, bei der Lehrkräfte gezielt für unterschiedliche Arten des formativen Assessments trainiert werden. Hier soll ökologisch valide überprüft werden, ob sich positive Effekte von Feedbackformen aus dem „Labor" in die Praxis übertragen lassen.

Im Folgenden gehen wir ausschließlich auf die Schritte 2 und 3 ein, die den Kern der ersten Projektphase bildeten.

2.2 Anlage des Surveys mit Rückmeldestudie

Im Mittelpunkt der Projektphase 2007–2009 stand eine Studie in 66 Realschulklassen und Gesamtschulkursen, die zum mittleren Abschluss führen. Sie verfolgte vor allem das Ziel, Testaufgaben zu entwickeln und psychometrisch zu skalieren sowie spezifische Kompetenzmodelle aufzustellen, die später im Laborexperiment und in der Inter-

ventionsstudie eingesetzt werden können. Bei der Kompetenzmodellierung sollte speziell untersucht werden, wie differenziert sich mathematische Teilkompetenzen (hier: Modellierungskompetenz und technische Kompetenz) innerhalb eingegrenzter Themenbereiche (hier: Satzgruppe des Pythagoras sowie Lineare Gleichungssysteme) erfassen lassen und wie sich diese Maße zu den breiten Kompetenzmodellen verhalten, die bei Bildungsstandards und Vergleichsarbeiten benutzt werden.

138 *Mathematikaufgaben* zu den beiden Themenbereichen und den zwei Kompetenzdimensionen wurden zum Teil unter Rückgriff auf Vorprojekte entwickelt; hinzu kamen 38 Items aus den Erhebungen zu nationalen Bildungsstandards[7]. Die Aufgaben wurden nach einem Youden-Square-Design auf 31 Testhefte verteilt, sodass beliebige Kombinationen vorkamen und die Position der Aufgaben rotiert wurde. Insgesamt 1560 Schüler der neunten Jahrgangsstufe bearbeiteten nach Zufall eines der Testhefte. Die doppelstündige Erhebung im jeweiligen Klassenverband erfolgte zwischen Mai und Juni 2008 unter der Verantwortung externer Testleiter. Offene Antworten wurden von trainierten Studierenden der Mathematikdidaktik ausgewertet. Zu allen Aufgaben wurden stichprobenartig Zweitkodierungen angefertigt; bei ungenügender Kodiererübereinstimmung wurden alle vorliegenden Schülerlösungen nachkodiert. Abschließend lag Kappa über alle Aufgaben und Kodierer hinweg bei sehr guten .93. Die Aufgaben wurden mit Hilfe der Software Conquest gemeinsam skaliert; einige zumeist sehr schwierige Items mussten aufgrund mangelnder Passung ausgeschlossen werden.

Im begleitenden *Schülerfragebogen* wurde vor dem Test die Testmotivation erhoben (Beispielitem: „Ich bin fest entschlossen, mich bei diesem Test voll anzustrengen."), danach wurden u.a. die wahrgenommene Bezugsnormorientierung der Lehrperson (Skalen „kriteriale Bezugsnormorientierung", „individuelle Bezugsnormorientierung") und Hintergrundvariablen wie Geschlecht und sozialer Status („Bücherfrage") erfasst. Im *Lehrerfragebogen* sollte für jeden Schüler das Ergebnis der Bearbeitung (richtig/falsch) bei vier Beispielaufgaben prognostiziert werden. Der Anteil korrekter Prognosen wurde als Indikator der diagnostischen Kompetenz verwendet. Ergänzend sollten die Lehrkräfte ihr eigenes professionelles Wissen im Bereich Diagnostik einschätzen (Markieritem: „Ich besuche Weiterbildungen oder informiere mich in der Literatur zum Thema Leistungsbeurteilung/Benotungskriterien."). Darüber hinaus wurden Aspekte der Praxis der Leistungsbeurteilung erhoben. Der Fragebogen wurde von 46 Lehrkräften vollständig ausgefüllt.

Etwa ein halbes Jahr nach dem Survey wurde eine *Rückmeldestudie* durchgeführt. Dabei wurde 167 Schülern aus 14 Klassen die individuelle Testleistung rückgemeldet, und zwar nach Zufall in einer von drei Formen: sozial vergleichend, kriterial oder prozessbezogen. Die *sozial vergleichende Bedingung* beinhaltete den Vergleich der individuellen Schülerleistung mit der durchschnittlichen Leistung der Klasse, getrennt für Modellierungs- und technische Kompetenz. In der *kriterialen Bedingung* wurde die Schülerleistung anhand von Kompetenzstufenmodellen – wiederum getrennt für Modellierungs- und technische Kompetenz – mit dem Lernziel für Realschüler der neunten

7 Wir danken dem Institut zur Qualitätsentwicklung im Bildungswesen (Humboldt-Universität zu Berlin) für die Überlassung dieses Materials.

Jahrgangsstufe verglichen. In der *prozessbezogenen Bedingung* wurden anhand von Beispielaufgaben spezifische Stärken und Schwächen sowie entsprechender Verbesserungs- und Übungsbedarf für beide Kompetenzdimensionen aufgezeigt. Anschließend wurde mittels Fragebogen die Wirkung des Feedbacks auf emotionale und motivationale Variablen (z.B. Zufriedenheit mit dem Ergebnis und Attribution) erhoben.

2.3 Fragestellungen

Mit Hilfe von Daten der Survey- und Rückmeldestudie sollen folgende Forschungsfragen beantwortet werden:

1. Bilden unterrichtsbezogene Tests auf der einen und allgemeine Bildungsstandard-bezogene Tests auf der anderen Seite psychometrisch eigenständige Leistungsdimensionen ab?
2. Welche Beurteilungspraxis dominiert im Alltag des Mathematikunterrichts? Wie stark unterscheiden sich dabei einzelne Klassen und inwieweit beeinflussen Lehrermerkmale wie z.B. diagnostische Kompetenz die Beurteilungspraxis?
3. Welche Formen der Leistungsbeurteilung hängen mit guten Leistungen bzw. hoher Motivation zusammen?
4. Wie wirken sich prozessbezogenes, kriteriales und sozial vergleichendes Feedback auf Zufriedenheit und Attribution aus?

3. Ergebnisse

3.1 Zur Modellierung mathematischer Kompetenzen für formatives und summatives Assessment

139 Testaufgaben aus der Survey-Studie konnten mit ausreichend gutem Modell-Fit auf einer gemeinsamen latenten Dimension abgebildet werden. Sie decken zwei Kompetenzen (technische bzw. Modellierungskompetenz) und drei Themenbereiche (Satzgruppe des Pythagoras, lineare Gleichungssysteme sowie allgemeine Bildungsstandard-bezogene Themen) ab. Dies spricht dafür, dass es prinzipiell möglich ist, themenspezifische (unterrichtsbezogene) und breit angelegte, standardbezogene Aufgaben in gemeinsamen Kompetenzmodellen abzubilden.

Auf der Basis von Aufgabenanalysen konnten aber auch spezifische Kompetenzmodelle für Teildimensionen (z.B. den Themenbereich „Pythagoras“) entwickelt werden. Die empirischen Daten ließen sich mit mehrdimensionalen Modellen besser abbilden als mit dem globalen eindimensionalen Modell. Die Leistungen in den beiden thematisch eingegrenzten Testteilen korrelierten messfehlerbereinigt untereinander zu .66, aber mit den Leistungen bei Bildungsstandard-Aufgaben zu .81 bzw. .77. Dies spricht dafür, dass thematisch fokussierte Leistungsmessungen, wie sie für formatives Assess-

ment gebraucht werden, zusätzliche und spezifische Informationen enthalten, die nicht mit einer summativen, standardbezogenen Messung abgedeckt sind.

3.2 Zur Praxis der Leistungsbeurteilung im Unterricht aus Lehrer- und Schülersicht

Aus Angaben der *Lehrkräfte* zur Praxis ihrer Leistungsbeurteilung lassen sich drei ausreichend reliable Skalen bilden, die untereinander nur geringfügig korrelieren:

- *Verbale Rückmeldungen* sind sehr häufig; sie werden beispielsweise bei der Tafelarbeit in zwei Drittel aller Klassen „immer" gegeben.
- Weniger häufig sind Maßnahmen, in denen es um die Vergabe von Noten oder zumindest um die explizite Bewertung einer Haus- oder Unterrichtsaufgabe durch die Lehrkraft geht (*lehrerzentrierte Beurteilungspraxis*).
- Verschiedene Praktiken, die eine *aktive Partizipation von Schülern* beinhalten (z.B. Selbsteinschätzung, Peer-Bewertung), bilden eine weitere Skala. Die Mittelwerte ihrer Items liegen jedoch zwischen „nie" und „manchmal". Beispielsweise findet Leistungsbewertung als expliziter Gegenstand des Unterrichts („Die Schüler bewerten ihre Arbeit anhand von Kriterien, die wir im Unterricht entwickelt haben.") nur in jeder fünften Mathematikklasse „manchmal" statt; noch seltener werden Portfolios oder Lerntagebücher eingesetzt.

Etwa die Hälfte der Befragten gibt an, Schüler mit Migrationshintergrund zumindest manchmal besonders mild zu bewerten. Die in manchen empirischen Studien (vgl. z.B. Klieme 2003) gefundene „positive Diskriminierung" ist also keineswegs eine implizite, sondern vielfach auch eine explizite Strategie – und zwar geschlechtsabhängig: Lehrerinnen neigen stärker als Lehrer dazu, Migranten oder auch Mädchen milder zu beurteilen. Darüber hinaus tendieren sie eher zu partizipativen Beurteilungspraktiken und verbalen Rückmeldungen.

Die befragten *Schüler* nehmen zumeist eine individuelle Bezugsnormorientierung der Lehrkräfte wahr („Wenn jemand seine Leistungen gegenüber früher verbessert, so wird er dafür von unserem Lehrer besonders gelobt."). Je stärker die Lehrkraft selbst die Rückmeldefunktion von Noten betont, und je stärker sie von partizipativen Formen der Leistungsbeurteilung berichtet, umso eher berichten die Schüler von dieser Art der Bezugsnormorientierung. Nur etwa 10% der Lehrkräfte definieren jedoch vor Klassenarbeiten bzw. Prüfungen explizite Kriterien.

Die *diagnostische Kompetenz* der Lehrkräfte (d.h. hier: die Treffsicherheit ihrer Prognose von Testleistungen) erweist sich als durchaus verhaltensrelevant: sie korrespondiert mit der Tendenz, Schülern verstärkt verbale Rückmeldungen zu geben. Selbst eingeschätztes diagnostisches Wissen hängt mit der „objektiven" diagnostischen Kompetenz nicht zusammen, geht aber mit partizipativen Beurteilungsverfahren und (aus Sicht der Schüler) mit einer individuellen Bezugsnormorientierung einher.

3.3 Zusammenhänge der Beurteilungspraxis mit Motivation und Leistung der Schüler

Nachdem somit gezeigt ist, dass unterschiedliche Praktiken der Leistungsbeurteilung identifiziert werden können, die mit Lehrermerkmalen korrelieren, soll die für das Forschungsprojekt zentrale Frage untersucht werden, wie diese Variablen ihrerseits mit Motivation und Leistung der Schüler zusammenhängen. Wir prüften dies mit Hilfe von Mehrebenenanalysen (s. Tabelle 1) und gingen davon aus, dass die diagnostische Kompetenz sowie das mathematische Anspruchs- und somit Anregungsniveau des Unterrichts[8] hinsichtlich der motivationalen Prozesse neutral sind, aber eine bessere kognitive Förderung ermöglichen. Allerdings liegen hier ausschließlich querschnittliche Daten vor, sodass keine Kausalzuschreibungen zulässig sind.

Einflussgröße	**Effekt (*β*-Koeffizient) auf …**	
	Testmotivation	**Testleistung**
Ebene 2: Klasse		
Diagnostische Kompetenz der Lehrperson	−0,01	0,02 *
Lehrerzentrierte Beurteilungspraxis	−0,16 *	−0,33 *
Anspruchsniveau des Unterrichts	0,01	0,26 *
Ebene 1: Schüler		
Geschlecht männlich	−0,25 ***	0,36 ***
Sozialer Status	0,01	0,11 ***
Familiensprache Deutsch	0,02	−0,29 ***
Schülerperzeption: individuelle Bezugsnorm	0,17 ***	0,13 **

* Angegeben sind standardisierte Regressionskoeffizienten.

*Tab. 1: Mehrebenenanalysen zu Effekten der Leistungsbeurteilungspraxis**

8 Wir verwenden hierfür einen Indikator, der auf Schülerwahrnehmungen aufbaut: den Klassenmittelwert bezüglich der Frage „Wie häufig löst ihr im Mathematikunterricht Textaufgaben?“. Da Problemlöse- und Modellierungsaufgaben im deutschen Mathematikunterricht umgangssprachlich als „Textaufgaben“ bezeichnet werden, spiegelt sich in diesem Indikator das von Schülern wahrgenommene Anspruchsniveau des Unterrichts.

Leistungsbeurteilung und Motivation: Eine stark lehrer- und notenzentrierte Leistungsbeurteilung ist in der Tat mit niedrigerer Testmotivation verbunden, während der persönliche Eindruck von Schülern, die Lehrperson orientiere sich an individuellen Lernfortschritten, mit höherer Motivation einhergeht.

Leistungsbeurteilung und Testleistung: Als Leistungskriterium wurde der Mittelwert aus der Leistung in den Bereichen Modellierungskompetenz und technische Kompetenz verwendet. Auch hier zeigt sich, dass lehrerzentrierte Leistungsbeurteilung negativ, die Wahrnehmung einer individuellen Bezugsnorm jedoch positiv mit der Testleistung zusammenhängt, und die Effekte des individuellen Hintergrunds bei der Leistung – wie aus der Unterrichtsforschung hinlänglich bekannt – noch stärker ausgeprägt sind als bei der Motivation. Anders als bei der Erklärung von Testmotivation sind nun aber auch diagnostische Kompetenz und mathematisches Anspruchsniveau wichtig.

3.4 Auswirkungen von Feedback

Auf der Grundlage des Forschungsstands zu Feedback wurde erwartet, dass prozessorientierte und kriteriale Rückmeldungen in motivationaler Hinsicht positiver aufgenommen werden als eine sozial vergleichende Rückmeldung. Außerdem wurde vermutet, dass die Art der Rückmeldung sich darauf auswirkt, ob Erfolg bzw. Misserfolg eher auf Begabung oder Zufall zurückgeführt, also intern oder extern attribuiert werden.

In der Tat ergaben die Analysen, dass eine *kriteriale* Form der Rückmeldung, verglichen mit sozial vergleichendem Feedback, die Zufriedenheit mit dem Rückmeldeergebnis und die Tendenz zur internalen Attribution („Begabung“ statt „Glück“) verstärkt. Allerdings zeigen sich derartige Effekte nicht bei der *prozessbezogenen* Rückmeldung. Dies könnte vor allem damit zusammenhängen, dass zwischen Test und Feedback in diesem Fall 5 bis 6 Monate lagen und im Anschluss an das Feedback kein weiterer Test zu bearbeiten war, für den die Anregungen hilfreich wären. Möglicherweise macht auch das Fehlen jeglicher Vergleichsinformation (die ja auch in der kriterialen Rückmeldung implizit enthalten ist, wenn alle möglichen Kompetenzstufen beschrieben werden) diese Rückmeldeform unattraktiv.

4. Zusammenfassung und Diskussion

In der ersten Teilstudie des Projekts Co^{2}CA konnten bereits substantielle Erkenntnisse zu den Kernfragestellungen gewonnen werden.

- In den hier untersuchten Realschulklassen dominiert nach Angaben der Lehrpersonen eine verbale Rückmeldekultur, verbunden mit verschiedenen lehrer- und notenzentrierten Beurteilungsformen. Partizipative Formen (Selbst- oder Peer-Evaluation) sind insgesamt selten, finden sich aber vergleichsweise häufig bei Lehrern und insbesondere Lehrerinnen, die angeben, sich gut mit diagnostischen Fragen auszuken-

nen. Diese Ergebnisse verweisen darauf, dass Beurteilungs- und Rückmeldekulturen von Lehrkraft zu Lehrkraft bzw. von Klasse zu Klasse variieren, sodass es sich lohnt, Ursachen und Folgen weiter zu prüfen.

- In Mehrebenenmodellen lassen sich tatsächlich Zusammenhänge mit Motivation und Leistung der Schüler identifizieren: lehrer- und notenzentrierte Beurteilungspraktiken gehen mit niedrigerer, eine aus Schülersicht wahrgenommene individuelle Bezugsnormorientierung der Lehrkraft hingegen mit höherer Motivation einher, während die diagnostische Kompetenz der Lehrperson (hier verstanden als hohe Treffsicherheit bei der Vorhersage von Schülerantworten) mit besseren Testleistungen der Schüler verbunden ist.
- Unterschiedliche Formen der Rückmeldung, die experimentell variiert wurden, wirkten sich erwartungsgemäß auf die Motivation der Schüler und auf deren Attribution der Testergebnisse aus, wobei insbesondere die kriteriale Rückmeldung auf der Basis von Kompetenzstufenmodellen signifikant bessere Effekte hatte als eine sozialnormorientierte Rückmeldung, wie sie in Schulklassen üblich ist.

Mit diesen Analysen, die hier nur in einem ersten, kursorischen Durchgang präsentiert werden konnten, beginnt das Projekt Co²CA seine Hauptfragestellung zu beantworten: die Frage nach dem Zusammenhang zwischen Leistungsbeurteilung/Feedback einerseits sowie Unterrichtsqualität und Lernergebnissen andererseits. Die Ergebnisse sind – auch wenn es sich hier nur um Querschnittsresultate handelt – kompatibel mit der Ausgangsthese des Projekts, dass eine argumentative, aktivierende, auf individuellen Bezugsnormen aufbauende Leistungsbeurteilung mit differenzierten (kriterialen bzw. prozessbezogenen) Rückmeldungen ein wichtiges Qualitätsmerkmal des Unterrichts darstellt. Für die Praxis ergibt sich als Schlussfolgerung, das Potential formativer Assessment-Praktiken stärker zu nutzen und bei summativen Testverfahren sorgfältig auf die Gestaltung der Rückmeldungen zu achten.

Literatur

Baker, E. (2007): The End(s) of testing. In: Educational Researcher 36, H. 6, S. 309–317.

Black, P.J./William, D. (1998): Assessment and Classroom Learning. In: Assessment in Education: Principles, Policy and Practice 5, H. 1, S. 7–74.

Brookhart, S.M. (1993): Teachers' grading practices: Meaning and values. In: Journal of Educational Measurement 30, S. 123–142.

Deci, E.L./Koestner, R./Ryan, R.M. (1999): A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. In: Psychological Bulletin 125, S. 627–668.

Deci, E.L./Schwartz, A.J./Sheinman, L./Ryan, R.M. (1981): An instrument to asses adults' orientations toward control versus autonomy with children: Reflections on intrinsic motivation and perceived competence. In: Journal of Educational Psychology 73, H. 5, S. 642–650.

Hattie, J./Timperley, H. (2007): The power of feedback. In: Review of Educational Research 77, H. 1, S. 81–112.

Heritage, M. (2007): Formative Assessment: What Do Teachers Need to Know and Do? In: Phi Delta Kappan 89, H. 2, S. 140–145.

Klieme, E. (2003): Benotungsmaßstäbe an Schulen: Pädagogische Praxis und institutionelle Bedingungen. Eine empirische Analyse auf der Basis der PISA-Studie. In: Döbert, H./von Kopp, B./Martini, R./Weiß, M. (Hrsg.): Bildung vor neuen Herausforderungen: historische Bezüge, rechtliche Aspekte, Steuerungsfragen, internationale Perspektiven. Neuwied/Kriftel: Luchterhand, S. 195–210.

Klieme, E./Pauli, C./Reusser, K. (2009): The Pythagoras Study. In: Janik, T./Seidel, T. (Hrsg.): The Power of Video Studies in Investigating Teaching and Learning in the Classroom. Münster: Waxmann, S. 137–160.

Koretz, D. (2008): Test-based Educational Accountability. Research Evidence and Implications. In: Zeitschrift für Pädagogik 54, H. 6, S. 777–790.

Pellegrino, J.W./Chudowsky, N./Glaser, R. (2001): Knowing what students know. The science and design of educational assessment. Washington, DC: National Academic Press.

Rakoczy, K./Klieme, E./Bürgermeister, A./Harks, B. (2008): The Interplay between Student Evaluation and Instruction: Grading and Feedback in Mathematics Classrooms. In: Zeitschrift für Psychologie/Journal of Psychology 216, H. 2, S. 111–124.

Sadler, D.R. (1989): Formative assessment and the design of instructional systems. In: Instructional Science 18, S. 119–144.

Schrader, F.-W. (^{3}2006): Diagnostische Kompetenz von Eltern und Lehrern. In: Rost, D.H. (Hrsg.): Handwörterbuch Pädagogische Psychologie. Weinheim/Basel: Beltz, S. 95–100.

Shavelson, R.J./Young, D.B./Ayala, C.C./Brandon, P.R./Furtak, E.M./Ruiz-Primo, M.A./Tomita, M.K./Yin, Y. (2008): On the Impact of Curriculum-Embedded Formative Assessment on Learning: A Collaboration between Curriculum and Assessment Developers. In: Applied Measurement in Education 21, S. 295–314.

Wilson, M. (2008): Cognitive Diagnosis Using Item Response Models. In: Zeitschrift für Psychologie/Journal of Psychology 216, S. 73–87.

Anschrift der Autor/innen

Prof. Dr. Eckhard Klieme, Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Schloßstraße 29, D-60486 Frankfurt a.M.
E-Mail: klieme@dipf.de

Anika Bürgermeister, M.A., Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Schloßstraße 29, D-60486 Frankfurt a.M.
E-Mail: buergermeister@dipf.de

Birgit Harks, Dipl. Psych., Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Schloßstraße 29, D-60486 Frankfurt a.M.
E-Mail: harks@dipf.de

Prof. Dr. Werner Blum, Universität Kassel; Fachbereich Mathematik, Heinrich-Plett-Str. 40, D-34132 Kassel
E-Mail: blum@mathematik.uni-kassel.de

Dr. Dominik Leiß, Leuphana Universität Lüneburg, Institut für Mathematik und ihre Didaktik, Scharnhorstr. 1, D-21335 Lüneburg
E-Mail: leiss@me.com

Dr. Katrin Rakoczy, Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Schloßstraße 29, D-60486 Frankfurt a.M.
E-Mail: rakoczy@dipf.de

Olga Kunina-Habenicht/Oliver Wilhelm/Franziska Matthes/André A. Rupp

Kognitive Diagnosemodelle: Theoretisches Potential und methodische Probleme

Projekt Kognitive Diagnosemodelle[1]

1. Kognitive Diagnosemodelle: Eine Einführung

Im Rahmen dieses Forschungsprojekts werden das theoretische Potential und die Schwierigkeiten neuartiger psychometrischer Modelle, sog. kognitiver Diagnosemodelle (CDMs) untersucht. Mit der Verwendung der CDMs sind im Wesentlichen drei Hoffnungen verknüpft. Zum einen sollen sie eine mehrdimensionale Abbildung der Kompetenzen ermöglichen, die auf einer theoretisch begründeten und vorab festgelegten Klassifikation der Aufgaben zu entsprechenden Fähigkeitsdimensionen basiert. Dies wird am Beispiel einiger Mathematikaufgaben erläutert.

Damit ein Schüler[2] eine bestimmte Aufgabe in einem Mathematiktest erfolgreich bearbeiten kann, muss er mehrere Teilprozesse korrekt ausführen. Das Lösen einer Sachaufgabe erfordert bspw. die Fertigkeit zum Leseverständnis und Modellieren. Modellieren bezeichnet die mentale Konstruktion der Problemsituation (des sog. realen Modells bzw. des Situationsmodells) und deren anschließende Übertragung in ein mathematisches Modell mit einer oder mehreren Operationen (vgl. Blum u.a. 2006). Die aufgestellte mathematische Rechnung muss schließlich durch die korrekte Anwendung von Rechenalgorithmen gelöst werden. Dabei können verschiedene Merkmale der Aufgabe eine Rolle spielen (z.B. kann in der Rechnung ein Zehner-Übertrag gefordert sein). Alle hier genannten Teilprozesse der Aufgabenbearbeitung lassen sich als mathematische *Teilfertigkeiten* auffassen.

Die für die Lösung der Aufgabe erforderlichen Teilkompetenzen werden in einer *Q*-Matrix repräsentiert (s. Tab. 1). Ist für eine Aufgabe nur ein Eintrag in der *Q*-Matrix verzeichnet, so fordert die entsprechende Aufgabe lediglich die Beherrschung eines der postulierten Teilprozesse. Aufgaben, die die Ausführung von mehr als nur einem Teilprozess verlangen, weisen multiple Einträge in der *Q*-Matrix – sog. *Mehrfachladungen* – auf. Eine *Q*-Matrix beschreibt somit ein Klassifikationssystem von Teilfertigkeiten einer übergeordneten Kompetenz.

Die zweite Erwartung an CDMs ist, dass sie eine statistisch begründete Klassifikation der Schüler bzgl. einer oder mehreren Kompetenzen in sog. Kompetenzprofilen er-

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: RU 1424/3-1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

2 Im Folgenden wird für Personenbezeichnungen beider Geschlechter nur die männliche Form verwendet.

lauben. Um dies zu ermöglichen, wird dem Schüler für jede postulierte Teilfertigkeit eine Ausprägung auf einer kategorialen Variable zugewiesen. Das bedeutet konkret, dass die Schüler für jede Teilfertigkeit in zwei Fähigkeitsgruppen klassifiziert werden (Schüler A beherrscht die Fähigkeit (sog. „masters") oder Schüler A beherrscht die Fähigkeit nicht (sog. „non-masters"). Eine solche statistische Klassifikation stellt selbstverständlich eine Vereinfachung dar. Mit der Formulierung „Schüler beherrscht die Fähigkeit in einer Grundrechenart" meinen wir, dass dieser Schüler in einem bestimmten Zahlenraum sicher Rechnungen dieser Grundrechenart (ggf. mit Übertrag und verschiedenen Positionen der unbekannten Operatoren) im Kopf ausrechnen kann. Er kann darüber hinaus diese Grundrechenart flexibel zur Lösung von Sachaufgaben anwenden und erkennen, wann eine Anwendung der inversen Operation erforderlich ist.

Schüler, die durch die CDMs als „non-master" eingestuft werden, können zwar einzelne Rechenaufgaben ausrechnen, haben jedoch Schwierigkeiten, die korrekte mathematische Operation zur Lösung von Sachaufgaben einzusetzen. Sie haben darüber hinaus Probleme zu erkennen, in welchen Fällen die Anwendung der inversen Operation erforderlich ist.

Über alle Teilfertigkeiten hinweg ergibt sich so ein Kompetenzprofil des Schülers, das an Schüler, Lehrkräfte und Eltern zurückgemeldet werden kann. Es wird somit möglich, auf spezifische Schwierigkeiten des Schülers einzugehen und weiteres Lernen optimiert zu fördern.

Schüler mit identischen Kompetenzprofilen werden in gleiche Kompetenzklassen zusammengefasst. Aus theoretischer Sicht ist denkbar, dass bestimmten Klassen mehr Schüler zugeordnet werden als anderen Klassen. Bspw. könnte es eine Kompetenzklasse geben, in der Kinder zwar schon sicher die Addition und Subtraktion im Zahlenraum bis 1000 im Kopf, aber noch nicht sicher die Multiplikation und Division im Kopf im Zahlenraum größer als 100 bzw. im Bereich des großen Ein-Mal-Eins beherrschen. Unplausibel wäre dagegen eine Kompetenzklasse mit dem Antwortmuster, bei dem Kinder korrekt Subtraktionsaufgaben lösen, aber Schwierigkeiten bei den Additionsaufgaben haben. Somit lassen sich einige der Kompetenzklassen zum Teil in eine geordnete Rangfolge bringen. So zeigen Schüler, die alle vier Grundrechenarten „beherrschen", insgesamt bessere Testleistungen als Kinder, die nur die Strichrechnung erfolgreich bewältigen. Die Anzahl der Schüler in verschiedenen Kompetenzklassen gibt dann einen Hinweis darauf, in welcher Reihenfolge die Teilfertigkeiten in der Regel erworben werden.

Die dritte Hoffnung, die mit CDMs verbunden ist, betrifft die Modellierung von nicht-kompensatorischen Modellen, die mit etablierten statistischen Methoden nicht geschätzt werden können. In *kompensatorischen Modellen* wird angenommen, dass Mängel in einer Teilfertigkeit durch Kenntnisse einer anderen Fertigkeit ausgeglichen (bzw. kompensiert) werden können. Nicht-kompensatorische Modelle hingegen postulieren, dass alle Teilfertigkeiten erforderlich sind, um eine Aufgabe korrekt lösen zu können. Da viele didaktische und psychologische Modelle davon ausgehen, dass alle postulierten Teilfertigkeiten für die Lösung einer Aufgabe erforderlich sind, erlauben die nicht-kompensatorischen CDMs im Gegensatz zu etablierten statistischen Methoden explizit die Prüfung dieser Modelle.

2. Kognitive Diagnosemodelle: Methodische Aspekte

Der folgende Abschnitt soll ein grundlegendes Verständnis für CDMs und deren Einordnung neben anderen, etablierten statistischen Methoden vermitteln. Zusammenfassend lassen sich CDMs als konfirmatorische probabilistische Item-Response (IRT) Modelle mit kategorialen latenten Variablen, die Mehrfachladungsstrukturen erlauben, beschreiben. Im Folgenden werden die Begriffe „konfirmatorisch" und „probabilistische IRT Modelle" näher erläutert. Zunächst wird hierfür beschrieben, was unter einem latenten Faktor zu verstehen ist.

Wie die meisten Personeneigenschaften in der Bildungsforschung sind Kompetenzen in der Regel nicht direkt beobachtbar; man kann auf diese nur indirekt aufgrund von beobachteten Aufgabenlösungen schließen. In diesem Fall spricht man von einem *theoretischen Konstrukt*, das in einem statistischen Modell durch eine sog. *latente Variable* bzw. einen *latenten Faktor* repräsentiert wird.

Konfirmatorische Modelle prüfen, ob das theoretisch angenommene Strukturmuster der untersuchten Kompetenz die empirisch erhobenen Daten ausreichend gut beschreibt. *Probabilistische IRT-Modelle* stellen eine breite Klasse von statistischen Modellen dar, die probabilistische (d.h. wahrscheinlichkeitsbezogene) Zusammenhänge zwischen dem beobachteten Antwortverhalten und der interessierenden latenten Fähigkeit der Probanden modellieren. In spezifischen IRT-Modellen werden bei der Schätzung der latenten Fähigkeitsverteilungen Aufgabenschwierigkeit, Trennschärfe der Aufgaben oder die Ratewahrscheinlichkeit berücksichtigt. Im Unterschied zu probabilistischen IRT-Modellen werden in CDMs an Stelle von kontinuierlichen, *kategoriale* latente Variablen angenommen. Kontinuierliche Variablen beschreiben den Grad der Beherrschung einer Fertigkeit auf einer Skala mit fließenden Übergängen. Kategoriale Variablen hingegen sind meistens dichotom, sie haben also nur zwei Ausprägungen wie bspw. Beherrschung und Nichtbeherrschung einer Fertigkeit.

In der neueren methodischen Literatur, insbesondere im angelsächsischen Raum, wurden in den letzten Jahren verschiedene CDMs entwickelt (vgl. für eine Übersicht Rupp/Templin/Henson in Druck; Rupp/Templin 2008; diBello/Roussos/Stout 2007; Leighton/Gierl 2007). Zu diesen Modellen zählen u.a. die DINA und NIDA Modelle (vgl. z.B. de la Torre 2009; Junker/Sijtsma 2001), die DINO und NIDO Modelle (vgl. z.B. Rupp/Templin/Henson, in Druck) sowie die flexiblen Modellklassen des „general diagnostic model" (vgl. z.B. von Davier 2005). Einen neuen integrativen Ansatz stellen log-lineare Modelle dar (vgl. Henson/Templin/Willse 2009). Bisher gibt es jedoch nur wenige erfolgreiche empirische Anwendungen dieser Modelle. In vielen Fällen handelt es sich um Reanalysen von Daten, die ursprünglich nur eine einzelne Dimension messen (sog. „Retrofitting"). Beim Retrofitting treten häufig sehr hohe Korrelationen zwischen den angenommenen latenten Teilfertigkeiten sowie Konvergenzprobleme auf. Diese Probleme können dafür sprechen, dass nicht alle postulierten Teilfertigkeiten statistisch und inhaltlich bedeutungsvoll voneinander separierbar sind.

3. Projektziele & Arbeitsprogramm

In dieser ersten Projektphase war das vorrangige Ziel, CDMs am Beispiel eines neu entwickelten Mathematiktests erfolgreich einzusetzen sowie diese Modelle mit anderen etablierten statistischen Auswertungsmethoden (IRT-Modellen und konfirmatorischen faktoranalytischen Modellen) zu vergleichen. Dieses Ziel war deswegen so entscheidend, weil der Nachweis des inkrementellen Nutzens der CDMs gegenüber etablierten statistischen Modellen umstritten ist und es bisher zu wenige Erfolgsfälle aus Anwendungen gibt, die kein Retrofitten darstellen (vgl. Wilhelm/Robitzsch 2009).

Zunächst stand die Entwicklung eines Klassifikationssystems mathematischer Fertigkeiten im Vordergrund. Anschließend wurden Aufgaben entsprechend dieses Klassifikationssystems entworfen. Die entwickelten Aufgaben wurden in einer ersten Pilotierungsstudie im April 2008 an 464 Grundschulkindern der dritten und vierten Klassen erprobt. Es folgte die Anpassung der Aufgaben und eine Normierung von Items mit zufriedenstellenden Itemschwierigkeiten und -trennschärfen in einer zweiten Studie im Juli 2008 an 747 Dritt- und Viertklässlern. In einer dritten Studie im November und Dezember 2008 wurden schließlich deutschlandweit Daten von 2123 Kindern der vierten Klassen an 47 Schulen erhoben, um das neu entwickelte Messinstrument abschließend zu normieren und anhand der Aufgaben zur Erfassung der Bildungsstandards Mathematik in der Grundschule sowie anhand des etablierten Mathematiktests DEMAT 3+ zu validieren. Im Folgenden soll der Prozess der Aufgabenentwicklung beschrieben werden.

4. Aufgabenentwicklung

Um den denkbaren Nutzen der CDMs optimal zu realisieren und um oben erörterte Schwierigkeiten beim Retrofitten zu vermeiden, wurden gezielt neue Aufgaben konstruiert, die auf einer vorab theoretisch definierten *Q*-Matrix basieren. Die Definition der *Q*-Matrix resultierte aus einem neu entwickelten Klassifikationssystem arithmetischer Kompetenzen. Die Entwicklung des Klassifikationssystems orientierte sich an Lehrplänen verschiedener Bundesländer, Bildungsstandards der Kultusministerkonferenz sowie an fachdidaktischer Literatur.

Um die Anforderungen der Aufgaben so eindeutig wie möglich zu halten, wurde bewusst auf die Verwendung komplexer offener Mathematikaufgaben verzichtet. Ein Problem komplexer offener Aufgaben besteht u.a. darin, dass beim Lösen solcher komplexer Aufgaben allgemeine kognitive Fähigkeiten und Sprachverständnis eine wichtige Rolle spielen, so dass unklar ist, inwieweit sich Kompetenzen wie „Kommunizieren“ oder „Argumentieren“ von solchen allgemeinen Fähigkeiten trennen lassen. Stattdessen wurden möglichst grundlegende Aufgaben zu den im Unterricht besonders intensiv behandelten Inhalten eingesetzt.

Mit diesen Prämissen und Einschränkungen versehen, wurde der Versuch unternommen, optimale Bedingungen für die erfolgreiche Schätzung theoretisch schlüssiger CDMs zu schaffen. Hierfür wurde für einfache Rechenaufgaben und Sachaufgaben ein

	X + 22 = 27	**113 + X = 204**	**106 – 8 = X**	**10 × 8 = X − 15**	**250 : 5 = X × 5**	**Sachaufgabe 4**
Addition	1	1	0	0	0	1
Subtraktion	0	0	1	1	0	1
Multiplikation	0	0	0	1	1	0
Division	0	0	0	0	1	0
Modellieren	0	0	0	0	0	1
Zehner-Übertrag	0	0	1	1	0	0
Zahlenraum größer 100	0	1	1	0	1	0
Umkehr-operation	1	1	0	1	1	0
Unbekannte Anfangsmenge	1	0	0	–	–	0
Unbekannte Teilmenge	0	1	0	–	–	0
Unbekannte Endmenge	0	0	1	–	–	1

Tab. 1: Q-Matrix für kontextfreie und kontextgebundene Aufgaben

atomistisches Klassifikationssystem entwickelt, das sich vornehmlich an der mathematischen Struktur der Aufgaben orientiert. Grundsätzlich ist denkbar, das vorliegende Aufgabenmaterial fachdidaktisch inspirierter zu klassifizieren und dabei die zugrunde liegenden kognitiven Denkhandlungen stärker zu betonen, anstatt sehr nah an Oberflächenaufgabenmerkmalen zu bleiben. Ein Ergebnis einer solchen Klassifikation könnten konkurrierende Messmodelle sein.

Drei zentrale Anforderungen an den neu entwickelten Mathematiktest wurden vorab als Fragestellungen formuliert: Die erste Fragestellung war, inwiefern der neu entwickelte Test ein reliables und valides Maß für die allgemeine mathematische Kompetenz liefert. Die zweite Fragestellung war, ob es der neu entwickelte Test auf einer elementareren Ebene erlaubt, die Beherrschung der vier Grundrechenarten und des Modellierens getrennt voneinander zu schätzen, auch wenn diese Teilkompetenzen voraussichtlich stark korreliert sind.

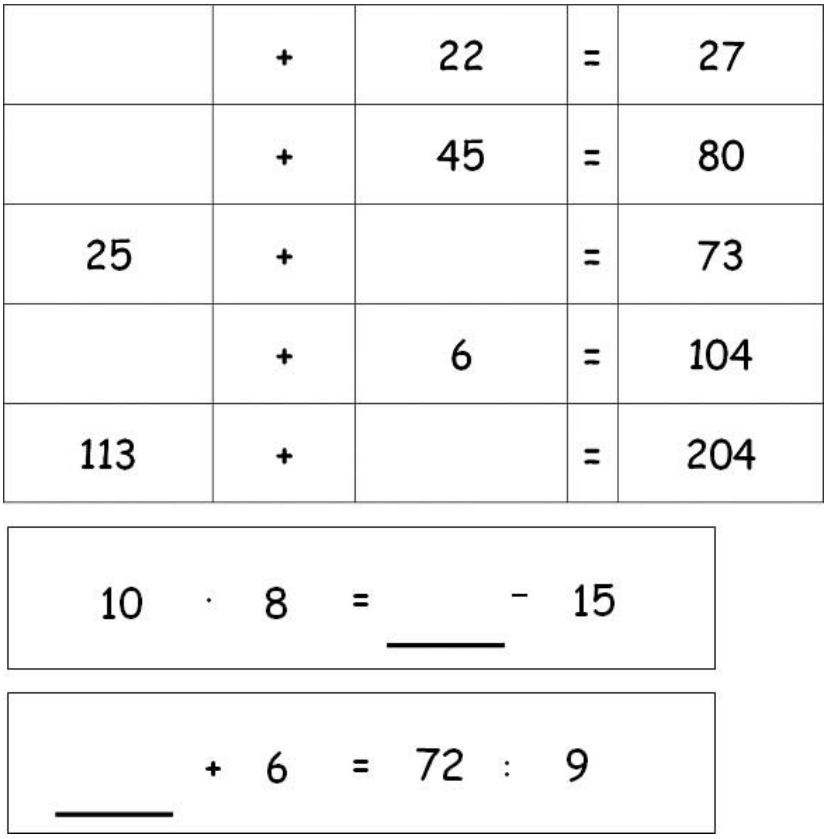


Abb. 1: Beispiel für kontextfreie einfache und komplexe Rechenaufgaben

Die dritte Fragestellung war, inwiefern mittels des Tests auf einer noch elementareren Ebene zwischen Fertigkeiten zur Bewältigung spezifischer Aufgabenanforderungen unterschieden werden kann. Die Auswahl der relevanten Aufgabenanforderungen beruht zum Teil auf den Arbeiten von Carpenter u.a. (1999). Die Aufgaben wurden hierbei u.a. danach unterschieden, ob Anfangs-, Teil- oder Endmengen unbekannt sind. Bei der Aufgabenkonstruktion wurde weiterhin die Anwesenheit oder Abwesenheit eines Übertrags („Zehner-Übertrag"), verschiedene Zahlenbereiche („Rechnen im Zahlenraum größer 100") und das Ausführen einer Umkehroperation (genauer einer Umkehroperation zu der Operation, die in der Aufgabe visuell gefordert wird) berücksichtigt. Die daraus resultierende Q-Matrix ist in Tabelle 1 dargestellt.

Aus diesem Prozess der Aufgabenentwicklung resultierten zwei Typen von Aufgaben. Der erste Typ beschreibt einfache und komplexe, kontextfreie Rechenaufgaben (s Abb. 1). Ein zweiter Typ von Aufgaben sind kontextgebundene Sachaufgaben (s. Abb. 2). In der dritten Studie im November und Dezember 2008 wurden darüber hinaus Aufgaben aus dem Inhaltsbereich „Messen und Größen" eingesetzt, die Umrechnungen und Anwendung arithmetischer Operationen auf verschiedene Maß-, Geld- und Längeneinheiten erforderten.

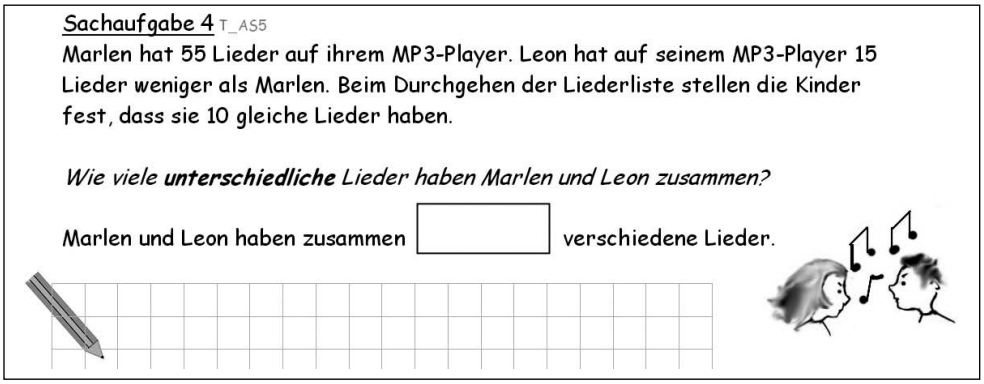


Abb. 2: Beispiel für eine kontextgebundene Sachaufgabe

5. Ergebnisse

Bei der Schätzung der CDMs ist die Anzahl und Beschaffenheit der theoretisch angenommenen relevanten Teilfertigkeiten von entscheidender Bedeutung. Um diese zu ermitteln, wurden im ersten Schritt mit Hilfe von konfirmatorischen Faktorenanalysen (CFA) (vgl. Kaplan 2000) drei denkbare Strukturmuster überprüft. Die geschätzten CFA-Modelle erlauben mittels verschiedener etablierter Modell-Fit-Indizes Aussagen über die Modellpassung und ermöglichen einen statistischen Vergleich der Modelle mittels des sog. Likelihood-Ratio-Tests.

Im ersten Modell wurde von einer allgemeinen arithmetischen Fertigkeit ausgegangen. Im zweiten Modell wurden drei verschiedene Fertigkeiten angenommen – Punktrechnung, Strichrechnung, Modellieren – wobei die Korrelationen zwischen allen drei Faktoren untereinander zugelassen wurden. Im dritten Modell wurde postuliert, dass die vier Grundrechenarten und das Modellieren fünf verschiedene, jedoch voneinander abhängige Konstrukte bzw. latente Faktoren darstellen.

Die Modellpassung für das erste Modell mit einer allgemeinen Arithmetikfähigkeit ist im Vergleich zu Modell 2 und Modell 3 deutlich schlechter. Aufgrund der hohen Korrelationen zwischen den latenten Faktoren ist es statistisch schwierig im Modell 3 zwischen Addition und Subtraktion bzw. Multiplikation und Division zu trennen. Die Datenlage in der Pilotierungsstudie spricht dafür, drei voneinander abhängige Konstrukte bzw. latente Faktoren zu unterscheiden: Punktrechnung, Strichrechnung und Modellieren (Modell 2). Darüber hinaus zeigen die Ergebnisse, dass Schüler in dritten und vierten Klassen große Schwierigkeiten mit Multiplikation und Division, insbesondere mit dem großen Ein-Mal-Eins haben.

Die zweite Frage, die in der Pilotierungsstudie beantwortet werden sollte, betraf den Vergleich der CDMs mit den korrespondierenden CFA-Modellen. Als ein Vertreter der CDMs wurde das *general diagnostic model* (GDM) gewählt, wobei hierbei das korrespondierende GDM zum CFA-Modell mit der besten Modellpassung geschätzt wurde.

Es zeigte sich, dass das GDM per se keine zusätzliche Information über die mehrdimensionalen CFA-Modelle hinaus bereitstellt. GDM erlaubt jedoch eine andere Aufbereitung der Information und erlaubt die direkte Schätzung der Kompetenzprofile der Schüler, die eine Auskunft darüber ermöglichen, welche Teilfertigkeiten vom Kind bereits beherrscht werden und welche noch nicht (vgl. Tab. 2). Diese Information kann potentiell für Rückmeldungen an Lehrkräfte und Eltern genutzt werden. Des Weiteren können die Häufigkeiten der beobachteten Kompetenzklassen (bspw. „Schüler beherrschen alle vier Rechenarten“ vs. „Schüler können addieren und subtrahieren, aber nicht ausreichend gut multiplizieren und dividieren“) Hinweise auf den Lernprozess des Erwerbs der arithmetischen Fertigkeiten liefern. Unsere Ergebnisse lassen vermuten, dass in der Regel zunächst Addition und Subtraktionskenntnisse erworben werden, die eine wichtige Voraussetzung für den Erwerb weiterer Teilfertigkeiten bilden. Aufbauend darauf erfolgt parallel der Erwerb der Modellierungskompetenz bzw. von Multiplikation und Division.

Eine ausführliche Beschreibung der hier berichteten Ergebnisse der Pilotierungsstudie sowie umfassende methodische Erläuterungen findet der interessierte Leser bei Kunina-Habenicht/Rupp/Wilhelm (2009).

In der vor kurzem abgeschlossenen Validierungsstudie wurde der neu entwickelte Mathematiktest zusammen mit den Bildungsstandardaufgaben in Mathematik für die Grundschule, DEMAT 3+ (vgl. Roick/Görlitz/Hasselhorn 2004) sowie dem kognitiven Fähigkeitstest (KFT; vgl. Heller/Perleth 1976) administriert. Erste Ergebnisse deuten auf mittelhohe Zusammenhänge zwischen dem neu entwickelten Mathematiktest und dem DEMAT 3+, den Bildungsstandardsaufgaben und der Mathematiknote hin. Ausführliche Informationen hierzu sind zu finden bei Kunina/Wilhelm/Rupp (2009).

In den nächsten Analyseschritten streben wir die Anwendung des neuen integrativen Ansatzes der log-linearen Modelle auf die Daten aus der Validierungsstudie an. Anschließend wird die Validität der so berechneten Kompetenzprofile anhand der Zusammenhänge mit den eingesetzten externen Messinstrumenten und mit der Mathematiknote geprüft. Ein weiteres Ziel in diesem Projekt betrifft die Untersuchung geeigneter Modell-Fit-Maße, die Auskunft über die Modellpassung und damit die Prüfung der Adäquatheit der CDMs geben.

Teilfertigkeiten					
Addition	**Subtraktion**	**Multiplikation**	**Division**	**Modellieren**	**3. Klasse**
0	0	0	0	0	23,5 %
1	1	1	1	1	21,0 %
1	1	0	0	1	14,2 %
1	1	1	1	0	6,6 %
1	1	0	0	0	6,3 %
1	0	0	0	0	6,0 %
1	1	1	0	1	4,6 %
69,2 %	54,5 %	44,8 %	37,2 %	48,5 %	82,2 %

Anmerkungen: 0 – Teilfertigkeit wird beherrscht; 1 – Teilfertigkeit wird nicht beherrscht. In der letzten Zeile ist der Anteil der Schüler angegeben, die die entsprechende Teilfertigkeit erfolgreich erworben haben. Genauere Informationen und Hinweise zu den statistischen Berechnungen sind zu finden bei Kunina-Habenicht/Rupp/Wilhelm (2009).

Tab. 2: Prozentualer Anteil von Schülern der 3. Klassen in ausgewählten Kompetenzklassen

6. Fazit: Zukunft der kognitiven Diagnosemodelle

Im abschließenden Abschnitt wollen wir eine subjektive Einschätzung der Vorzüge und Probleme geben, die aus unserer Sicht derzeit mit den CDMs bestehen.

Eine Hoffnung, die mit den CDMs verbunden ist, betrifft die statistische und damit scheinbar objektive Klassifikation der Schüler in Kompetenzklassen. Diese Klassifikation ist jedoch nur dann bedeutungsvoll, wenn die *Q*-Matrix korrekt und reliabel definiert wurde. Da die Definition der *Q*-Matrix jedoch idealerweise über ein Expertentreffen erfolgt, bei denen die Experten Konsens über die Definition der *Q*-Matrix sowie Klassifikation der Aufgaben in die *Q*-Matrix finden sollen, ist die Subjektivität der Klassifikation der Schüler bei den CDMs im Prozess der Definition der *Q*-Matrix „versteckt“ (vgl. Gorin 2009).

Die Definition der *Q*-Matrix – oder ggf. konkurrierender *Q*-Matrizen – sollte nach Möglichkeit auf der Grundlage einer psychologisch oder fachdidaktisch fundierten Theorie basieren. In der Literatur wird häufig bemängelt, dass in vielen Forschungsbereichen die theoretische Fundierung der zugrunde liegenden Teilfertigkeiten und kognitiven Prozesse gar nicht oder kaum ausgereift sind (vgl. Maris/Bechger 2009; Wilhelm/Robitzsch 2009). Ob die Orientierung bei der Erzeugung einer Q-Matrix wie in dieser Arbeit vorrangig über Aufgabenattribute oder über vermeintlich ablaufende Denkprozesse erfolgt, ist unseres Erachtens zunächst zweitrangig. Entscheidend ist – auch im direkten Modellvergleich – Modelle mit guter Passung und domänengerechter nomothetischer Spanne zu etablieren. Ein weiterer kritischer Punkt betrifft die Validität der Kompetenzprofile, die CDMs liefern, die bisher kaum untersucht worden ist (vgl. Sinharay/Haberman 2009).

Aus statistischer Perspektive bestehen bei der Schätzung der CDMs noch viele ungeklärte Fragen. Ein wesentliches Problem betrifft die Tatsache, dass die Komplexität der Modelle und damit die Anzahl der zu schätzenden Parameter mit der Anzahl der angenommenen latenten Faktoren deutlich ansteigen. Dies führt häufig zu Schätzproblemen, die sich bspw. in mangelnder Reliabilität, großen Standardfehlern der Parameterschätzungen oder ausbleibender Konvergenz äußern können (vgl. Maris/Bechger 2009). Ferner ist bisher nur wenig über die Beurteilung der Güte der Modellpassung bei den CDMs – im Vergleich zu den CFAs – bekannt (vgl. Levy 2009).

In diesem Projekt untersuchen wir einige der genannten problematischen Aspekte und hoffen eine überzeugende Anwendung für empirische Daten zu präsentieren, die klare Vorteile der CDMs gegenüber bisherigen etablierten statistischen Modellen aufzeigt.

Literatur

Blum, W./Drüke-Noe, C./Hartung, R./Köller, O. (Hrsg.) (2006): Bildungsstandards Mathematik: konkret. Berlin: Cornelsen.

Carpenter, T.P./Fennema, E./Franke, M.L./Empson, S.B./Levi, L.W. (1999): Children's mathematics: Cognitively guided instruction. Portsmouth, NH: Heinemann.

de la Torre, J. (2009): DINA model and parameter estimation: A didactic. In: Journal of Educational and Behavioral Statistics 34, S. 115–130.

di Bello, L.V./Roussos, L.A./Stout, W. (2007): Review of cognitively diagnostic assessment and a summary of psychometric models. In: Rao, C.R./Sinharay, S. (Hrsg.): Handbook of Statistics, Vol. 26. Psychometrics. Amsterdam: Elsevier, S. 979–1030.

Gorin, J.S. (2009): Diagnostic classification model: Are they necessary? Commentary on Rupp and Templin (2008). In: Measurement: Interdisciplinary Research and Perspectives 7, S. 30–33.

Heller, K.A./Perleth, Ch. (1976): Kognitiver Fähigkeitstest für 4.–12. Klassen (KFT 4–12+). Göttingen: Hogrefe.

Henson, R./Templin, J./Willse, J. (2009): Defining a Family of Cognitive Diagnosis Models Using Log-Linear Models with Latent Variables. In: Psychometrika 74, H. 2, S. 191–210.

Junker, B.W./Sijtsma, K. (2001): Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. In: Applied Psychological Measurement 25, S. 258–272.

Kaplan, D. (2000): Structural equation modelling: Foundations and extensions. Advanced Quantitative Techniques in the Social Sciences series. London: Sage.
Kunina-Habenicht, O./Rupp, A.A./Wilhelm, O. (2009): A Practical Illustration of Multidimensional Diagnostic Skills Profiling: Comparing Results from Confirmatory Factor Analysis and Diagnostic Classification Models. In: Studies in Educational Evaluation 35 (2), S. 64–70.
Kunina, O./Wilhelm, O./Rupp A.A. (2009): Validity of Proficiency Profiles in Arithmetic Ability based on Cognitive Diagnosis Models. Vortrag gehalten beim annual meeting of American Educational Research Association (AERA) in San Diego USA, 13.–17. April.
Leighton, J./Gierl, M. (Hrsg.) (2007): Cognitive diagnostic assessment for education: Theory and practice. Cambridge: Cambridge University Press.
Levy, R. (2009): Evidentiary reasoning in diagnostic classification models. In: Measurement: Interdisciplinary Research and Perspectives 7, S. 36–41.
Maris, G./Bechger, T. (2009): Equivalent diagnostic classification models. In: Measurement: Interdisciplinary Research and Perspectives 7, S. 41–46.
Roick, T./Görlitz, D./Hasselhorn, M. (2004): Deutscher Mathematiktest für dritte Klassen (DEMAT 3+). Göttingen: Beltz Test.
Rupp, A.A./Templin, J. (2008): Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. In: Measurement: Interdisciplinary Research and Perspectives 6, S. 219–262.
Rupp, A.A./Templin, J./Henson, R. (2010): Diagnostic measurement: Theory, methods, and applications. New York: The Guilford Press.
Sinharay, S./Haberman, S.J. (2009): How much can we reliably know about what examinees know? In: Measurement: Interdisciplinary Research and Perspectives 7, S. 46–49.
von Davier, M. (2005): A general diagnostic model applied to language testing data (Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service.
Wilhelm, O./Robitzsch, A. (2009): Have cognitive diagnostic models delivered their good? Some substantial and methodological concerns. In: Measurement: Interdisciplinary Research and Perspectives 7, S. 53–57.

Anschrift der Autor/innen

Olga Kunina-Habenicht, Goethe-Universität Frankfurt, Institut für Psychologie, Arbeitsbereich Pädagogische Psychologie, Senckenberganlage 15, D-60325 Frankfurt a.M.
E-Mail: kunina@paed.psych.uni-frankfurt.de

Oliver Wilhelm, Universität Duisburg-Essen, Berliner Platz 6–8, CO 8/17, D-45127 Essen
E-Mail: oliver.wilhelm@uni-due.de

Franziska Matthes, Humboldt Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Unter den Linden 6, D-10099 Berlin
E-Mail: franziska.matthes@student.hu-berlin.de

André A. Rupp, Department of Measurement, Statistics, and Evaluation (EDMS), University of Maryland, USA
E-Mail: ruppandr@umd.edu

Aiso Heinze

Mathematische Kompetenz modellieren und diagnostizieren

Eine Diskussion der Forschungsprojekte des DFG-Schwerpunktprogramms „Kompetenzmodelle“ aus mathematikdidaktischer Sicht

Review

1. Einleitung

Mit der Einrichtung des DFG-Schwerpunktprogramms „Kompetenzmodelle“ ist das zentrale Ziel verbunden, Kompetenzmodelle zu entwickeln und empirisch zu prüfen, auf deren Basis sich anschließend valide Messinstrumente konstruieren lassen. Letztere sollen dabei entweder für eine individuelle Kompetenzdiagnostik und damit zur Förderung individueller Lernprozesse dienen oder im Sinne des Assessment für Untersuchungen zum Monitoring von Bildungsinstitutionen und Bildungssystemen geeignet sein. Diesem Ziel entsprechend strukturiert sich das Forschungsprogramm in verschiedene Bereiche, welche die gesamte Bandbreite, ausgehend von der Entwicklung theoretischer Kompetenzmodelle über die Konstruktion adäquater psychometrischer Messmodelle und Verfahren zur empirischen Erfassung von Kompetenzen bis hin zur Nutzung von diagnostischen Informationen, umfasst.

Im Folgenden wird auf die in diesem Band dargestellten ersten Ergebnisse aus Forschungsprojekten eingegangen, die sich auf Kompetenzen in der Domäne Mathematik beziehen. Sie werden dabei insbesondere aus mathematikdidaktischer Perspektive betrachtet und dahingehend diskutiert, inwieweit sie das Potenzial zu einer Verbesserung des Lehrens und Lernens von Mathematik auf der Ebene des Individuums bzw. auf der Ebene von Bildungsinstitutionen oder des Bildungssystems haben.

2. Die Forschungsprojekte mit Bezug zur Domäne Mathematik

Die Ziele der in diesem Band beschriebenen mathematikrelevanten Forschungsprojekte decken die gesamte Bandbreite des Schwerpunktprogramms ab. Während es in dem Projekt *Heureko* darum geht, ein Kompetenzstrukturmodell zum mathematischen Problemlösen zu entwickeln, stehen in den drei Projekten *MAT*, *Kognitive Diagnosemodelle* und *Regelgeleitete Itementwicklung* eher psychometrische Modelle und Messverfahren im Vordergrund, die am Beispiel der Domäne Mathematik realisiert werden. Das Projekt *CoCa* schließlich untersucht die mögliche Verwendung von Kompetenzmessungen für die Unterrichtspraxis. Aus Sicht der Mathematikdidaktik stehen diese fünf Projekte damit exemplarisch für die gesamte Breite des Schwerpunktprogramms und den

zu erwartenden Nutzen an Erkenntnissen und Methoden zur Verbesserung der Lehr-Lern-Bedingungen für das Fach Mathematik. Vor diesem Hintergrund sollen die in diesem Band dargestellten ersten Ergebnisse dieser Projekte diskutiert werden.

2.1 Heureko: Repräsentationswechsel beim Problemlösen mit Funktionen – Identifikation von Kompetenzprofilen auf der Basis eines Kompetenzstrukturmodells

Das Projekt *Heureko* beschäftigt sich mit der Kompetenzstruktur des Problemlösens im mathematischen Inhaltsbereich „Wachstum und Veränderung" und betrachtet dabei vor allem die Fähigkeit des Repräsentationswechsels. Dieser sehr spezielle Kompetenzbereich umfasst aus mathematikdidaktischer Sicht wesentliche Aspekte des mathematischen Kompetenzerwerbs. So ist der Inhaltsbereich „Wachstum und Veränderung" zentral im Curriculum der Sekundarstufe, und das Problemlösen wird seit jeher als bedeutendes Bildungsziel des Mathematikunterrichts angesehen. Die Fähigkeit, verschiedene Repräsentationen von abstrakten Konzepten zu nutzen und zwischen ihnen zu wechseln, gilt als ein wesentlicher Aspekt mathematischer Kompetenz.

Das vorgestellte Kompetenzmodell umfasst vier Dimensionen, die ein mathematisches Verarbeiten innerhalb einer grafischen bzw. numerischen Repräsentation sowie den Wechsel zwischen einer situativen Repräsentation (Text/Bild) und einer numerischen bzw. grafischen Repräsentation beschreiben. Die Dimensionen wurden anhand von 80 Items operationalisiert und durch Daten von Schülerinnen und Schülern aus 37 Gymnasialklassen bestätigt. Kritisch anzumerken ist dabei – und dies wird auch von den Autorinnen und Autoren so gesehen – dass die Reliabilität einzelner Skalen unbefriedigend ist und hier Nachbesserungsbedarf besteht.

Als Ergebnis einer weiteren explorativen Studie wird die Identifikation von Kompetenzprofilen berichtet. Auf Basis einer latenten Klassenanalyse können sechs Kompetenzprofile in der Stichprobe ausgemacht werden, die in den vier Kompetenzdimensionen verschiedene Stärken und Schwächen aufweisen. Zur Erklärung wurden vermutete Zusammenhänge zu den individuellen kognitiven Grundfähigkeiten untersucht, die aber nicht bestätigt werden konnten. Hier sollten weitere individuelle Variablen sowie Unterrichtsvariablen einbezogen werden, die mögliche Einflussfaktoren für die Herausbildung der Kompetenzprofile darstellen könnten. So wäre bei der vorhandenen Stichprobengröße eine Analyse von Klasseneffekten (als Indikator für den Einfluss des Mathematikunterrichts) ein naheliegender nächster Schritt.

2.2 MAT – Multidimensionale adaptive Kompetenzdiagnostik: Ergebnisse zur Messeffizienz

Das Projekt *MAT* beschäftigt sich mit einem speziellen Messverfahren, welches exemplarisch an der Domäne Mathematik untersucht wird. Kern des Projektes ist die Analyse

der Messeffizienz von multidimensionalen adaptiven Testverfahren im Vergleich zu eindimensionalen adaptiven Testverfahren bzw. konventionellen Testverfahren. Die Frage ist dabei, ob bei gleichbleibender Messpräzision der Aufwand der Testungen reduziert werden kann.

Ein Effizienzvorteil des multidimensionalen adaptiven Testens (MAT) gegenüber dem eindimensionalen adaptiven Testen ist zu erwarten, wenn die betrachteten Kompetenzdimensionen hoch korreliert sind, da dann die Redundanz der Informationen in den zu erhebenden Daten reduziert werden kann. Wie in dem Beitrag berichtet wird, zeigen Simulationsstudien mit künstlich generierten Datensätzen genau dieses Ergebnis. Bei gleichbleibender Messpräzision kann bei einem fünfdimensionalen Modell mit einer angenommenen latenten Korrelation von .85 zwischen den Dimensionen die Messeffizienz gegenüber dem konventionellen Testverfahren um den Faktor 3,5 erhöht werden und auch gegenüber dem eindimensionalen adaptiven Testverfahren signifikant gesteigert werden. Wie die Autoren anmerken, handelt es sich hier um die obere Grenze der Effizienzsteigerung, da in der Simulation ein „optimaler" Itempool verwendet wurde, der in der Realität so nicht generiert werden kann.

Die Untersuchung im Rahmen von *MAT* ist der Grundlagenforschung zuzuordnen, da ein Messverfahren auf Basis von Simulationsstudien untersucht wird und demnach keinerlei reale Daten bzw. Kompetenzstrukturmodelle betrachtet werden. Dennoch steht im Hintergrund eine mögliche Anwendung des Verfahrens im Bereich der Mathematik auf Basis der theoretischen multidimensionalen Modelle mathematischer Kompetenz, wie sie in den Bildungsstandards beschrieben werden. Aus mathematikdidaktischer Sicht ist vor allem interessant, dass die Kompetenzmodelle direkt in multidimensionale psychometrische IRT-Modelle abgebildet werden können und durch das MAT eine effiziente Messung der komplexen Modelle möglich ist. Allerdings müssen beim MAT die Testaufgaben am Computer bearbeitet werden, womit bisher eine Reduzierung der Itemformate auf Multiple-Choice und geschlossene Items verbunden ist. Die Frage, ob in Zukunft computerbasierte Kompetenzmessungen möglich sein werden, die eine vergleichbare Messpräzision wie konventionelle offene Items aufweisen, ist derzeit noch offen.

2.3 Kognitive Diagnosemodelle: Theoretisches Potential und methodische Probleme

Dieser Beitrag diskutiert die praktische Realisierbarkeit der Modellierung von Daten aus diagnostischen Tests durch kognitive Diagnosemodelle. Unter kognitiven Diagnosemodellen werden konfirmatorische probabilistische IRT-Modelle verstanden, die statt kontinuierlicher kategoriale (in diesem Fall dichotome) Variablen annehmen. Entsprechend werden nicht mehr Kompetenzkomponenten im Grad ihrer Ausprägung modelliert, sondern allein ihr Vorhandensein oder Nichtvorhandensein. Die Grundlage für die Generierung kognitiver Diagnosemodelle ist, dass a priori ein Klassifikationssystem von Komponenten der zu betrachtenden Kompetenz erstellt wird. Dessen Operationali-

sierung wird als Beziehung von Testaufgaben und Kompetenzkomponenten in einer *Q*-Matrix dargestellt. Als Domäne wurde in dem Projekt die Arithmetik der Primarstufe gewählt, als Stichprobe Schülerinnen und Schüler der Jahrgangsstufen 3 und 4.

Wie die Autorinnen und Autoren selbst anmerken, steht und fällt der Nutzen des kognitiven Diagnosemodells mit der Klassifikation der Kompetenzkomponenten und der Wahl der *Q*-Matrix. Entsprechend ist der diagnostische Nutzen dieser Modelle darauf beschränkt, was vorher an theoretischen fachdidaktischen Annahmen eingegeben wurde. Das dargestellte Beispiel zur Diagnose der Kompetenz in Arithmetik kann nur als ein erstes Herantasten an die praktische Realisierung angesehen werden, da die Ergebnisse aus mathematikdidaktischer Sicht wenig bedeutsam sind. Eine Identifizierung der latenten Faktoren „Strichrechnung", „Punktrechnung" und „Modellieren" und eine damit verbundene Besetzung von Kompetenzklassen, auf deren Basis die Autorinnen und Autoren vermuten, dass zunächst Additions- und Subtraktionskenntnisse erworben werden, welche dann eine wichtige Voraussetzung für den Erwerb weiterer Teilfertigkeiten bilden, liefern kaum neue Erkenntnisse. Die kognitiven Diagnosemodelle werden sich daran messen lassen müssen, inwieweit sie Kompetenzen zu mathematischen Teilbereichen sinnvoll abbilden können. Dabei spielt nicht nur die grobe mathematische Struktur eine Rolle, sondern vielmehr die Frage nach verschiedenen Grundvorstellungen zu mathematischen Begriffen oder zur Kenntnis von verschiedenen Lösungsstrategien. Sollte es gelingen, die bisherigen diagnostischen Testinstrumente, die oft vor dem Hintergrund einer Defizitorientierung entwickelt wurden (z.B. typische Fehler beim schriftlichen Addieren), durch spezifische Diagnosemodelle zu Kompetenzen in verschiedenen mathematischen Teilbereichen zu ergänzen, so wäre dies aus mathematikdidaktischer Sicht ein gewinnbringender Schritt.

2.4 Regelgeleitete Itementwicklung: Konstruktion von statistischen Textaufgaben: Anwendung von linear logistischen Testmodellen, Itemcloning und optimalem Design

Auch das Projekt *Regelgeleitete Itementwicklung* beschäftigt sich mit der Diagnose von Kompetenzen, diesmal konkretisiert am Beispiel von statistischen Textaufgaben für Lernende in der gymnasialen Oberstufe bzw. dem Studium. Das übergeordnete Ziel besteht in der Entwicklung eines automatischen Aufgabengenerators, der für eine Person während der Testphase adaptiv und in Echtzeit Testaufgaben zur Diagnose von unzureichend entwickelten Kompentenzkomponenten generiert. Die Aufgabengenerierung geschieht dabei einerseits regelgeleitet nach schwierigkeitsbestimmenden Merkmalen und andererseits durch Itemcloning, d.h. durch Änderung von Oberflächenmerkmalen bei Beibehaltung der schwierigkeitsbestimmenden Merkmale.

Auch bei diesem Projekt kommt es darauf an, auf Basis von Modellen der zu untersuchenden Kompetenz einen entsprechenden Input für den Aufgabengenerator zu leisten. Aus mathematikdidaktischer Perspektive wäre es interessant, inwieweit nicht nur sog. eingekleidete statistische Textaufgaben zu vergleichsweise wenig komplexen Kon-

texten generiert werden können, wie in dem Beitrag an Beispielen aufgezeigt wird, sondern ob auch der Bereich der realistischen Modellierungsaufgaben, welche als Operationalisierung der Kompetenz „Modellieren“ für die Leitidee „Daten und Zufall“ aufgefasst werden können, abgedeckt werden kann.

2.5 Co²CA: Nutzung und Auswirkung der Kompetenzmessung in mathematischen Lehr-Lern-Prozessen

Das Projekt Co²CA wendet sich der Frage zu, inwieweit Kompetenzmessungen zur Unterstützung von Lernprozessen im Mathematikunterricht verwendet werden können. In dem Projekt werden Kompetenzmessungen auf Basis eines Kompetenzmodells in Verbindung mit der individuellen Kompetenzrückmeldung untersucht. Auf Basis einer Stichprobe von 66 Schulklassen der 9. Jahrgangsstufe wurden zu drei Inhaltsbereichen sowie zu zwei Kompetenzbereichen Testaufgaben generiert und Daten erhoben.

Auf Grundlage der Kompetenzmessung wurden in einer experimentellen Studie drei Formen der individuellen Rückmeldung in parallelisierten Teilstichproben implementiert. Zum einen eine kriteriale Rückmeldung auf Basis von Kompetenzniveaus, zum zweiten eine prozessbezogene Rückmeldung, in der individuelle Lösungsprozesse kommentiert wurden und zum dritten eine Rückmeldung auf Basis der sozialen Bezugsnorm in der jeweiligen Klasse. Die Ergebnisse zur individuellen Zufriedenheit mit der Rückmeldung weisen positive Effekte zugunsten der kriterialen Rückmeldung im Vergleich zur Rückmeldung auf Basis der sozialen Bezugsnorm auf. Für die Form der prozessbezogenen Rückmeldung waren solche Effekte nicht nachzuweisen. Es wird vermutet, dass Letzteres ggf. auf das Fehlen jeglicher Vergleichsmöglichkeit der individuellen Leistungen zurückgeführt werden kann. Etwas problematisch an der Studie ist, dass zwischen Datenerhebung und Rückmeldung ein sehr langer Zeitraum von ca. 5–6 Monaten lag.

Co²CA zeigt eine einfache Möglichkeit auf, wie Verfahren zur Kompetenzmessung für eine direkte Nutzung im Mathematikunterricht verwendet werden können. Von großem Interesse ist dabei, inwieweit die individuellen Kompetenzrückmeldungen auch Einfluss auf die Kompetenzentwicklung der Schülerinnen und Schüler haben. Auch die Effekte einer kontinuierlichen und zeitnahen, d.h. lernbegleitenden Kompetenzrückmeldung wären ein lohnendes Untersuchungsziel.

3. Schlussbemerkung

Wie in der Einleitung zu Abschnitt 2 erwähnt, stehen die fünf diskutierten Projekte aus Sicht der Mathematikdidaktik in gewisser Weise exemplarisch für die gesamte Breite des Schwerpunktprogramms und den zu erwartenden Nutzen. Betrachtet man die Projekte als Grundlagenforschung, so ist einsichtig, dass in den meisten Fällen kaum mit unmittelbar verwendbaren Erkenntnissen zu rechnen ist. Deutlich wird bei der Diskus-

sion der einzelnen Beiträge, dass neben den innovativen psychometrischen Aspekten die domänenspezifische Analyse der Kompetenzstruktur ein wesentlicher Schlüssel für eine gewinnbringende Nutzung der berichteten und noch zu erwartenden Forschungsergebnisse sein wird. Diagnostische Kompetenzmodelle, Verfahren zur Kompetenzmessung, Rückmeldung von Kompetenzen – all dies ist in einer Domäne nur sinnvoll, wenn bekannt ist, was genau diagnostiziert, gemessen und rückgemeldet werden soll. Dies spricht dafür, die sich ergänzenden Expertisen von Wissenschaftlerinnen und Wissenschaftlern aus den Fachdidaktiken, der Bildungsforschung und der Psychometrie in Zukunft noch stärker zusammenzuführen. Schließlich gibt es in diesem Feld – das zeigen die diskutierten Projekte – noch viel zu tun.

Anschrift des Autors

Prof. Dr. Aiso Heinze, Abteilung Didaktik der Mathematik, Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN), Olshausenstraße 62, D-24098 Kiel
E-Mail: heinze@ipn.uni-kiel.de

Tobias Viering/Hans E. Fischer/Knut Neumann

Die Entwicklung physikalischer Kompetenz in der Sekundarstufe I

Projekt Physikalische Kompetenz[1]

1. Fragestellung und theoretischer Ansatz

In den Nationalen Bildungsstandards (NBS) für den Mittleren Schulabschluss in Physik (vgl. KMK 2005) sind die Bildungsziele des Faches für den Abschluss der Sekundarstufe I in Form von Kompetenzen formuliert (vgl. Klieme u.a. 2003). Um den Entwicklungsstand einzelner Schülerinnen und Schüler hinsichtlich der postulierten Bildungsziele im Verlauf der Sekundarstufe I feststellen zu können und damit eine individuelle Förderung zu ermöglichen, werden entsprechende Diagnoseinstrumente benötigt. Der Zusammenhang zwischen abstrakten Bildungszielen und konkreten Aufgaben in Diagnoseinstrumenten wird durch Kompetenzmodelle hergestellt (ebd.). Kompetenzmodelle, die Ausprägungen in verschiedenen Kompetenzbereichen beschreiben, werden als Kompetenzstrukturmodelle bezeichnet. Dagegen bilden Kompetenzentwicklungsmodelle ab, wie sich Kompetenzstrukturen verändern (vgl. Schecker/Parchmann 2006). Ziel des in diesem Beitrag vorgestellten Projekts ist es, ein Entwicklungsmodell theoretisch herzuleiten und empirisch zu prüfen. Grundlage ist ein bereits erprobtes Strukturmodell physikalischer Kompetenz.

1.1. Modellierung physikalischer Kompetenz

Für das Fach Physik in der Sekundarstufe I diskutierte Kompetenzstrukturmodelle gehen auf die Konkretisierung des Konzepts naturwissenschaftlicher Grundbildung durch Bybee (1997) zurück. Ausgehend von dieser Beschreibung leiten Klieme u.a. (2000) aus den TIMSS-Daten post hoc ein Modell naturwissenschaftlicher Kompetenz ab. Die vorgenommene Zuordnung der naturwissenschaftlichen Testaufgaben zu den postulier-

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: NE 1368/2-1, 2-2) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

ten Kompetenzniveaus konnte in einem Experten-Rating allerdings nicht bestätigt werden (vgl. Klieme 2000). Auch ein erneuter Versuch, im Rahmen von PISA 2000 ein Modell naturwissenschaftlicher Kompetenz post hoc zu beschreiben, schlug fehl: Wieder gelang die Zuordnung der Aufgaben zu den beschriebenen Niveaus nicht zufriedenstellend (vgl. Prenzel u.a. 2001, S. 202ff.). Im Rahmen der nationalen Zusatzerhebung PISA-E konnten zwar mit einem a priori entwickelten Kompetenzstrukturmodell verschiedene kognitive Teilkompetenzen zufriedenstellend unterschieden werden (vgl. Prenzel u.a. 2001, S. 225ff.), Ausprägungen in den Teilkompetenzen wurden jedoch norm- und nicht kriterienbezogen definiert (vgl. Neumann u.a. 2007).

Ein Modell naturwissenschaftlicher Kompetenz, das diese Beschränkung überwinden soll, wird von Schecker/Parchmann (2006) vorgeschlagen. Auf der Grundlage der NBS für die drei naturwissenschaftlichen Fächer führen sie die Ergebnisse fachdidaktischer Forschung zu einem für alle drei Fächer gültigen Modell in den Dimensionen *Inhaltsbereich, Prozess, Kontext, Ausprägung* und *Kognitive Anforderung* zusammen. Empirische Untersuchungen zeigen, dass sich die wesentlichen Komponenten des Modells bestätigen lassen, eine systematische Unterscheidung der Komponenten oder sogar eine kriterienorientierte Unterscheidung von Kompetenzniveaus bisher jedoch nicht möglich ist (vgl. Einhaus 2007; Schmidt 2008).

Ebenfalls bezogen auf die Vorgaben der NBS entwickelt Kauertz (2007), ausgehend von einem Modell der Vernetzung von Fachinhalten (vgl. Fischer u.a. 2006), ein sogenanntes Inhaltsstrukturmodell zur Erklärung der Schwierigkeit von Physikaufgaben in drei Dimensionen: *Leitidee*, *Kognitive Aktivität* und *Komplexität.* Die Dimension *Leitidee* umfasst die in den NBS zur Strukturierung des Kompetenzbereichs *Fachwissen* benannten Basiskonzepte *Energie*, *Wechselwirkung*, *System* und *Materie* (vgl. KMK 2005). Die Dimension *Kognitive Aktivität* bezieht sich auf kognitive Verarbeitungsstrategien, die als *Erinnern*, *Strukturieren* und *Explorieren* bezeichnet werden. Die Dimension *Komplexität* umfasst sechs hierarchisch geordnete Komplexitätsniveaus: *Ein Fakt* (1), *Mehrere Fakten* (2), *Ein Zusammenhang* (3), *Mehrere unverbundene Zusammenhänge* (4), *Mehrere verbundene Zusammenhänge* (5), *Übergeordnetes Konzept* (6). Als Fakten werden dabei kleinste physikalische Sinneinheiten, wie z.B. Beobachtungen bezeichnet. Als Zusammenhänge gelten mögliche Beziehungen zwischen Fakten. Mit *Übergeordnetes Konzept* ist gemeint, dass die Schülerin bzw. der Schüler über so viele Zusammenhänge zwischen Fakten verfügen kann, dass sich eine neue Qualität von Wissen herausgebildet hat, ein konzeptuelles Verständnis (vgl. Kauertz 2007). Die empirische Validierung des Modells zeigt: Für jede Leitidee besitzt die *Komplexität* (der erwarteten Lösung) einen schwierigkeitserzeugenden Einfluss; die einzelnen Leitideen wirken sich dabei jeweils unterschiedlich auf die Schwierigkeit aus. Ein Einfluss der kognitiven Aktivitäten auf die Schwierigkeit kann nicht nachgewiesen werden. Die Komplexitätsniveaus *Mehrere Fakten* und *Mehrere unverbundene Zusammenhänge* zeigen eine hohe Streuung bei der Schwierigkeit der jeweiligen Aufgaben, was wahrscheinlich in der stark variierenden Zahl der Fakten bzw. Zusammenhänge begründet ist (vgl. ebd.). Daher wurden diese Niveaus für Folgeuntersuchungen auf *Zwei Fakten* bzw. *Zwei Zusammenhänge* begrenzt und die Niveaus *Mehrere unverbundene Zusammenhänge*

und *Mehrere verbundene Zusammenhänge* zusammengefasst. Das so modifizierte Modell bildet den Kern des Modells, das zur Normierung der Nationalen Bildungsstandards für den Mittleren Schulabschluss genutzt wird (vgl. Walpuski u.a. 2008).

Kompetenzentwicklung wird in frühen Kompetenzstrukturmodellen (vgl. Bybee 1997; Klieme u.a. 2000; Prenzel u.a. 2001) als Erwerb von Fähigkeiten auf höheren Kompetenzniveaus angenommen. Da sich diese Kompetenzmodelle jedoch nur als eingeschränkt valide herausgestellt haben, wurde diese Annahme nicht weiter empirisch untersucht. Für die neueren, auf Grundlage der NBS entwickelten Modelle (vgl. Schecker/Parchmann 2006; Kauertz 2007; Walpuski u.a. 2008) wird Kompetenzentwicklung bisher nicht thematisiert. Die Feststellung von Schecker und Parchmann, dass empirisch „… bisher gar nicht geklärt ist, in welcher Weise und in welcher Verknüpfung sich die Ausprägungen naturwissenschaftlicher Kompetenz beim Individuum zeitlich entwickeln“ (Schecker/Parchmann 2006, S. 57), hat deshalb weiterhin Bestand.

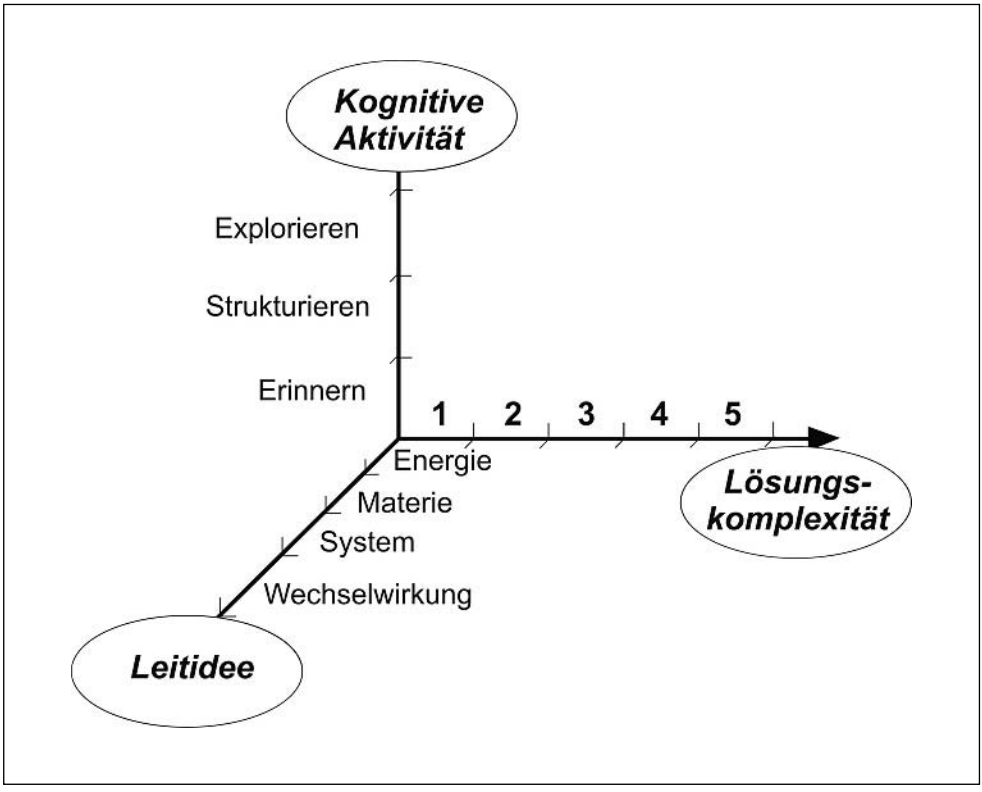


Abb. 1: Strukturmodell physikalischer Kompetenz

1.2 Entwicklung physikalischer Kompetenz

Als Ausgangspunkt für die theoretische Beschreibung der Entwicklung physikalischer Kompetenz wird das von Kauertz (2007) entwickelte und empirisch validierte Kompetenzstrukturmodell in der beschriebenen weiterentwickelten Form gewählt (s. Abb. 1).

Versteht man Kompetenzentwicklung in diesem Modell als das Fortschreiten von einem Kompetenzniveau zum nächsten, entspricht das einer stärkeren Vernetzung der Wissensbasis. Schülerinnen und Schüler verknüpfen Fakten zu Zusammenhängen und einzelne Zusammenhänge zu komplexen Netzwerken, die einem konzeptuellen Verständnis entsprechen (vgl. Neumann u.a. 2007). Gleichzeitig kann Kompetenzentwicklung aber auch auf jedem einzelnen Kompetenzniveau stattfinden: Schülerinnen und Schüler erwerben neue Fakten und stellen neue Zusammenhänge zu diesen Fakten her. Insgesamt vergrößert sich ihre Wissensbasis, sie wird differenzierter. Das konzeptuelle Verständnis, das mit dieser Wissensbasis verknüpft ist, verändert sich. Dies steht im Einklang mit Ansätzen der Beschreibung des Wandels begrifflichen Verständnisses (vgl. Wellman/Gelman 1998), wobei nicht das Verständnis einzelner Begriffe oder deren Vernetzung betrachtet wird, sondern vielmehr das Begriffsnetz selbst mit dem Verständnis eines physikalischen Konzepts gleichgesetzt wird.

Die Untersuchung des Verständnisses physikalischer Konzepte wird in der physikdidaktischen Forschung unter dem Begriff „Schülervorstellungen“ subsumiert. Der Schwerpunkt lag dabei bisher auf der Erfassung verschiedener Vorstellungen von physikalischen Konzepten (vgl. Vosniadou 2008). Besonders intensiv untersucht wurden zentrale Konzepte wie *Energie* (vgl. z.B. Duit 1986) oder *Materie* (vgl. z.B. Andersson 1990). Ausgehend von diesen Vorarbeiten analysieren neuere Untersuchungen, wie sich das Verständnis dieser Konzepte über die Schulzeit hinweg entwickelt: Für das Energiekonzept postulieren Liu und McKeough (2005) eine hierarchische Anordnung von vier inhaltsspezifischen Entwicklungsstufen, die für den mittleren Schulabschnitt relevant sind: *Energieformen und -quellen, Energieumwandlung und -transport, Energieentwertung* und *Energieerhaltung*. Durch Zuordnung der energiebezogenen Aufgaben der TIMSS-Untersuchung (vgl. Harmon u.a. 1997) zu diesen Entwicklungsstufen und durch Analyse der TIMSS-Daten gelangen Liu und McKeough (2005) zu dem Schluss, dass sich das Verständnis des Energiekonzepts entsprechend der von ihnen postulierten Hierarchie entwickelt. Liu und Lesniak (2006) analysieren die Entwicklung des Materiekonzepts auf der Grundlage von Daten, die dagegen mit einem spezifischen Testinstrument erhoben wurden. Sie beobachten eine Entwicklung des Materiekonzepts, die im Gegensatz zu der des Energiekonzepts für verschiedene Aspekte des Materiekonzepts parallel verläuft.

Um diese Ergebnisse zu berücksichtigen, wird das Kompetenzstrukturmodell (s. Abb. 1) zunächst um eine Dimension *Konzeptentwicklung* erweitert.

Bei der Erfassung des Entwicklungsstandes physikalischer Kompetenz durch Aufgaben ist zudem zu berücksichtigen, dass der Informationsgehalt des Aufgabentextes eine Rolle spielt. Es wird angenommen, dass einer Schülerin bzw. einem Schüler mit umfangreicher und vernetzter Wissensbasis weniger komplexe Informationen im Aufgabentext gegebenen werden müssen, damit er bzw. sie die Aufgabe auf dem geforderte Komplexi-

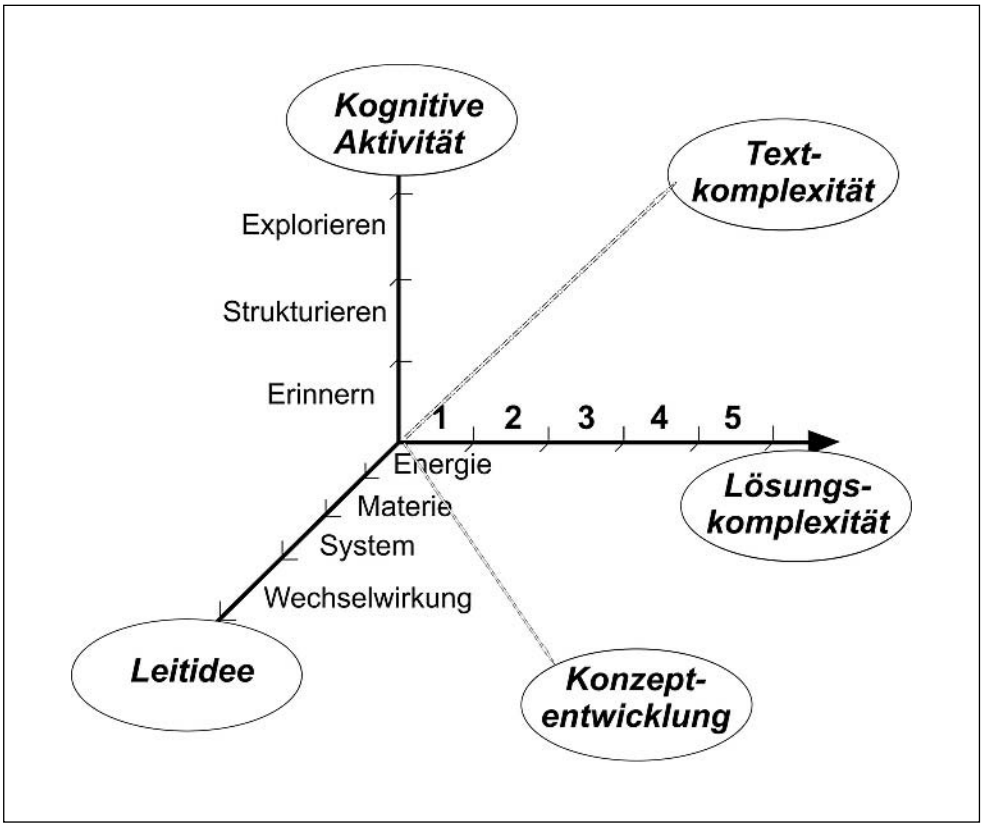


Abb. 2: Kompetenzentwicklungsmodell für den Bereich Fachwissen

tätsniveau lösen kann. Damit muss nicht nur der Aufgabenlösung sondern auch dem Aufgabentext ein spezifisches Komplexitätsniveau zugeordnet werden. Zusätzlich zur Dimension *Komplexität* – im Folgenden genauer als *Lösungskomplexität* bezeichnet – wird das Modell um die Dimension *Textkomplexität* erweitert. Es ergibt sich ein Entwicklungsmodell physikalischer Kompetenz in fünf Dimensionen (s. Abb. 2).

2. Forschungsdesign und Methoden

Zur empirischen Prüfung des vorgeschlagenen Kompetenzentwicklungsmodells (vgl. Abb. 2) muss das Modell durch Aufgaben operationalisiert werden. Dabei wird in diesem Projekt zunächst auf die Entwicklung des Verständnisses des Energiekonzepts fokussiert. Liu/McKeough (2005) unterscheiden die oben benannten vier Entwicklungsstufen des Energiekonzepts, wobei im Rahmen der Sekundarstufe I eine Entwicklung von einem Verständnis des Energiekonzepts auf dem Niveau von Energieformen und -quellen hin zu Energieerhaltung angenommen wird.

Da in dieser Untersuchung ausschließlich auf das konzeptuelle Verständnis der Schülerinnen und Schüler fokussiert werden soll, werden nur Aufgaben eingesetzt, bei denen die Komplexität der erwarteten Lösung dem Niveau *Übergeordnetes Konzept* entspricht. Die Lösungskomplexität der zu entwickelnden Aufgaben ist daher konstant. Der Einfachheit halber können Lösungskomplexität und Textkomplexität zur *Aufgabenkomplexität* als ein Maß für die Schwierigkeit der Aufgaben zusammengefasst werden: Die Aufgabenkomplexität ist damit als Differenz von Lösungskomplexität und Textkomplexität definiert. Durch Variation der Komplexität des Aufgabentextes kann die Aufgabenkomplexität die Werte 1 (maximale Information) bis 4 (minimale Information) annehmen. Die kognitive Aktivität wird auf *Explorieren* festgelegt. Die Beschränkung auf eine kognitive Aktivität hält den Aufwand für die Prüfung des Modells in einem vertretbaren Rahmen. *Explorieren* wird deshalb gewählt, weil es sich auf die Anwendung vorhandener Wissensstrukturen auf unbekannte Inhalte bezieht. *Erinnern* und *Strukturierung* bezeichnen dagegen die Reproduktion vorhandener Wissensstrukturen ohne weitere kognitive Verarbeitung bzw. die Umstrukturierung vorhandener Wissensstrukturen (vgl. Kauertz 2007). Der Entwicklungsstand der physikalischen Kompetenz einer Schülerin bzw. eines Schülers bezogen auf die Konzeptualisierung eines Basiskonzepts sollte sich entsprechend in Aufgaben ausdrücken, die *Explorieren* erfordern.

Das Entwicklungsmodell kann somit durch eine 4×4 – Matrix mit den Dimensionen *Konzeptentwicklung* und *Aufgabenkomplexität* repräsentiert werden, die durch Testaufgaben zu operationalisieren ist.

Für die Prüfung des vorgeschlagenen Kompetenzentwicklungsmodells lassen sich folgende Hypothesen ableiten:

- Dieselben Schülerinnen und Schüler lösen in höheren Jahrgangsstufen mit im Mittel gleicher Wahrscheinlichkeit schwierigere Aufgaben, d.h. Aufgaben auf höherer Entwicklungsstufe und Aufgabenkomplexität.
- Dieselben Schülerinnen und Schüler lösen in höheren Jahrgangsstufen mit höherer Wahrscheinlichkeit im Mittel gleich schwierige Aufgaben, d.h. Aufgaben auf gleicher Entwicklungsstufe und Aufgabenkomplexität.

Die Untersuchung ist als Längsschnitt von Jahrgangsstufe 6 bis Jahrgangsstufe 9 angelegt. Vorab werden die Aufgaben in einem Querschnitt mit N = 1200 Schülerinnen und Schülern der Jahrgangsstufen 6, 8 und 10 normiert. Dabei werden die kognitive Fähigkeit und Lesefähigkeit mit den Subskalen Q1 und N1 des KFT $4 - 12 + R$ (vgl. Heller/Perleth 2000) und des LGVT 6–12 nach Schlagmüller und Schneider (2007) kontrolliert. Die im Rahmen der Normierung getesteten Schülerinnen und Schüler der Jahrgangsstufe 6 werden in einem Längsschnitt in Jahrgangsstufe 8 und Jahrgangsstufe 9 jeweils noch einmal getestet. Dabei wird ihnen eine Auswahl der im Quasilängsschnitt normierten Aufgaben erneut vorgelegt. Auch kognitive Fähigkeiten und Lesefähigkeit werden erneut erfasst. Zusätzlich werden die auf Grundlage des Kompetenzmodells festgestellten Entwicklungsverläufe durch wiederholte strukturierte Interviews validiert.

Da den Aufgaben als Operationalisierung des zu prüfenden Kompetenzentwicklungsmodells eine zentrale Bedeutung zukommt, wird bei der Aufgabenentwicklung besonderes Augenmerk auf eine möglichst gute Modellpassung gelegt. Anhand einer Anleitung werden in einem Durchlauf jeweils 16 Aufgaben (entsprechend einer vollständigen Matrix) konstruiert (zur Verwendung von Konstruktionsanleitungen bei der Entwicklung von Aufgaben vgl. Kauertz 2007). Zu Beginn wird der Kontext festgelegt, z.B. ein Skater, der in einer Halfpipe hin und her rollt oder ein fliegendes Flugzeug. Eine Aufar-

Einem Auto geht während der Fahrt auf ebener Strecke das Benzin aus.

Dabei tritt Reibung auf. Dadurch wird die Bewegungsenergie des Autos in thermische Energie umgewandelt.

Warum bleibt ein Auto stehen, dem das Benzin ausgeht?

☐ Solange das Auto angetrieben wird, tritt keine Reibung auf. Sie setzt erst ein, wenn das Benzin ausgeht. Dann wird Bewegungsenergie in thermische Energie umgewandelt, bis das Auto stehen bleibt.

☐ Wenn das Auto ohne Benzin nicht mehr angetrieben wird, hat das Auto auch keine Bewegungsenergie mehr. Ohne Bewegungsenergie kann das Auto die Reibung nicht mehr überwinden und bleibt stehen.

☐ Durch Reibung wird Bewegungsenergie in thermische Energie umgewandelt. Wenn das Auto ohne Benzin nicht mehr angetrieben wird, nimmt die Bewegungsenergie immer weiter ab, bis das Auto schließlich stehen bleibt.

☐ Da das Auto ohne Benzin nicht mehr angetrieben wird, drehen sich die Räder immer langsamer. Wenn sie stillstehen, tritt Reibung auf, und die Bewegungsenergie wird in thermische Energie umgewandelt.

Abb. 3: Beispielaufgabe

beitung der Sachstruktur des gewählten Kontextes unter dem Basiskonzept Energie ist die Grundlage für die nächsten Schritte. Von der Beschreibung ausgehend werden getrennte Aufgaben für die einzelnen Stufen der Konzeptentwicklung konstruiert. Dabei unterscheiden sich Aufgaben zu einem bestimmten Kontext und einer bestimmten Entwicklungsstufe nur hinsichtlich ihrer Aufgabenkomplexität. Während also Situation, Fragestellung und vorgegebene Antwortalternativen unverändert bleiben, wird die Komplexität der lösungsrelevanten Information im Aufgabenstamm systematisch über alle vier Aufgabenkomplexitäten variiert. Ein spezieller Kontext liefert demnach 16 Aufgaben, für jede Kombination der vier Entwicklungsstufen und Aufgabenkomplexitäten eine. In Abbildung 3 ist beispielhaft eine Aufgabe der Entwicklungsstufe 3 (Energieentwertung) und der Aufgabenkomplexität 2 (ein Fakt und ein Zusammenhang in den zusätzlichen Informationen) aus dem Kontext *Auto* dargestellt.

Eine Pilotierungsstudie ergab erste statistisch relevante Informationen über die Schwierigkeit der Testaufgaben. Es wurden 32 ausgewählte Aufgaben, zwei aus jeder der 16 Zellen des Modells, mit N = 395 Schülerinnen und Schülern aus 15 Klassen der Jahrgänge 7 bis 11 an Gymnasien pilotiert. Die Aufgaben wurden auf zwei Testhefte zu je 20 Aufgaben verteilt; davon kamen in beiden Heften acht Aufgaben als Ankeraufgaben für eine Raschskalierung zum Einsatz. Zusätzlich wurden entsprechend des Forschungsdesigns die kognitiven Fähigkeiten und die Lesefähigkeit kontrolliert.

3. Ergebnisse und Diskussion

Die im Rahmen der Pilotstudie erhobenen Daten wurden auf ein dichotomes Raschmodell angepasst. Aufgaben mit einer Lösungshäufigkeit unter 15% oder über 85%, einem WMNSQ-Fitwert außerhalb des Intervalls [0,8; 1,2] oder einem *T*-Wert größer als 2,0 wurden aus der Analyse ausgeschlossen. Das betraf drei Aufgaben mit unzureichender Lösungshäufigkeit und eine Aufgabe mit zu hohem *T*-Wert. Mit den verbleibenden 28 Aufgaben ergab sich eine Reliabilität von $\alpha = 0{,}68$.

Beim Vergleich der Aufgabenschwierigkeiten in Abhängigkeit von den Entwicklungsstufen (s. Abb. 4) kann erwartungskonform ein Ansteigen der Aufgabenschwierigkeit mit den Entwicklungsstufen festgestellt werden. Drei Aufgaben, die der Entwicklungsstufe *Energieformen und -quellen* zugeordnet sind, weisen hingegen eine zu hohe Schwierigkeit auf. Eine Überprüfung dieser drei Aufgaben hat gezeigt, dass sie im Gegensatz zu den anderen Aufgaben dieser Stufe nicht allein das konzeptuelle Verständnis von Energieformen abfragen, d.h. ob sie ausschließlich prüfen, ob die Schülerin bzw. der Schüler in der Lage war, einer gegebenen Situation die relevante Energieform zuzuschreiben. Zur erfolgreichen Bearbeitung ist implizit zusätzlich eine Konzeptualisierung des Energiekonzepts auf der Stufe *Energieumwandlung* notwendig. Die Aufgaben operationalisieren das Modell also nicht adäquat und müssen entsprechend überarbeitet werden. Sie sind im Folgenden aus der Analyse ausgenommen.

Es ergibt sich ein statistisch bedeutsamer Zusammenhang zwischen Aufgabenschwierigkeit δ und Entwicklungsstufe λ von $\tau = 0{,}465$ ($p < 0{,}01$). Eine Varianzanalyse

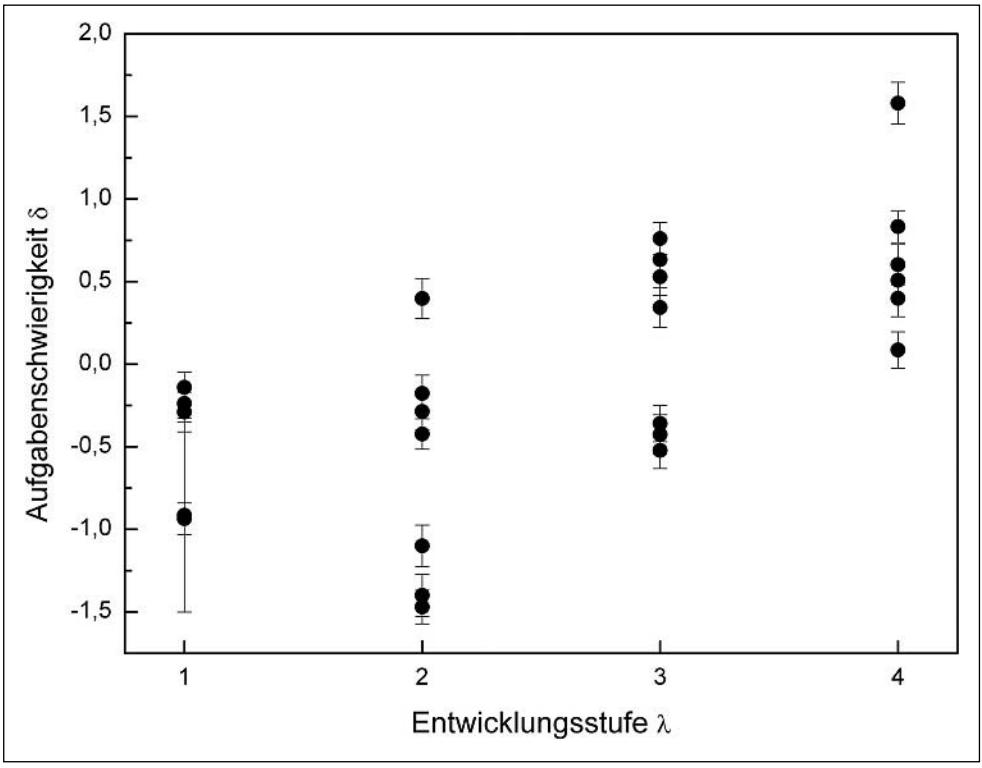


Abb. 4: Aufgabenschwierigkeiten und Entwicklungsstufen

liefert eine Varianzaufklärung von 50% ($F(3,21) = 7{,}09$; $p < 0{,}01$ (zweiseitig); $\eta^2 = 0{,}50$) durch die Entwicklungsstufen.

Abbildung 5 zeigt die Aufgabenschwierigkeit δ in Abhängigkeit von der Aufgabenkomplexität ξ. Im Gegensatz zu den Entwicklungsstufen zeigen sich hier keine statistisch bedeutsamen Zusammenhänge. Der Grund hierfür liegt möglicherweise in der verhältnismäßig kleinen Zahl von Aufgaben pro Stufe der Aufgabenkomplexität und in der relativ kleinen Stichprobe.

Schließlich wurde der Einfluss der Jahrgangsstufe J auf den Schätzer β des Fähigkeitsparameters unter Kontrolle kognitiver Fähigkeiten und der Lesefähigkeit untersucht. Die quantitativen Ergebnisse dazu sind in Tabelle 1 wiedergegeben. Es kann ein Anstieg der Schülerfähigkeit mit der Schulzeit konstatiert werden, der, wie erwartet, von kognitiven Fähigkeiten dominiert wird (vgl. Weinert/Helmke 1995). Der geringe Effekt der Lesefähigkeit weist auf gut verständliche Aufgabentexte hin, sodass ein hohes Leseverständnis keinen großen Vorteil beim Bearbeiten liefert. Ebenso spielt die Lesegeschwindigkeit aufgrund ausreichend gewährter Bearbeitungszeit augenscheinlich nur eine nebensächliche Rolle.

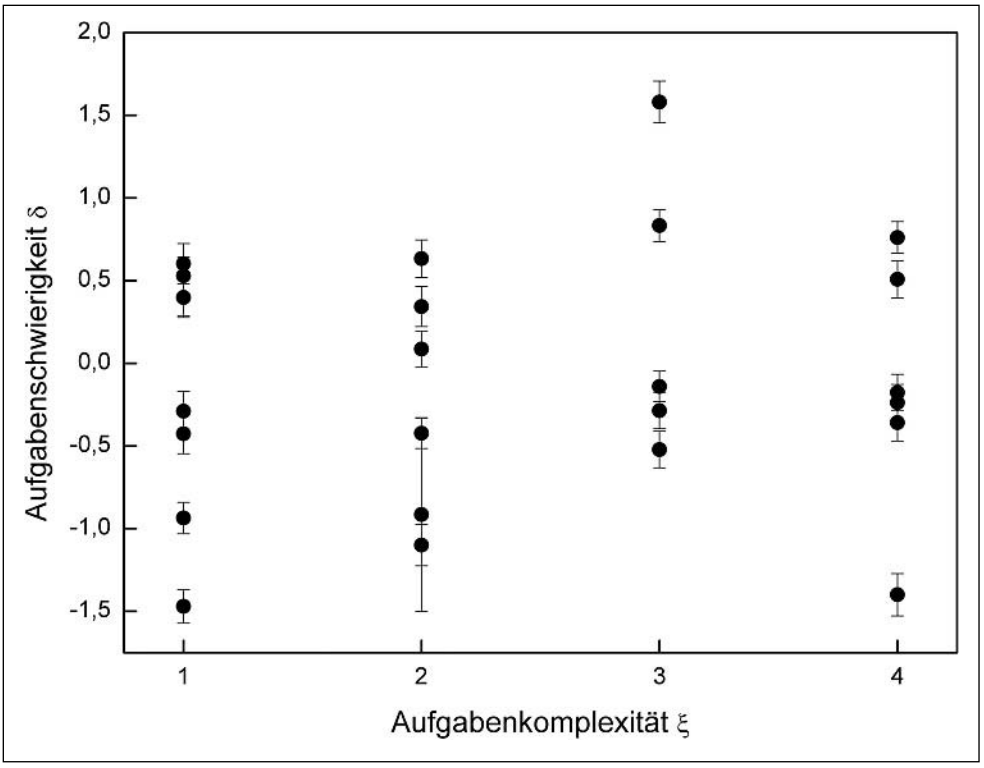


Abb. 5: Aufgabenschwierigkeiten und Aufgabenkomplexitäten

	F	**df**	**Sig.**	**part. η^2**
Korrigiertes Modell	17,54	7 (379)	$p < 0{,}01$	0,248
Konstanter Term	2,11	1	$p = 0{,}15$	0,006
Kognitive Fähigkeiten	32,84	1	$p < 0{,}01$	0,081
Leseverständnis	6,81	1	$p < 0{,}01$	0,018
Lesegeschwindigkeit	5,11	1	$p < 0{,}05$	0,014
Jahrgang J	4,20	4	$p < 0{,}01$	0,043

Tab. 1: Kovarianzanalyse der Schülerfähigkeit

Zusammenfassend lässt sich festhalten, dass die Testaufgaben, wie theoretisch erwartet, eine mit der Entwicklungsstufe wachsende Schwierigkeit zeigen. Ein Anwachsen der Aufgabenschwierigkeit mit der Aufgabenkomplexität konnte hingegen nicht nachgewiesen werden. Hier gilt es die Ergebnisse aus der Normierung mit einer angemessenen Stichprobengröße abzuwarten.

4. Erkenntnisgewinn

Im Rahmen des in diesem Beitrag beschriebenen Projekts wurde ein Modell der Entwicklung physikalischer Kompetenz für die Sekundarstufe I in den Dimensionen Konzeptentwicklung und Aufgabenkomplexität theoretisch begründet. Das Modell ist dabei so konstruiert, dass es prinzipiell anschlussfähig an die angrenzenden Bildungsabschnitte ist.

Eine erste Pilotierung ausgewählter Aufgaben bestätigt prinzipielle Annahmen des Modells. Dabei zeigte sich vor allem, dass Aufgaben höherer Entwicklungsstufen auch von Schülerinnen und Schülern höherer Jahrgänge erfolgreich bearbeitet werden konnten. Die umfassende Bestätigung des Modells an einer größeren Stichprobe findet im Sommer 2009 statt. Anschließend folgt die empirische Prüfung des Modells in einem echten Längsschnitt. In dessen Rahmen werden die in der ersten Phase in der 6. Jahrgangsstufe getesteten Schülerinnen und Schüler in den Jahrgängen 8 und 9 erneut getestet. Zusätzlich sollen die auf diese Art und Weise erfassten Entwicklungsverläufe durch strukturierte Interviews validiert werden.

Damit werden neben einem Beitrag zur Beschreibung und Erklärung der Entwicklung physikalischer Kompetenz vor allem auch Instrumente zur Diagnose verschiedener Entwicklungsstände und -verläufe einzelner Schülerinnen und Schüler bereitgestellt.

Literatur

Andersson, B.R. (1990): Pupils' conceptions of matter and its transformations (age 12–16). In: Lijnse, P.L./Licht, P./de Vos, W./Waarlo, A.J. (Hrsg.): Relating macroscopic phenomena to microscopic particles: A central problem in secondary Science Education. Utrecht: CD-Press, S. 12–35.

Bybee, R.W. (1997): Toward an understanding of scientific literacy. In: Gräber, W./Bolte, C. (Hrsg.): Scientific literacy, an international symposium. Kiel: IPN, S. 37–68.

Duit, R. (1986): Der Energiebegriff im Physikunterricht. Kiel: IPN.

Einhaus, E. (2007): Schülerkompetenzen im Bereich Wärmelehre. Berlin: Logos.

Fischer, H.E./Glemnitz, I./Kauertz, A./Sumfleth, E. (2006): Auf Wissen aufbauen – kumulatives Lernen in Chemie und Physik. In: Kircher, E./Girwidz, R./Häußler, P. (Hrsg.): Physikdidaktik. Theorie und Praxis. Heidelberg: Springer.

Harmon, M./Smith, T.A./Martin, M.O./Kelly, D.L./Beaton, A.E./Mullis, I.V.S. u.a. (1997): Performance Assessment in IEA's Third International Mathematics and Science Study. TIMSS International Study Center: Boston College.

Heller, K.A./Perleth, C. (2000): Kognitiver Fähigkeitstest für 4.–12. Klassen, Revision (KFT 4 − 12 + R). Göttingen: Hogrefe.

Kauertz, A. (2007): Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben. Berlin: Logos.

Klieme, E. (2000): Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretische Grundlagen, Kompetenzstufen und Unterrichtsschwerpunkt. In: Baumert, J./Bos, W./Lehmann, R. (Hrsg.): TIMSS/III Band 2. Opladen: Leske+Buderich, S. 57–128.

Klieme, E./Avenarius, H./Blum, W./Döbrich, P./Gruber, H./Prenzel, M. u.a. (2003): Expertise zur Entwicklung nationaler Bildungsstandards. Berlin: Bundesministerium für Bildung und Forschung (BMBF).

Klieme, E./Baumert, J./Köller, O./Bos, W. (2000): Mathematische und naturwissenschaftliche Grundbildung: Konzeptuelle Grundlagen und die Erfassung und Skalierung von Kompetenzen. In: Baumert, J./Bos, W./Lehmann, R. (Hrsg.): TIMSS/III Band 1. Opladen: Leske + Budrich, S. 85–133.

Liu, X./Lesniak, K. (2006): Progression in children's understanding of the matter concept from elementary to high school. In: Journal of Research in Science Teaching 43, S. 320–347.

Liu, X./McKeough, A. (2005): Developmental growth in students' concept of energy: Analysis from selected items from the TIMSS database. In: Journal of Research in Science Teaching 45, S. 493–517.

Neumann, K./Kauertz, A./Lau, A./Notarp, H./Fischer, H.E. (2007): Die Modellierung physikalischer Kompetenz und ihrer Entwicklung. In: Zeitschrift für Didaktik der Naturwissenschaften 13, S. 125–143.

Prenzel, M./Rost, J./Senkbeil, M./Häußler, P./Klopp, A. (2001): Naturwissenschaftliche Grundbildung: Testkonzeption und Ergebnisse. In: Baumert, J. u.a. (Hrsg.): PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske+Budrich, S. 191–248.

Schecker, H./Parchmann, I. (2006): Modellierung naturwissenschaftlicher Kompetenz. In: Zeitschrift für Didaktik der Naturwissenschaften 12, S. 45–66.

Schlagmüller, M./Schneider, W. (2007): Lesegeschwindigkeits- und -verständnistest für die Klassenstufen 6–12. Göttingen: Hogrefe.

Schmidt, M. (2008): Kompetenzmodellierung und -diagnostik im Themengebiet Energie der Sekundarstufe I. Berlin: Logos.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [KMK]. (2005): Bildungsstandards im Fach Physik für den Mittleren Schulabschluss. München: Luchterhand.

Vosniadou, S. (2008): Handbook of Research on Conceptual Change. Mahwah, NJ: Lawrence Erlbaum.

Walpuski, M./Kampa, N./Kauertz, A./Wellnitz, N. (2008): Evaluation der Bildungsstandards in den Naturwissenschaften. In: Der mathematische und naturwissenschaftliche Unterricht 61, S. 223–226.

Weinert, F.E./Helmke, A. (1995): Interclassroom differences in Instructional Quality and Interindividual Differences in Cognitive Development. In: Educational Psychologist 30, S. 15–20.

Wellman, H.M./Gelman, S.A. (1998): Knowledge acquisitation in foundational domains. In: Kuhn, D./Siegler, R.S. (Hrsg.): Handbook of child psychology. New York: Wiley, S. 523–573.

Anschriften der Autoren

Dipl.-Phys. Tobias Viering, Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik an der Universität Kiel, Didaktik der Physik, Olshausenstr. 62, D-24098 Kiel
E-Mail: viering@ipn-kiel.de

Dr. Hans E. Fischer, Universität Duisburg-Essen, Fachbereich Physik, Schützenbahn 70, D-45127 Essen
E-Mail: hans.fischer@uni-due.de

Dr. Knut Neumann, Leibniz-Institut für die Pädagogik der Naturwissenschaften (IPN) Kiel, Didaktik der Physik, Olshausenstraße 62, D-24098 Kiel
E-Mail: neumann@ipn.uni-kiel.de

Renate Soellner/Stefan Huber/Norbert Lenartz/Georg Rudinger

Facetten der Gesundheitskompetenz – eine Expertenbefragung

Projekt Gesundheitskompetenz[1]

1. Fragestellungen und theoretischer Ansatz

Der Bildungs- und Erziehungsauftrag von Schulen umfasst nicht nur die Vermittlung von Fachkompetenzen, sondern auch die Prävention und Förderung gesundheitsbezogener Kompetenzen (vgl. Schulgesetz NRW 2005, Erster Teil, §2 Absatz 5). Die große Bedeutung, die der Förderung gesundheitsbezogener Kompetenzen (kurz: Gesundheitskompetenz) im schulischen Kontext mittlerweile beigemessen wird, spiegelt sich in vielfältigen Programmen oder Wettbewerben zur ‚Gesunden Schule' wider, die in verschiedenen Bundesländern (z.B. Brandenburg, Berlin, Hessen) angeboten und u.a. von der Robert Bosch Stiftung gefördert werden.

Auch wenn das Konzept der Gesundheitskompetenz zunehmend von Wissenschaftler/-innen unterschiedlichster Disziplinen sowie von Verantwortlichen aus Politik und Gesellschaft aufgegriffen wird, besteht bis heute weder ein Konsens, wie der Begriff der Gesundheitskompetenz (engl. *health literacy*) zu definieren ist, noch eine Einigkeit bezüglich der Frage, welche Fähigkeiten und Fertigkeiten diese Kompetenz konstituieren (vgl. Soellner u.a. 2009). Generell lassen sich bei der Erforschung der Gesundheitskompetenz zwei Paradigmen unterscheiden (vgl. Pleasant/Kuruvilla 2008). Der *klinische* Ansatz versteht Gesundheitskompetenz als

> *„... the ability to read and comprehend prescription bottles, appointment slips, and the other essential health-related materials required to successfully function as a patient"* (American Medical Association 1999, S. 552).

Dieser Auffassung von Gesundheitskompetenz als „Gesundheits-Alphabetisierung" steht der *public-health*-Ansatz gegenüber, der erstmals von Nutbeam (2000) ausformuliert wurde. Er erweitert die eng gefasste Konzeption des klinischen Ansatzes um einen aktiven und konstruktiven Umgang mit gesundheitsbezogenen Informationen und versteht Gesundheitskompetenz als

> *„...the cognitive and social skills which determine the motivation and ability of individuals to gain access to, understand and use information in ways which promote and maintain good health"* (WHO 1998, S. 10).

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: SO 899/1-1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen" (SPP 1293).

Im Rahmen des *public-health*-Ansatzes schlägt Nutbeam (2000) ein Stufenmodell der Gesundheitskompetenz vor, das drei aufeinander aufbauende Formen der Gesundheitskompetenz postuliert: Die *funktionale* Gesundheitskompetenz umfasst basale kognitive Fähigkeiten, die Menschen in die Lage versetzen, grundlegende Informationen zur Gesundheit zu verstehen und in Handeln umzusetzen. Die *kommunikative, interaktive* Gesundheitskompetenz umfasst fortgeschrittene kognitive und soziale Fähigkeiten, die es dem Individuum erlauben, eine aktive Rolle im sozialen und gesellschaftlichen Umfeld einzunehmen. Die *kritische* Gesundheitskompetenz ermöglicht darüber hinaus die kritische Auseinandersetzung mit gesundheitsbezogenen Informationen und dem Gesundheitssystem. Diese Konzeption lässt zahlreiche Parallelen und Anknüpfungspunkte zum Kompetenzbegriff von Tippelt, Mandl und Straka (2003) erkennen, die Kompetenz in Fachkompetenz, methodisch-instrumentelle, sozial-kommunikative und personale Kompetenz sowie Basiswissen unterteilen.

Kritiker werfen dem Modell von Nutbeam vor, dass es lediglich altbekannte Konstrukte neu verpacke, ohne wirklich zur Klärung des Begriffes der Gesundheitskompetenz beizutragen (vgl. Tones 2002). Zudem wurde der stufenartige Aufbau des Modells bislang empirisch nicht überprüft. Lediglich für die unterste Stufe der funktionalen Gesundheitskompetenz liegen umfassende Forschungsarbeiten vor. Mittlerweile wurden zwar erste Versuche unternommen, auch die interaktive und kritische Form der Gesundheitskompetenz der empirischen Messung zugänglich zu machen (vgl. Steckelberg u.a. 2009; Ishikawa u.a. 2008), doch existieren bisher keine Erhebungsinstrumente, die in der Lage wären, das Konstrukt umfassend abzubilden. Will man das Konstrukt in all seinen Facetten empirisch erfassen, benötigt man zunächst ein Kompetenz*struktur*modell, welches diejenigen Fähigkeiten und Fertigkeiten ausweist, die notwendig sind, um gesundheitskompetent entscheiden und handeln zu können (vgl. Soellner u.a. 2009).

Das Projekt „Gesundheitskompetenz – Modellentwicklung und Validierung" verfolgt das Ziel, ein Modell zu entwickeln, das Aufschluss über die innere Struktur und Zusammensetzung der Gesundheitskompetenz geben sowie dessen Beziehungen zu verwandten Konstrukten kognitiver Art aufzeigen soll. Dabei soll der Begriff der Gesundheitskompetenz einem kognitiv-orientierten Kompetenzbegriff zugeführt werden (vgl. Weinert 2001). Kompetenzen werden dementsprechend als kontextspezifische kognitive Leistungsdispositionen verstanden, die sich funktional auf bestimmte Klassen von Situationen und Anforderungen beziehen lassen. Gesundheitskompetenz wird in diesem Zusammenhang als eine wissensbasierte Kompetenz betrachtet, die primär durch Kultur, Bildung und Erziehung vermittelt wird.

2. Methodisches Vorgehen

Auf Basis einer systematischen Literaturrecherche wurde zunächst eine Arbeitsdefinition entwickelt, die Gesundheitskompetenz als eine Sammlung von Fähigkeiten und Fertigkeiten versteht, über die jemand verfügen muss, um im Alltag und im Umgang mit dem Gesundheitssystem so handeln zu können, dass es sich positiv auf seine Gesundheit

und sein Wohlbefinden auswirkt. Diese Arbeitsdefinition bildete die Grundlage für eine mehrstufige Expertenbefragung, die nach der Methode des *concept mapping* durchgeführt wurde. *Concept mapping* bezeichnet im Allgemeinen Techniken zur strukturierten, visuell-räumlichen Darstellung von Wissen und Informationen, z.B. in Form von Hierarchien, Clustern oder netzartigen Strukturen (vgl. Cox 1999; Wiegmann u.a. 1992). Das Vorgehen der hier vorliegenden Expertenbefragung orientierte sich an der *concept-mapping*-Methode, wie sie Trochim (1989) vorschlägt. Dabei wird in einem ersten Schritt ein Brainstorming durchgeführt, mit dem Ziel möglichst viele Assoziationen zu einer Fokusfrage zu generieren, die den interessierenden Wissensbereich repräsentieren. Anschließend werden die Befragten gebeten, die gesammelten Aussagen so zu strukturieren, dass sich inhaltlich kohärente Kategorien ergeben. Zur Analyse der auf diese Weise gewonnenen Daten wird eine Kombination aus multidimensionaler Skalierung (MDS) und hierarchischer Clusteranalyse angewandt. Im Folgenden werden die einzelnen Schritte der Expertenbefragung und der Datenanalyse näher beschrieben.

Zur Expertenbefragung wurden sowohl Praktiker/innen aus dem Gesundheitswesen als auch Wissenschaftler/innen aus der Gesundheits- und Kompetenzforschung eingeladen.[2] Von den angeschriebenen 250 Expert(inn)en nahmen $N = 99$ am Brainstorming teil. Männer und Frauen waren in etwa zu gleichen Teilen vertreten (52,3% männlich) mit einem durchschnittlichen Alter von 45 Jahren ($SD = 9{,}5$; Min = 28, Max = 69). 48,8% der Teilnehmenden gaben als fachlichen Hintergrund Psychologie an, gefolgt von Expert(inn)en aus der Medizin (19,8%) und der Erziehungswissenschaft/Pädagogik (14%). Um eine möglichst umfassende Sammlung an Fähigkeiten und Fertigkeiten zu erhalten, die die Expert(inn)en mit der oben genannten Definition der Gesundheitskompetenz assoziieren, wurde ihnen folgende Fokusfrage vorgelegt:

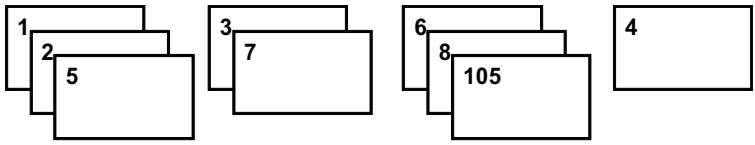
> *„Über welche Fähigkeiten und Fertigkeiten muss jemand verfügen, um im Alltag und im Umgang mit dem Gesundheitssystem so handeln zu können, dass es sich positiv auf seine Gesundheit und sein Wohlbefinden auswirkt?“*

Das Ergebnis des Brainstormings war eine Sammlung von 244 Aussagen zu Gesundheitskompetenz. Da viele dieser Aussagen mehr als einen inhaltlichen Aspekt enthielten, wurden sie zunächst in mehrere Einzelaussagen zerlegt, was zu einem Pool von 382 Aussagen führte. Zur Vorbereitung der Sortieraufgabe wurden diese 382 Aussagen um Redundanzen bereinigt und aufgrund theoretischer Überlegungen reduziert, sodass daraus ein Set von 105 Aussagen resultierte, das die Gesamtheit der im Brainstorming generierten Aussagen inhaltlich zufriedenstellend abbildete.

Zur Teilnahme an der nachfolgenden Sortier-Phase wurden erneut alle Expert(inn)en aus der Brainstormingphase eingeladen. Insgesamt 27 Personen beteiligten sich an der Sortieraufgabe. Das durchschnittliche Alter betrug 40 Jahre ($SD = 7{,}9$, Min = 25, Max = 57), 53,8% der Teilnehmenden waren männlich. Im Vergleich zur Brainstor-

2 Die Befragung wurde online mit Hilfe der Software CS Global 4.0 durchgeführt, welche von der Firma Concept Systems bezogen werden kann (www.conceptsystems.com).

ming-Phase war der Anteil an Psycholog(inn)en mit 81,5% deutlich erhöht. Die Sortieraufgabe konnte entweder online oder mit Hilfe von Karteikarten, die den Befragten auf Wunsch zugesandt wurden, bearbeitet werden. Die Aufgabe bestand darin, die 105 Aussagen zur Gesundheitskompetenz nach inhaltlichen Gesichtspunkten zu kategorisieren und diese Kategorien mit einem passenden Label zu versehen. Es wurden keine Sortierkriterien oder Kategorien vorgegeben, sondern lediglich darauf hingewiesen, dass mehr als eine Kategorie gebildet werden solle und eine Kategorie „Rest" bzw. „Sonstiges" nicht zulässig sei. Die Anzahl der Kategorien, die die Expert(inn)en bildeten, schwankte zwischen 4 und 25 ($M = 12{,}07$; $SD = 6{,}13$). Wurden zwei Aussagen von den Befragten in dieselbe Kategorie sortiert, wurde dem Aussagenpaar der Ähnlichkeitswert 1 zugewiesen, wurden sie unterschiedlichen Kategorien zugeordnet wurden, erhielt das Aussagenpaar den Ähnlichkeitswert 0. So ergab sich für jede/n Befragte/n eine binäre symmetrische Ähnlichkeitsmatrix ($m = n = 105$, siehe Abb. 1). Diese 27 individuellen Ähnlichkeitsmatrizen wurden nachfolgend auf Gruppenebene aggregiert, um so geteilte Wissensbestände abbilden zu können. Die daraus resultierende symmetrische Gruppenähnlichkeitsmatrix ($m = n = 105$) enthielt für jedes Aussagenpaar einen Wert zwischen 0 (min. Ähnlichkeit, d.h. keine/r der Befragten sortierte ein Aussagenpaar zusammen) und 27 (max. Ähnlichkeit, d.h. alle Befragten sortierten die Aussagen in einer Kategorie zusammen).



Item	1	2	3	4	5	6	7	8	…	105
1	1	1	0	0	1	0	0	…	…	0
2	1	1	0	0	1	0	0	…	…	0
3	0	0	1	0	0	0	1	…	…	0
4	0	0	0	1	0	0	0	…	…	0
5	1	1	0	0	1	0	0	…	…	0
6	0	0	0	0	0	1	0	…	…	1
7	0	0	1	0	0	0	1	…	…	0
8	0	0	0	0	0	1	0	…	…	1
…	…	…	…	…	…	…	…	…	…	…
105	0	0	0	0	0	1	0	…	…	1

Abb. 1: Beispielhafte Darstellung einer individuellen Sortierung (oben) und daraus resultierende binäre Ähnlichkeitsmatrix (unten)

Auf der Basis dieser aggregierten Ähnlichkeitsmatrix wurde anschließend eine nichtmetrische 2-dimensionale MDS durchgeführt. Die Interpretation der MDS-Lösung erfolgt beim *concept mapping* nach Trochim (1989) nicht – wie allgemein üblich – entlang der Dimensionen, sondern durch die nachträgliche Partitionierung des Raumes in Regionen konzeptuell zusammengehöriger Punkte (vgl. Borg/Staufenbiel 2007). In *CS Global 4.0* ist für die Regionalisierung standardmäßig eine hierarchische Clusteranalyse nach Ward auf der Grundlage der quadrierten euklidischen Distanz der *x/y* Koordinaten aus der MDS Lösung implementiert. Details zur Datenanalyse mittels *CS Global 4.0* finden sich bei Kane und Trochim (2007).

Um zu überprüfen, ob das Hinzufügen einer weiteren Dimension die Passung, die Interpretierbarkeit und die Kohärenz der Lösung verbessert, wurden die Daten zusätzlich mit dem in *SPSS 15.0* implementierten PROXSCAL-Algorithmus analysiert. Für die hierarchische Clusteranalyse, die im Anschluss an die 3-dimensionale MDS berechnet wurde, wurde das *average-linkage*-Verfahren als Fusionierungsalgorithmus gewählt, da das Ward-Verfahren die Bildung gleich großer Gruppen begünstigt, was hier nicht zwingend erschien.

3. Ergebnisse und Diskussion

Ein wichtiges Gütekriterium für die Bewertung von MDS-Lösungen ist der Stresswert *S*. Dieser gibt an, wie gut eine MDS-Konfiguration die Ursprungsdaten repräsentiert, wobei ein Wert von Null bedeutet, dass die Ursprungsdaten perfekt abgebildet werden. Je höher der Stresswert, desto schlechter ist die MDS-Konfiguration in der Lage, die Daten adäquat abzubilden (vgl. Borg/Staufenbiel 2007). Kruskal schlägt für die Beurteilung von Stresswerten folgende Unterteilung vor (vgl. Janssen/Laatz 2005, S. 571):

- $S \geq 0.2$ schlechte Übereinstimmung
- $0.2 \geq S \geq 0.1$ befriedigende Übereinstimmung
- $0.1 \geq S \geq 0.05$ gute Übereinstimmung
- $0.05 \geq S \geq 0.025$ hervorragende Übereinstimmung
- $0.025 \geq S \geq 0.00$ perfekte Übereinstimmung

Diese Wertebereiche stellen jedoch nur einen Orientierungsrahmen für die Bewertung von Stresswerten dar, da der Stresswert von einer ganze Reihe von Aspekten wie z.B. der Anzahl der zu repräsentierenden Objekte, der Anzahl an Dimensionen und dem Fehleranteil in den Daten abhängig ist. Beispielsweise erhöht sich der Stresswert mit der Anzahl *m* zu repräsentierender Objekte, weil die Zahl der abzubildenden Distanzen zwischen den Objekten fast quadratisch mit *m* ansteigt (vgl. Borg/Staufenbiel 2007). Für *concept-mapping*-Untersuchungen, die in der Regel einen großen *m*-Wert aufweisen, wird in einer Meta-Analyse von Trochim (1993) ein durchschnittlicher Stresswert von 0.285 ($SD = .004$) berichtet.

Die mit Global CS berechnete 2-dimensionale MDS-Lösung wies einen Stresswert von 0.23 auf, was einen guten Fit der Daten, verglichen mit dem bei *concept-mapping*-Untersuchungen zu erwartenden mittleren Stresswert, indiziert. Der Stresswert für die mit SPSS berechnete 3-dimensionale Lösung lag bei 0.015. Auch dieser Wert weist auf einen sehr guten Fit der Daten, entsprechend der Kategorisierung nach Kruskal, hin. Die Clusterlösung auf der Grundlage der 3-dimensionalen MDS-Lösung erwies sich jedoch eindeutiger interpretierbar als die auf der Grundlage der 2-dimensionalen Lösung. Folgende neun Cluster konnten identifiziert werden (siehe Abb. 2):

(1) die Fähigkeit zu Selbstregulation und Selbstdisziplin;
(2) die Fähigkeit zur Wahrnehmung der eigenen Bedürfnisse und Gefühle sowie ein hohes Körperbewusstsein;
(3) die Bereitschaft zur Verantwortungsübernahme für die eigene Gesundheit;
(4) gesundheitsbezogene Grundfertigkeiten, insbesondere die Fähigkeit, gesundheitsrelevante Texte lesen und verstehen zu können (*literacy*) und gesundheitsrelevante mathematische Aufgabenstellungen lösen zu können (*numeracy*);
(5) die Fähigkeit, gesundheitsrelevante Informationen angemessen interpretieren und nutzen zu können, wozu auch ein bestimmtes Maß an medizinisch-biologischem Grundwissen nötig ist;
(6) die Fähigkeit, sich gesundheitsrelevante Informationen beschaffen zu können;

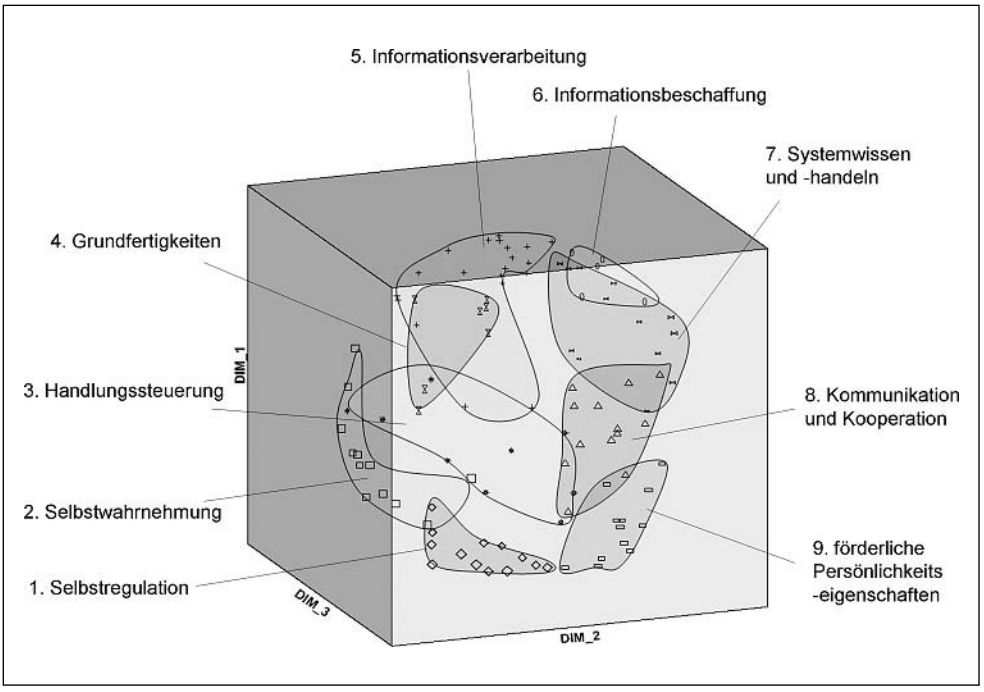


Abb.2: 3-dimensionales concept map

(7) die Fähigkeit, innerhalb des Gesundheitssystems navigieren und handeln zu können, sowie über das dazu notwendige Systemwissen zu verfügen;
(8) die Fähigkeit zur Kommunikation und Kooperation bezüglich gesundheitsrelevanter Inhalte;
(9) förderliche Persönlichkeitseigenschaften.

Auf der Grundlage dieses *concept maps* sowie theoretischer Überlegungen wurde ein erstes hypothetisches Strukturmodell der Gesundheitskompetenz entwickelt, das in Abbildung 3 wiedergegeben ist. Gesundheitskompetenz wird dabei als ein Netz aus grundlegenden Fertigkeiten (literacy/numeracy), Handlungskompetenz, Wissen und Motivation verstanden. Handlungskompetenz kann in die vier Kompetenzbereiche (1) Navigieren und Handeln im Gesundheitssystem, (2) Kommunikation und Kooperation, (3) Informationsbeschaffung und -verarbeitung sowie (4) Selbstwahrnehmung und Selbstregulation unterteilt werden. Die Wissenskomponente wird durch System- und Gesundheitswissen repräsentiert, wobei Systemwissen im *concept map* in Cluster (7) und Gesundheitswissen in Cluster (5) enthalten ist. Cluster (3), Bereitschaft zur Verantwortungsübernahme für die eigene Gesundheit, bildet den motivationalen Teil der Gesundheitskompetenz. Persönlichkeitseigenschaften sind im Sinne des Modells keine Strukturkomponenten.

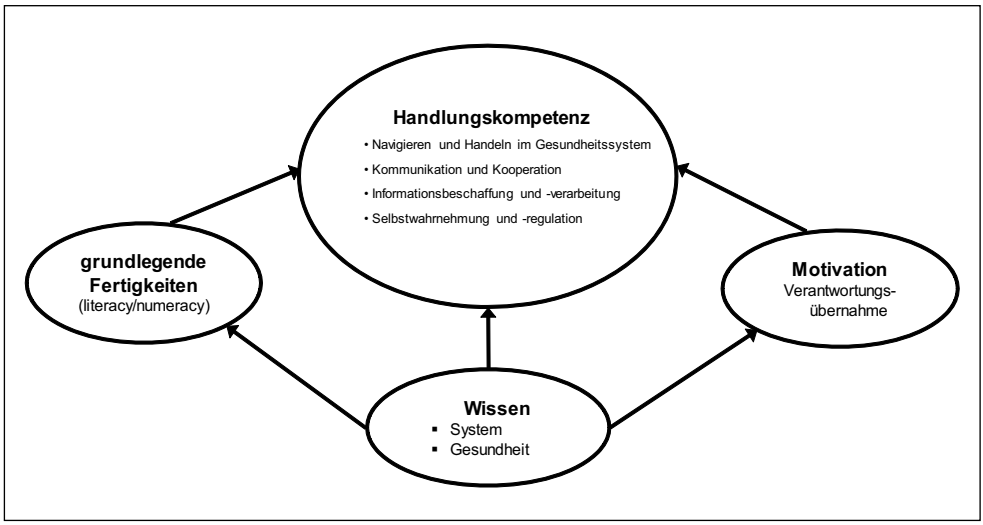


Abb. 3: Hypothetisches Strukturmodell der Gesundheitskompetenz

Die Komponenten des Modells sind kognitiv konzipiert, auch wenn einzelne Strukturkomponenten, insbesondere die motivationale Komponente zur Verantwortungsübernahme für die eigene Gesundheit den Rahmen eines engen Kognitionsbegriffs verlassen. Das hier dargestellte Strukturmodell stellt eine Erweiterung bereits existierender Modelle der Gesundheitskompetenz dar, insofern es vormals getrennt betrachtete Komponenten wie Grundfertigkeiten, Kommunikation, Informationsverarbeitung und Hand-

lungsbereitschaft (im Sinne der Verantwortungsübernahme für die Gesundheit) integriert und miteinander in Beziehung setzt (vgl. Nutbeam 2000; Schulz/Nakamoto 2005; Kriegesmann u.a. 2005; vgl. auch Soellner u.a. 2009). Eine derart ganzheitliche Betrachtung des Konzeptes unter Einbezug verschiedenster notwendiger Strukturkomponenten wurde bislang noch nicht geleistet. Im Vergleich mit bisherigen Modellen der Gesundheitskompetenz weist das in diesem Forschungsprojekt erarbeitete Modell wesentliche Besonderheiten auf:

(1) Das Modell wurde umfassend und systematisch unter Einbezug von Expert(inn)en aus dem Gesundheitsbereich und der Kompetenzforschung konzipiert.
(2) Das Konstrukt der Gesundheitskompetenz wurde auf Basis eines Kompetenzbegriffs entwickelt, welcher dem Modell einen klaren theoretischen Rahmen gibt.
(3) Es wurde ein Gesamtmodell der Gesundheitskompetenz geschaffen, welches über die heterogene Vielfalt gesundheitsrelevanter Situationen anwendbar sein soll, aber dennoch situations- und kontextspezifisch ausformuliert werden kann (i.S. eines Wechselspiels zwischen den Strukturkomponenten und den jeweiligen situations- und kontextspezifischen Anforderungen).
(4) Das vorgelegte Modell erweitert bisherige Modelle der Gesundheitskompetenz in seiner Gesamtheit und weist diese als Teilbereiche der Gesundheitskompetenz aus.
(5) Dabei ergänzt es bisherige Modelle, v.a. um die Kompetenzbereiche der Selbstregulation und Selbstwahrnehmung sowie um den motivationalen Aspekt zur aktiven Verantwortungsübernahme für die eigene Gesundheit.

Durch ein explizit situationsübergreifend formuliertes Modell und dem damit einhergehenden höheren Abstraktionsniveau sollte eine umfassende Konzeption relevanter Fähigkeiten und Fertigkeiten gewährleistet werden. Eine vorzeitige Einengung des Modells auf spezifische, eng umrissene Gesundheitssituationen sollte somit im Sinne der Neudefinition des Begriffs *health literacy* durch die WHO vermieden werden. Gleichzeitig gehen mit einer derart übergreifenden Betrachtung jedoch auch Nachteile einher:

> *„...The more general a competency or strategy (i.e. the greater the range of different types of situations to which it applies) the smaller the contribution of this competency or strategy to the solution of demanding problems"* (Weinert 2001, S. 53).

Damit zeigt sich ein Dilemma im Prozess der Modellierung der Gesundheitskompetenz: Wird der Begriff von Anfang an im Rahmen enger, situationsspezifischer Kontexte definiert, besteht die Gefahr, dass das jeweilige Modell keine Relevanz für eine Vielzahl heterogener, gesundheitsrelevanter Situationen und Anforderungen hat. Vielmehr müssten zahlreiche Submodelle der Gesundheitskompetenz entwickelt werden, die mehr oder weniger unverbunden nebeneinander stünden. Andererseits läuft ein konsequent situationsübergreifendes Modell aufgrund des hohen Abstraktionsgrades Gefahr, nur wenig Ansatzpunkte für eine empirische Überprüfbarkeit zu liefern.

Ein Ausweg aus diesem Dilemma bestand im Rahmen unseres Projektes darin, zunächst ein umfassendes, situationsübergreifendes Gesamtmodell der Gesundheitskompetenz zu entwickeln, welches möglichst alle, für die verschiedenen gesundheitsrelevanten Kontexte bedeutenden Kompetenzen ausweist, um in weiteren Arbeitsschritten dann die Struktur dieses Modells in konkreten, eng definierten Situationen zu prüfen und situationsspezifische Anforderungsprofile zu erarbeiten. Somit wird davon ausgegangen, dass die Fähigkeit für ein erfolgreiches und somit gesundheitskompetentes Handeln in verschiedenen Situationen durch verschiedene Strukturkomponenten determiniert wird. Eine solche Betrachtungsweise gewährleistet, dass das Modell alle für gesundheitsrelevantes Verhalten notwendigen kognitiven Kompetenzen ausweist. Gleichzeitig wird eine nachfolgende, kontextspezifische Ausformulierung des Gesamtmodells ermöglicht, welche für die Identifikation einzelner, in diesen Kontexten relevanter Kompetenzanteile notwendig ist.

4. Erkenntnisgewinn

Das im Rahmen des Projekts „Gesundheitskompetenz: Modellentwicklung und Validierung“ entwickelte Kompetenzstrukturmodell stellt eine Integration und Erweiterung bisheriger Modelle der Gesundheitskompetenz dar. Das Modell hilft die Facetten der Gesundheitskompetenz umfassend abzubilden und trägt damit zur Klärung des Begriffs bei. Es erweitert bisherige Gesundheitskompetenzmodelle um die Kompetenzbereiche Selbstregulation und Selbstwahrnehmung sowie um die Bereitschaft und Fähigkeit zur Verantwortungsübernahme. Das Modell lässt sich schlüssig interpretieren und ist mit anderen psychologischen Theorien in Einklang zu bringen. So decken die Komponenten Selbstregulation, Selbstwahrnehmung und die Verantwortungsübernahme für die eigene Gesundheit beispielsweise zentrale Konzepte sozialpsychologischer Theorien der Selbstregulation ab (vgl. Zimmermann 2005). Das Modell liefert darüber hinaus eine Grundlage für die Entwicklung von Kompetenzniveaumodellen der Gesundheitskompetenz und deren Erfassung mit Hilfe der Item-Response-Theorie.

Literatur

American Medical Association, Ad Hoc Committee on Health Literacy (1999): Health literacy: report of the Council on Scientific Affairs. In: Journal of the American Medical Association 281, S. 552–557. http://www.cmaj.ca/cgi/ijlink?linkType=ABST&journalCode=jama&resid=281/6/552 [24.07.2009].

Borg, I./Staufenbiel, T. (^{4}2007): Lehrbuch Theorien und Methoden der Skalierung. Bern: Hogrefe & Huber.

Cox, R. (1999): Representation construction, externalised cognition and individual differences. In: Learning and Instruction 9, S. 343–363.

Ishikawa, H./Nomura, K./Sato, M./Yano, E. (2008): Developing a measure of communicative and critical health literacy: A pilot study of Japanese office workers. In: Health Promotion International 23, S. 269–274.

Janssen, J./Laatz, W. ([5]2005): Statistische Datenanalyse mit SPSS für Windows: eine anwendungsorientierte Einführung in das Basissystem und das Modul exakte Tests. Berlin: Springer.

Kane, M./Trochim, W. (2007): Concept mapping for planning and evaluation. Thousand Oaks: Sage.

Kriegesmann, B./Kottmann, M./Masurek, L./Nowak, U. (2005): Kompetenz für eine nachhaltige Beschäftigungsfähigkeit. (Schriftenreihe der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin: Forschungsbericht, Fb 1038). Bremerhaven: Wirtschaftsverlag NW Verlag für neue Wissenschaft GmbH.

Nutbeam, D. (2000): Health literacy as a public health goal: A challenge for contemporary health education and communication strategies into the 21st century. In: Health Promotion International 15, H. 3, S. 259–267.

Pleasant, A./Kuruvilla, S. (2008): A tale of two health literacies: Public health and clinical approaches to health literacy. In: Health Promotion International 23, H. 2, S. 152–159.

Schulgesetz des Landes Nordrhein-Westfalen in der Fassung vom 15.02.2005. Erster Teil §2 Absatz 5.

Schulz, P./Nakamoto, K. (2005): Emerging themes in health literacy. In: Studies in Communication Sciences 5, H. 2, S. 1–10.

Soellner, R./Huber, S./Lenartz, N./Rudinger, G. (2009): Gesundheitskompetenz – ein vielschichtiger Begriff. In: Zeitschrift für Gesundheitspsychologie 17, H. 3, S. 105–113.

Steckelberg, A./Hülfenhaus, C./Kasper, J./Rost, J./Mühlhauser, I. (2009): How to measure critical health competences: Development and validation of the Critical Health Competence Test (CHC Test). In: Advances in Health Sciences Education 14, H. 1, S. 11–22.

Tippelt, R./Mandl, H./Straka, G. (2003): Entwicklung und Erfassung von Kompetenz in der Wissensgesellschaft – Bildungs- und wissenstheoretische Perspektiven. In: Gogolin, I./Tippelt, R. (Hrsg.): Innovation durch Bildung. Beiträge zum 18. Kongress der Deutschen Gesellschaft für Erziehungswissenschaft. Opladen: Leske+Budrich, S. 349–369.

Tones, K. (2002): Health literacy: New wine in old bottles? In: Health Education Research 17, H. 3, S. 287–290.

Trochim, W. (1989): An introduction to concept mapping for planning and evaluation. In: Evaluation and Program Planning 12, S. 1–16.

Trochim, W. (1993): Reliability of Concept Mapping. Vortrag gehalten bei der Annual Conference of the American Evaluation Association, Dallas, Texas, November, 1993.

Weinert, F.E. (2001): Concept of competence: A conceptual clarification. In: Rychen, D.S./Salganik, L.H. (Hrsg.): Defining and selecting key competencies. Göttingen: Hogrefe, S. 45–65.

Wiegmann, D.A./Dansereau, D.F./McCagg, E.C./Rewey, K.L./Pitre, U. (1992): Effects of knowledge map characteristics on information processing. In: Contemporary Educational Psychology 17, S. 136–155.

WHO (1998): Health promotion glossary. http://www.who.int/hpr/NPH/docs/hp_glossary_en.pdf [24.07.2009].

Zimmerman, B.J. (2005): Attaining self-regulation: a social cognitive perspectice. In: Boekaerts, M./Pintrich, P./Zeidner, M. (Hrsg.): Handbook of self-regulation. San Diego: Academic Press, S. 13–35.

Anschrift der Autor/innen

Prof. Dr. Renate Soellner, Universität Hildesheim, Institut für Psychologie,
Marienburger Platz 22, D-31141 Hildesheim
E-Mail: soellner@uni-hildesheim.de

Dipl.-Psych. Stefan Huber, Universität Hildesheim, Institut für Psychologie,
Marienburger Platz 22, D-31141 Hildesheim
E-Mail: stefan.huber@uni-hildesheim.de

Dipl.-Psych. Norbert Lenartz, Freie Universität Berlin, Arbeitsbereich Evaluation,
Qualitätssicherung und Qualitätsmanagement in Erziehungswissenschaft und Psychologie,
Habelschwerdter Allee 45, D-14195 Berlin
E-Mail: norbert.lenartz@fu-berlin.de

Prof. Dr. Georg Rudinger, Rheinische Friedrich-Wilhelms Universität Bonn,
Institut für Psychologie, Methodenlehre und Diagnostik, Kaiser-Karl-Ring 9, D-53111 Bonn
E-Mail: rudinger@uni-bonn.de

Ilonca Hardy/Thilo Kleickmann/Susanne Koerber/Daniela Mayer/
Kornelia Möller/Judith Pollmeier/Knut Schwippert/Beate Sodian

Die Modellierung naturwissenschaftlicher Kompetenz im Grundschulalter

Projekt Science-P[1]

Science-P ist ein im Rahmen des DFG-Schwerpunktprogrammes „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen" angesiedeltes Projekt mit dem Ziel der theoretischen Modellierung und psychometrischen Erfassung der naturwissenschaftlichen Kompetenzentwicklung von der zweiten bis vierten Klassenstufe. Dieser Beitrag thematisiert die theoretischen Grundlagen sowie die Entwicklung gruppentestfähiger Aufgaben in diesem Bereich.

1. Kompetenzmodelle in den Naturwissenschaften

Kompetenzmodelle in den Naturwissenschaften unterscheiden verschiedene Komponenten naturwissenschaftlicher Kompetenz (z.B. PISA 2006; vgl. Prenzel u.a. 2007). So differenzieren Duit, Häußler und Prenzel (2001) zwischen inhaltlich-konzeptuellen Komponenten, naturwissenschaftlichen Methoden und Denkweisen, Wissenschaftsverständnis (*Nature of Science*) und gesellschaftlichem Bezug. Die in der Vergangenheit postulierten Modelle wurden allerdings wegen ihrer überwiegend normativen Orientierung bzw. post-hoc ermittelter Kompetenzstufen als unzureichend kritisiert (vgl. z.B. Rost u.a. 2004). Angesichts der Forderung nach einer systematischen und modellbasierten Generierung von Testitems (vgl. Klieme u.a. 2003) und den Problemen bei der Operationalisierung und Überprüfung n-achsiger Kompetenzmodelle (vgl. z.B. Kauertz/Fischer 2006) fokussieren wir bei der Entwicklung des Kompetenzmodells zunächst auf die Modellierung von zwei ausgewählten Dimensionen naturwissenschaftlicher Kompetenz: 1) „naturwissenschaftliches Wissen" (d.h. inhaltlich/konzeptuelles Wissen) und 2) „Wissen über die Naturwissenschaften" (d.h. Wissen über naturwissenschaftliche Methoden und Wissenschaftsverständnis). Diese finden sich nicht nur in der Beschreibung der in der PISA 2006-Studie erfassten Kompetenzen und in Ansätzen zur *Scientific Literacy* wieder (vgl. Prenzel u.a. 2007); sie sind auch Bestandteil unterschiedlicher für den Grundschulbereich formulierter Standards (vgl. Gesellschaft für Didaktik des Sachunterrichts 2002).

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: MO 942/4-1/2, SO 213/29-1/2, SCHW 890/3-1/2) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen" (SPP 1293).

Das Grundschulalter wird in bisherigen Arbeiten nur selten berücksichtigt, und auch zur Dimensionalität naturwissenschaftlicher Kompetenz liegen kaum empirische Ergebnisse vor. Jedoch belegen Studien für das Sekundarschulalter funktionale Zusammenhänge zwischen Wissen über Naturwissenschaften und dem Erwerb naturwissenschaftlichen Wissens (vgl. z.B. Stathopoulou/Vosniadou 2007).

2. Kompetenzentwicklung als konzeptuelle Umstrukturierung

In Übereinstimmung mit Ergebnissen der Konzeptwechselforschung in unterschiedlichen Inhaltsgebieten und Altersgruppen (vgl. z.B. Vosniadou/Baltas/Vamvakoussi 2007) schlagen wir drei Kompetenzniveaus in den von uns untersuchten Dimensionen vor: (1) Naive Vorstellungen (Fehlvorstellungen), die einer empirischen Prüfung in unterschiedlichen Kontexten nicht standhalten. (2) Zwischenvorstellungen, mit denen Phänomene begrenzt erklärt werden können. (3) Wissenschaftliche Vorstellungen, die auf in der Wissenschaft geteilten Konzepten beruhen.

Diese Kompetenzniveaus können auch durch Erkenntnisse begrifflicher Entwicklung, die für naturwissenschaftliches Lernen relevant sind, begründet werden. So ziehen Kinder[2] beispielsweise zunächst sogenannte naive Konzepte (Niveau 1) zur Erklärung naturwissenschaftlicher Phänomenen heran, die im Laufe der Entwicklung eine fundamentale Umstrukturierung erfahren (vgl. z.B. Carey 1991). Dieses naive Wissen wird zunächst in vielen Alltagssituationen verstärkt, da es Kindern eine sinnvoll erscheinende Strukturierung und Vorhersage von Ereignissen ermöglicht. Naive Konzepte erfassen meist jedoch nicht oder nur teilweise die den Phänomenen zugrundeliegenden Mechanismen naturwissenschaftlicher Erklärungen. Naive Konzepte scheinen besonders im Hinblick auf außerschulische Kontexte und Transferkontexte schwer durch Unterricht veränderbar zu sein (vgl. Wandersee/Mintzes/Novak 1994). In vielen naturwissenschaftlichen Inhaltsgebieten können sogenannte Alltagsvorstellungen (Zwischenvorstellungen, Niveau 2) von naiven Vorstellungen unterschieden werden. Diese sind belastbare, ausbaufähige Vorstellungen zur Erklärung naturwissenschaftlicher Phänomene, wie beispielsweise das Materialkonzept im Kontext des „Schwimmen und Sinkens“ von Gegenständen (vgl. Hardy u.a. 2006; Tytler 2000). Erst durch Unterricht werden auch in der Grundschule erste wissenschaftliche Vorstellungen aufgebaut (Niveau 3). In vielen Fällen resultiert aus naturwissenschaftlichem Unterricht jedoch auch fragmentiertes oder nicht integriertes Wissen, welches das simultane Halten verschiedener Vorstellungen beinhaltet. Damit kann ein zusätzliches, höheres Kompetenzniveau (Niveau 3+) in der Integration von Vorstellungen im Sinne der simultanen Ablehnung von naiven Vorstellungen und der Annahme von wissenschaftlichen Vorstellungen gesehen werden.

Auch in Studien zum Wissenschaftsverständnis können die drei beschriebenen Niveaus unterschieden werden. Auf Niveau 1 (naive Vorstellung) wird Wissenschaft als

2 Bei der nachfolgenden Nennung von Personen sind jeweils beide Geschlechter gemeint.

Aktivität bzw. als objektivistisches Sammeln von Fakten verstanden, auf einer etwas höheren Ebene (Niveau 2, Zwischenvorstellung) als Herstellung einfacher kausaler Zusammenhänge. Auf Niveau 3 (wissenschaftliche Vorstellung) sehen Schüler Wissenschaft als Suche nach Erklärungen und wissenschaftliches Wissen als das Ergebnis der Prüfung von Hypothesen. Auf einem weiteren Niveau (das im Grundschulalter meist nicht erreicht wird) wird der zyklische und kumulative Charakter der Bildung, Prüfung und Revision von Theorien erkannt (vgl. Carey u.a. 1989; Sodian u.a. 2006).

2.1 Inhalte der Dimension naturwissenschaftliches Wissen

Die Mehrheit der Primarstufenlehrpläne fokussiert bei den naturwissenschaftlichen Themen auf den konzeptuell anspruchsvollen Bereich „Materie" (z.B. Luft, Verdunsten/Kondensieren) und auf Themen wie Magnetismus, Auftriebs- und Gewichtskraft unter dem Aspekt der Wechselwirkung. Wir konzentrieren uns daher auf die Inhaltsgebiete „Schwimmen und Sinken" und „Verdunstung/Kondensation".

Schwimmen und Sinken: Als Vorläufer des Verständnisses von Schwimmen und Sinken können ein Verständnis von Materie und der Gewichtsbegriff angesehen werden (vgl. Carey 1991). Naive Erklärungen zum Schwimmen und Sinken im Grundschulalter sind häufig durch eine eindimensionale Fokussierung auf Aspekte des Objekts bzw. der Luft gekennzeichnet und somit nicht mit der das Objekt umgebenden Flüssigkeit verbunden (vgl. z.B. Tytler/Peterson 2004). Erst durch Instruktion werden fortgeschrittenere Erklärungsansätze mit Dichtevergleich und Auftriebskraft auch im Grundschulalter systematisch in verschiedenen Kontexten angewendet (vgl. Hardy u.a. 2006).

Verdunstung und Kondensation: Der Zusammenhang zwischen Verdunstung und Kondensation erfordert eine Vorstellung unterschiedlicher Zustände (Übergang von flüssig zu gasförmig). Naive Erklärungen zum Verdunsten geben häufig an, dass Wasser einfach verschwindet bzw. vom Boden absorbiert wird (vgl. z.B. Tytler 2000). Die Vorstellung, dass Wasser „nach oben" transferiert wird, beinhaltet hingegen bereits den Ansatz der Konservierung von Masse. Erst mit ca. 12 Jahren wird eine „Umwandlung von flüssigem Wasser in gasförmiges Wasser" (bzw. Auflösung in Teilchen) als Erklärung für die Verdunstung angegeben und eine Unterscheidung in Bezug auf die Zustandsform (flüssig/gasförmig) getroffen.

2.2 Inhalte der Dimension Wissen über Naturwissenschaften

Wissen über Naturwissenschaften umfasst a) die epistemologischen Überzeugungen über die Natur wissenschaftlichen Wissens und b) das Verständnis naturwissenschaftlicher Methoden (z.B. Datengewinnung, Dateninterpretation). Wissen über Naturwissenschaften wird, wenn überhaupt, erst im Sekundarschulunterricht gelehrt. Jedoch wird in der aktuellen naturwissenschaftsdidaktischen Debatte die Notwendigkeit der Einbet-

tung dieser Thematik in den breiteren Kontext eines adäquaten Verständnisses der Konstruktion naturwissenschaftlichen Wissens betont (vgl. Windschitl/Thompson/Braaten 2008) und durch neuere entwicklungspsychologische Befunde unterstrichen.

Wissenschaftsverständnis (Nature of Science). Das Wissenschaftsverständnis von Schülern der Sekundarstufe I entspricht meist einer unreflektierten epistemologischen Position, die durch die mangelnde Differenzierung zwischen Theorien/Hypothesen einerseits und empirischer Evidenz andererseits sowie durch ein unzureichendes Verständnis des zyklischen und kumulativen Charakters naturwissenschaftlichen Wissens gekennzeichnet ist (vgl. Carey u.a. 1989). Neuere Studien zum Wissenschaftsverständnis zeigen allerdings, dass selbst das Verstehensniveau von Grundschulkindern durch Interventionen eines wissenschaftstheoretisch orientierten Unterrichts signifikant angehoben werden kann und dass dieser Unterricht außerdem Effekte auf die Fähigkeit zur Produktion kontrollierter Experimente hat (vgl. Sodian u.a. 2006).

Methodenkompetenzen. Im Mittelpunkt unseres Interesses stehen das intuitive Verständnis experimenteller Designs (z.B. Variablenkontrollstrategien), sowie die Fähigkeit zur Interpretation von Kovariationsdaten im Hinblick auf eine zu evaluierende Hypothese. In neuerer entwicklungspsychologischer Forschung wurde ein grundlegendes Verständnis von Experimentierstrategien, Hypothesenprüfung und Evidenzevaluation im Vor- und Grundschulalter gefunden (vgl. Zimmerman 2007). So verstehen bereits Erstklässler den Unterschied zwischen der Produktion und dem Testen von Effekten und zeigen damit ein Grundverständnis der Hypothesenprüfung (vgl. Sodian/Zaitchik/Carey 1991) und bei fünf- bis siebenjährigen Kindern konnte ein rudimentäres Hypothese-Evidenz Verständnis nachgewiesen werden (vgl. z.B. Koerber u.a. 2005). Weiterhin sind bereits bei Drittklässlern Experimentierstrategien mit langfristigem Erfolg trainierbar (vgl. Strand-Cary/Klahr 2008).

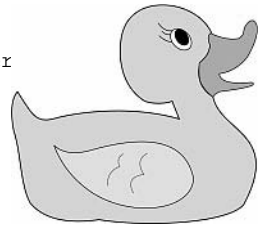
Obwohl es also Hinweise darauf gibt, dass im Grundschulalter sowohl im Bereich des naturwissenschaftlichen Wissens als auch im Bereich des Wissens über Naturwissenschaften grundlegende Umstrukturierungen von naiven Vorstellungen durch geeigneten Unterricht erfolgen können, ist die empirische Forschungslage insbesondere zur langfristigen individuellen Entwicklung von Vorstellungen im Bereich der Naturwissenschaften defizitär. Längsschnittuntersuchungen an größeren Stichproben mit gruppentestgeeigneten Instrumenten fehlen. Das aufwändige methodische Vorgehen in Interviewstudien wird zwar durch die Einbettung in sozio-kulturelle Theorien der Konzeptentwicklung von manchen Autoren gerechtfertigt (vgl. z.B. Mason 2007), bringt aber häufig mit sich, dass Entwicklungsverläufe lediglich anhand von querschnittlichen Vergleichen mit geringen Stichprobenzahlen postuliert werden. Ein zentrales Anliegen unseres Projekts ist deshalb die Konstruktion von gruppentestfähigen Instrumenten zur Erfassung der Dimensionen *naturwissenschaftliches Wissen* und *Wissen über Naturwissenschaften*, welche psychometrischen Gütekriterien entsprechen und zur längsschnittlichen Modellierung naturwissenschaftlicher Kompetenz im Grundschulalter eingesetzt werden können.

3. Die Erfassung naturwissenschaftlicher Kompetenz mit grundschulgemäßen Testverfahren

Bei der Entwicklung von Testverfahren und der empirischen Testung des postulierten Kompetenzmodells gehen wir nach dem von Wilson (2005) vorgeschlagenen Ansatz des *construct modeling* vor. Dieser Ansatz beschreibt verschiedene produktive und reflektive Schritte bei der Entwicklung eines Messinstrumentes. Das zu messende Konstrukt wird als kontinuierliche, latente Variable verstanden, bei der verschiedene qualitative Niveaus unterschieden werden. Ausgangspunkt der Instrumentenentwicklung bildet eine sog. *construct map*, in der die Ausprägungen des Konstruktes auf den Kompetenzniveaus spezifiziert werden. Die Einzelitems der *construct map* sind individuelle Realisierungen ihrer Zellinhalte. Bei der Ausarbeitung der *construct map* konnte zum einen auf die Ergebnisse eigener Forschung zurückgegriffen werden (vgl. z.B. Hardy u.a. 2006), zum anderen wurden relevante Schülervorstellungen aus der fachdidaktischen

Gummiente

Instruktion des Testleiters:
Ich habe hier eine Gummiente. Mit so einer Gummiente kann man in der Badewanne spielen. Wie ihr seht,schwimmt die Gummiente oben auf dem Wasser (Demonstration!).



Woran liegt es, dass die Gummiente schwimmt?

Kreuze nach jeder Antwort ‚Richtig' oder ‚Falsch' an!

	Richtig	Falsch
1. Die Gummiente schwimmt, weil sie innen hohl ist.	☐	☐
2. Die Gummiente schwimmt, weil das Wasser sie nach oben drückt.	☐	☐
3. Die Gummiente schwimmt, weil sie sehr leicht ist.	☐	☐
Was ist die beste Antwort?	Nr. _____	

Abb. 1: Multiple-Select-Aufgabe mit anschließendem Multiple-Choice für die Komponente Schwimmen und Sinken. Antwortalternative 1 entspricht einer Zwischenvorstellung, Antwort 2 einer wissenschaftlichen und Antwort 3 einer naiven Vorstellung.

Forschung wie aus der entwicklungspsychologischen Forschung zusammengestellt (vgl. z.B. Bar/Galili 1994; vgl. Zimmerman 2007). Des Weiteren wurden relevante Kindervorstellungen in offenen Interviewfragen für die Formulierung der Antwortalternativen gesammelt.

Herr Müller baut Flugzeuge und möchte, dass sie möglichst wenig Treibstoff verbrauchen. Jetzt hat er verschiedene Ideen, wovon der Treibstoffverbrauch abhängen könnte:	
Er denkt: Ein Flugzeug kann eine runde Nase oder eine spitze Nase haben.	runde Nase  spitze Nase 
Er denkt: Die Höhenruder können oben oder unten angebracht werden.	Höhenruder oben  Höhenruder unten 
Er denkt: Ein Flugzeug kann doppelte Flügel oder einfache Flügel haben.	doppelte Flügel  einfache Flügel 
Herr Müller überlegt: Es könnte daran liegen, ob die **Höhenruder oben oder unten** angebracht werden.	

Was soll Herr Müller tun, um herauszufinden, ob die Stellung der Höhenruder wichtig oder egal für den Treibstoffverbrauch ist?

Kreuze die <u>beste</u> Antwort an!

1. Herr Müller muss ein paar Flugzeuge bauen und schauen, ob sie wenig Treibstoff verbrauchen.	
2. Herr Müller muss zwei Flugzeuge bauen, eines mit dem Höhenruder oben und eines mit dem Höhenruder unten. Sie müssen aber sonst gleich sein.	
3. Herr Müller muss zwei ganz unterschiedliche Flugzeuge bauen, bei denen er Nase, Flügel und Höhenruder unterschiedlich macht.	

Abb. 2: Multiple-Choice-Aufgabe für die Komponente Methodenkompetenz. Antwortalternative 1: Naive Vorstellung (Produktion von Effekten), Antwort 2: wissenschaftliche Vorstellung (kontrolliertes Experiment); Antwort 3: Zwischenvorstellung (kontrastiver Test).

Das *Item Design* betrifft die verschiedenen Aufgabenformate, wie *Forced-Choice-Aufgaben*, welche die Wahl der besseren von zwei vorgegebenen Antwortalternativen erfordern und damit auf den kritischen Übergang zwischen zwei Kompetenzniveaus fokussieren. Neben Aufgaben mit offenem Antwortformat gibt es einen weiteren Aufgabentypus, in dem eine wissenschaftliche Erklärung und mehrere naive Erklärungen separat als richtig oder falsch zu beurteilen sind. Dies zielt auf die Integration des Verständnisses, also auf die Frage, ob neben der Annahme der wissenschaftlichen Erklärung auch gleichzeitig die Ablehnung der naiven Erklärungen geleistet werden kann. Bei Multiple-Select-Aufgaben müssen drei Antwortalternativen, die jeweils eines der drei postulierten Kompetenzniveaus repräsentieren, getrennt beurteilt werden. Im Anschluss ist zusätzlich die beste Alternative auszuwählen (Multiple-Choice-Antwort). Dieses Format deckt folglich simultan die gesamte Spannbreite der drei Kompetenzniveaus ab. Bei der Formulierung von Aufgabenstämmen für die Dimension *naturwissenschaftliches Wissen* konnten für die Komponente *Schwimmen und Sinken* einige Aufgabenstämme aus Tests übernommen werden (vgl. Hardy u.a. 2006). Zur Itementwicklung für die Komponente *Verdunstung und Kondensation* wurden weitere eigene Vorarbeiten herangezogen. Die entlehnten Aufgabenstämme wurden überarbeitet und an die besonderen Erfordernisse der Testdurchführung mit Zweitklässlern angepasst. Weitere Aufgabenstämme wurden vollständig neu konzipiert (vgl. Abbildung 1).

Zur Komponente *Wissenschaftsverständnis* wurden Aufgabenstämme sowohl zum abstrakt-deklarativen als auch zum kontextualisierten Wissenschaftsverständnis konstruiert. Dabei dienten Vorarbeiten aus Interviewstudien als Orientierung (vgl. Carey u.a. 1989). Beispiele für kontextualisierte Items sind Interpretationskonflikte zwischen zwei Wissenschaftlern, z.B. über die Genese einer Krankheit (Anlage vs. Umwelt Theorien). Die Aufgaben basieren z.T. auf dem sogenannten „Hexerei-Interview" der LOGIK-Studie (vgl. z.B. Bullock/Sodian/Koerber 2009). Weitere Aufgaben wurden in Anlehnung an das „Nature Nurture Interview" (vgl. Thoermer/Sodian 2002) entwickelt. Für die Komponente *Methodenkompetenz* wurden aufbauend auf entwicklungspsychologischen Vorarbeiten Aufgabenstämme zu Experimentierstrategien konstruiert. Dazu zählen die Unterscheidung zwischen Hypothesenprüfung und Effektproduktion, die Wahl eines konklusiven Tests, eines kontrastiven bzw. kontrollierten Experiments sowie das Verständnis verschiedener Aspekte des experimentellen Designs. Eine Beispielaufgabe ist in Abbildung 2 dargestellt.

Bei der Durchführung der Tests wird eine ausreichende Betreuung durch Assistenten insbesondere für die Zweitklässler gewährleistet. Nach einer Erläuterung des Untersuchungszwecks wird durch die gemeinsame Bearbeitung von Beispielaufgaben sichergestellt, dass alle Probanden das Antwortformat verstehen, wobei alle Aufgaben jeweils von der Versuchsleitung vorgelesen werden und durch Projektion der behandelten Aufgabe sowie ggf. kurzen Demonstrationen begleitet werden.

4. Itempilotierung und Messmodell

Die Überarbeitung der Aufgaben erfolgte in einem gestuften Prozess: Zu Beginn standen qualitative Aufgabenerprobungen mit kleinen Stichproben, worauf eine Prä-Pilotierung der entwickelten Aufgaben beider Dimensionen an einer Stichprobe von $N = 1274$ folgte. Die Bearbeitungszeiten und Beobachtungen hinsichtlich der Durchführung bestätigten die Machbarkeit von Gruppentestungen in beiden Altersgruppen der zweiten und vierten Klasse. Darüber hinaus wurden die psychometrischen Eigenschaften der Einzelitems und 21 Testhefte für beide Dimensionen getrennt ausgewertet (vgl. hierzu Kleickmann u.a. im Druck; Koerber u.a. im Druck). Beispielsweise wurden bei Kleickmann u.a. (im Druck) bei einer Substichprobe von $N = 100$ Kindern bei einem Testheft Hinweise sowohl auf adäquate Itemschwierigkeiten und Reliabilitäten gefunden als auch darauf, dass Antworten auf einem höheren Niveau tatsächlich häufiger bei Viert- als bei Zweitklässlern auftreten. In Validierungsstudien (Ergebnisse liegen derzeit noch nicht vor) wird zudem untersucht, inwieweit frei produzierte Schülerantworten mit der Aufgabenbearbeitung im Fragebogen zusammenhängen. Dabei bearbeiten jeweils ca. 70 Drittklässler ein Testheft aus Items der Prä-Pilotierung sowie Fragen im Einzelinterview, um die konvergente Validität dieser Aufgaben zu prüfen. Zur Ermittlung der diskriminanten Validität werden weitere Variablen wie kognitive Fähigkeit und Leseverständnis erfasst. Nach Auswertung der Prä-Pilotierungsstudie nach psychometrischen Eigenschaften und Augenscheinvalidität resultierte aus einem Itempool von insgesamt 167 Aufgaben mit 370 Einzelitems eine Anzahl von $N = 244$ Items für die Dimension naturwissenschaftliches Wissen und $N = 126$ Items für die Dimension Wissen über Naturwissenschaften, die in einer Querschnittserhebung in der zweiten, dritten und vierten Klasse mit $N = 900$ Kindern eingesetzt werden sollen, um die Dimensionalität der beiden Bereiche zu überprüfen. Dabei erfassen wir als Kontrollvariablen die kognitive Fähigkeit, das Leseverständnis, die Schulleistung in den Fächern Sachunterricht, Deutsch und Mathematik (Noten) und den sozioökonomischen Status sowie den Migrationshintergrund der Kinder.

Wie unter Punkt 2 ausgeführt, umfasst das empirisch zu prüfende Modell naturwissenschaftlicher Kompetenz in der Grundschule ein Struktur- und ein Niveau-Modell. Sowohl für die Dimension *naturwissenschaftliches Wissen* als auch für die Dimension *Wissen über Naturwissenschaften* werden drei hierarchisch geordnete Kompetenzniveaus postuliert: Naive Vorstellungen, Zwischenvorstellungen und wissenschaftliche Vorstellungen. Die statistische Modellierung soll anhand einer Variante des Rasch-Modells (Partial-Credit-Modell; Rost u.a. 2004) erfolgen. Die Annahme einer hierarchischen Gliederung naturwissenschaftlicher Kompetenz in drei Niveaus wird geprüft, indem die empirischen Itemschwierigkeiten bzw. die Schwellenparameter im Falle der Partial-Credit-Aufgaben zu den im Kompetenzmodell postulierten Kompetenzniveaus in Beziehung gesetzt werden. Obwohl von einer Überlappung der Kompetenzniveaus auszugehen ist, werden für beide Kompetenzdimensionen substantielle Korrelationen zwischen Itemschwierigkeiten und Kompetenzniveaus erwartet.

Durch die Anwendung eines Raschmodells in Verbindung mit einem Multi-Matrix-Design wird es möglich, die Daten zu den beiden Kompetenzdimensionen aufeinander

zu beziehen und die Annahme zu prüfen, dass konzeptuelles Wissen in grundlegenden Inhaltsbereichen und formales Wissen zwei Dimensionen naturwissenschaftlicher Kompetenz im Grundschulalter bilden. Dies wird anhand von ein- und mehrdimensionalen Raschmodellen erfolgen. Dabei vermuten wir Zusammenhänge zwischen den beiden Dimensionen insbesondere für das obere Kompetenzniveau. Wegen der anzunehmenden Bereichsspezifität konzeptuellen naturwissenschaftlichen Wissens wird in weiterführenden Analysen geprüft, ob die beiden Inhaltsbereiche Schwimmen und Sinken und Verdunstung/Kondensation tatsächlich zwei Dimensionen bilden oder besser eindimensional modelliert werden. Diese Analysen werden insbesondere Aufschluss geben über den in der Entwicklungspsychologie und der Fachdidaktik diskutierten Zusammenhang von bereichsspezifischem und bereichsübergreifendem Wissen im Grundschulalter.

Literatur

Bar, V./Galili, I. (1994): Stages on children's views about evaporation. In: International Journal of Science Education 16, H. 2, S. 157–174.

Bullock, M./Sodian, B./Koerber, S. (2009): Doing experiments and understanding science: Development of scientific reasoning from childhood to adulthood. In: Schneider, W./Bullock, M. (Hrsg.): Human development from early childhood to early adulthood. Findings from the Munich Longitudinal Study. Mahwah, NJ: Erlbaum, S. 173–197.

Carey, S. (1991): Knowledge acquisition: Enrichment or conceptual change? In: Carey, S./Gelman, R. (Hrsg.): The epigenesis of mind: Essays on biology and cognition. Hillsdale, NJ: Lawrence Erlbaum Associates.

Carey, S./Evans, R./Honda, M./Jay, E./Unger, C. (1989): An experiment is when you try it and see if it works. A study of junior high school students' understanding of the construction of scientific knowledge. In: International Journal of Science Education 11, S. 514–529.

Duit, R./Häußler, P./Prenzel, M. (2001): Schulleistungen im Bereich der naturwissenschaftlichen Bildung. In: Weinert, F. (Hrsg.): Leistungsmessungen in Schulen. Weinheim: Beltz, S. 169–186.

Gesellschaft für Didaktik des Sachunterrichts (2002): Perspektivrahmen Sachunterricht. Bad Heilbrunn: Klinkhardt.

Hardy, I./Jonen, A./Möller, K./Stern, E. (2006): Effects of instructional support within constructivist learning environments for elementary school students' understanding of „Floating and Sinking". Journal of Educational Psychology 98, H. 2, S. 307–326.

Kauertz, A./Fischer, H.E. (2006): Assessing students level of knowledge and analysing the reasons for learning difficulties in physics by Rasch Analysis. In: Xiufeng, L./Boone, W.E. (Hrsg.): Applications of Rasch Measurement in Science Education. USA: Jam Press, S. 212–246.

Kleickmann, T./Hardy, I./Möller, K./Pollmeier, J./Tröbst, S. (im Druck): Die Modellierung naturwissenschaftlicher Kompetenz im Grundschulalter: Theoretische Konzeption und Testkonstruktion. In: Zeitschrift für Didaktik der Naturwissenschaften.

Klieme, E./Avenarius, H./Blum, W./Döbrich, P./Gruber, H./Prenzel, M./Reiss, K./Riquarts, K./Rost, J./Tenorth, H.-E./Vollmer, H. (2003): Expertise zur Entwicklung nationaler Bildungsstandards. Berlin: BMBF.

Koerber, S./Sodian, B./Thoermer, C./Nett, U. (2005): Scientific reasoning in young children: Preschoolers' ability to evaluate covariation evidence. In: Swiss Journal of Psychology 64, H. 3, S. 141–152.

Koerber, S./Sodian, B./Kropf, N./Mayer, D./Schwippert, K. (im Druck): Die Entwicklung des wissenschaftlichen Denkens im Grundschulalter: Theorieverständnis, Experimentierstrategien, Dateninterpretation. In: Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie.

Mason, L. (2007): Introduction: Bridging the cognitive and sociocultural approaches in research on conceptual change: Is it feasible? In: Educational Psychologist 42, H. 1, S. 1–7.

Prenzel, M./Artelt, C./Baumert, J./Blum, W./Hammann, M./Klieme, E./Pekrun, R. (2007): PISA 2006 in Deutschland – Die Ergebnisse der dritten internationalen Vergleichsstudie. Münster: Waxmann.

Rost, J./Prenzel, M./Carstensen, C.H./Senkbeil, M./Groß, K. (2004): Naturwissenschaftliche Bildung in Deutschland – Methoden und Ergebnisse von PISA 2000. Wiesbaden: Verlag für Sozialwissenschaften.

Sodian, B./Jonen, A./Thoermer, C./Kircher, E. (2006): Die Natur der Naturwissenschaften verstehen – Implementierung wissenschaftstheoretischen Unterrichts in der Grundschule. In: Prenzel, M./Allolio-Näcke, J. (Hrsg.): Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms. Münster: Waxmann, S. 147–160.

Sodian, B./Zaitchik, D./Carey, S. (1991): Young children's differentiation of hypothetical beliefs from evidence. In: Child Development 6, S. 753–766.

Stathopoulou, C./Vosniadou, S. (2007): Exploring the relationship between physics-related epistemological beliefs and physics understanding. In: Contemporary Educational Psychology 32, S. 255–281.

Strand-Cary, M./Klahr, D. (2008): Developing elementary science skills: Instructional effectiveness and independence. In: Cognitive Development 23, H. 4, S. 488–511.

Thoermer, C./Sodian, B. (2002): Science undergraduates' and graduates' epistemologies of science: The notion of interpretive frameworks. In: New Ideas in Psychology 20, S. 263–283.

Tytler, R. (2000): A comparison of year 1 and year 6 students' conceptions of evaporation and condensation: Dimensions of conceptual progression. In: International Journal of Science Education 22, H. 5, S. 447–467.

Tytler, R./Peterson, S. (2004): From „try it and see" to strategic exploration: Characterizing young children's scientific reasoning. In: Journal of Research in Science Teaching 41, H. 1, S. 94–118.

Vosniadou, S./Baltas, A./Vamvakoussi, X. (2007): Re-framing the conceptual change approach in learning and instruction. Amsterdam: Elsevier Science.

Wandersee, J./Mintzes, J./Novak, J. (1994): Research on alternative conceptions in science. In: Gabel, D. (Hrsg.): Handbook of Research on Science Teaching and Learning. New York: Macmillan Publishing Company, S. 177–210.

Wilson, M. (2005): Constructing Measures. An item-response modelling approach. Mahwah: Lawrence Erlbaum.

Windschitl, M./Thompson, J./Braaten, M. (2008): Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. In: Science Education 92, H. 5, S. 941–967.

Zimmerman, C. (2007): The development of scientific thinking skills in elementary and middle school. In: Developmental Review 27, H. 2, S. 172–223.

Anschrift der Autor/innen

Prof. Dr. Ilonca Hardy, Goethe-Universität Frankfurt, Senckenberganlage 15,
D-60054 Frankfurt am Main
E-Mail: hardy@em.uni-frankfurt.de

Dr. Thilo Kleickmann, Max-Planck- Institut für Bildungsforschung, Lentzeallee 94,
D-14195 Berlin
E-Mail: kleickmann@mpib-berlin.mpg.de

Prof. Dr. Susanne Koerber, Pädagogische Hochschule Freiburg, Fakultät 1, Institut für
Psychologie, Kunzenweg 15, D-79117 Freiburg
E-Mail: susanne.koerber@ph-freiburg.de

Dipl. Psych. Daniela Mayer, Ludwig-Maximilians-Universität München,
Lehrstuhl für Entwicklungspsychologie, Leopoldstr. 13, D-80802 München
E-Mail: daniela.mayer@psy.lmu.de

Prof. Dr. Kornelia Möller, Seminar für Didaktik des Sachunterrichts, Fachbereich Physik,
Westfälische Wilhelms-Universität Münster, Leonardo Campus 11, D-48149 Münster
E-Mail: sachunterricht@uni-muenster.de

Dipl. Psych. Judith Pollmeier, Seminar für Didaktik des Sachunterrichts, Fachbereich Physik,
Westfälische Wilhelms-Universität Münster, Leonardo Campus 11, D-48149 Münster
E-Mail: j.pollmeier@uni-muenster.de

Prof. Dr. Knut Schwippert, Sektion 1: Allgemeine, Interkulturelle und International
vergleichende Erziehungswissenschaft, Universität Hamburg, Binderstraße 34,
D-20146 Hamburg
E-Mail: knut.schwippert@uni-hamburg.de

Prof. Dr. Beate Sodian, Ludwig-Maximilians-Universität München,
Lehrstuhl für Entwicklungspsychologie, Leopoldstr. 13, D-80802 München
E-Mail: beate.sodian@psy.lmu.de

Nina Roczen/Florian G. Kaiser/Franz X. Bogner

Umweltkompetenz – Modellierung, Entwicklung und Förderung

Projekt Umweltkompetenz[1]

Bislang wird im Rahmen von Umweltbildung zumeist auf eine Förderung ökologieunspezifischer, allgemeiner Fähigkeiten, zum Beispiel der Befähigung zu kritischem Denken oder zum Problemlösen (vgl. De Haan 2006; Kyburz-Graber 2004) abgezielt. Da diese allgemeinen Fähigkeiten in der Regel theoretisch abgeleitet sind und empirisch bislang keine Verhaltenswirksamkeit nachgewiesen werden konnte, sind sie für die Handlungskompetenz der/des Einzelnen vergleichsweise irrelevant. Im Zentrum unseres Forschungsprojekts steht dagegen die Entwicklung eines auf empirisch bestätigt verhaltenswirksamen, ökologiespezifischen Fähigkeiten basierendes Strukturmodell der individuellen Umweltkompetenz sowie der Vergleich von systematisch geförderten und spontanen Entwicklungsverläufen dieser Umweltkompetenz und der mit ihr verbundenen ökologiespezifischen Fähigkeiten.

Wir möchten zunächst die theoretischen Hintergründe unseres Umweltkompetenzmodells vorstellen. Anschließend werden wir auf unsere konkreten Forschungsfragen sowie das daraus abgeleitete Forschungsdesign eingehen. Schließlich stellen wir die bisher durchgeführten (Vor-)Arbeiten dar und diskutieren mögliche methodische, theoretische sowie praktische Erkenntnisgewinne.

1. Umweltkompetenz – Theoretischer Hintergrund

Das oberste Ziel von Bildung ist erfolgreiches Verhalten und Problemlösen im individuellen Alltag und weniger die bloße Aneignung reinen Faktenwissens oder das erfolgreiche Bestehen eines Leistungstests (vgl. McClelland 1973; OECD 2003). Aus dieser Grundidee speist sich die Definition von Kompetenzen als diejenigen Fähigkeiten und Merkmale, die man zum erfolgreichen Handeln im Alltag benötigt (vgl. Weinert 2001). Übertragen auf die Umweltbildung heißt das konkret, dass Bildungsmaßnahmen auf die gezielte Förderung von ökologischem Handeln hinwirken müssen. Um solchermaßen „zielgerichtetes ökologisches Verhalten“ zu verbessern, sollten diejenigen Fähigkeiten identifiziert und gefördert werden, die erwiesenermaßen zu diesen Handlungen befähigen und motivieren (vgl. Kaiser/Roczen/Bogner 2008). Dieser Logik folgend haben wir

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: BO 944/5-1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

zum einen Umwelthandlungskompetenz zum Zielkriterium erhoben, zum anderen ließen wir diejenigen intellektuellen und motivationalen Fähigkeiten in unser Modell einfließen, die empirisch nachweisbar positiv mit Umwelthandeln zusammenhängen und die sich zudem eignen, im Rahmen der Umweltbildung gefördert zu werden.

Umweltwissen ist eine intellektuelle Fähigkeit, die klassischer Gegenstand von Umweltbildungsmaßnahmen und erwiesenermaßen eine notwendige Vorbedingung von Umwelthandeln ist. Als motivationale Komponente, die den oben genannten Kriterien der Verhaltenswirksamkeit sowie der Formbarkeit durch Bildung entspricht, erachten wir die individuelle „Verbundenheit mit der Natur".

1.1 Umwelthandlungskompetenz

Umwelthandlungskompetenz verstehen wir als Disposition zu „zielgerichtetem ökologischem Verhalten": Je ausgeprägter diese Handlungsdisposition ist, desto mehr setzt eine Person auch aufwändige Verhaltensmittel zur Realisierung ihrer persönlichen Umweltschutzziele ein (vgl. z.B. Kaiser/Wilson 2004). Psychometrisch kann eine so verstandene Umwelthandlungskompetenz mit dem Raschmodell modelliert werden. Die entsprechenden Indikatoren umfassen Verhaltensweisen aus sechs unterschiedlichen ökologie-relevanten Bereichen (vgl. Energiesparen, Mobilität, Müllvermeidung, Recycling, Konsumverhalten und indirektes Umweltverhalten). Die auf diesem Prinzip entwickelte und mehrfach angewendete Skala erwies sich als ein Instrument, auf dessen Grundlage sich die Disposition zu allgemein ökologischem Handeln bei Jugendlichen und bei Erwachsenen reliabel und valide erfassen lässt (vgl. z.B. Kaiser/Wilson 2004; Kaiser/Oerke/Bogner 2007).

1.2 Umweltwissen

Umweltwissen wird allgemein als eine notwendige, jedoch nicht hinreichende Bedingung für die Entwicklung von Umwelthandlungskompetenz betrachtet (vgl. Gardner/Stern 2002; Schultz 2002b). Obwohl Umweltwissen aus motivationaler Sicht als nicht besonders relevant für individuelles Verhalten angesehen wird (vgl. Hines/Hungerford/Tomera 1986/87), könnte es indirekt Verhalten inspirieren, indem es Bewusstsein schafft und Gründe für ökologisches Verhalten liefert. Wir unterscheiden drei verschiedene Arten von Umweltwissen: Umweltsystem-, Handlungs- und Wirksamkeitswissen. (1) *Umweltsystemwissen* entspricht Wissen über die geltenden Zusammenhänge in Ökosystemen sowie über Ursachen von Umweltproblemen. Ein typisches Beispiel für diese Form des Wissens ist die Kenntnis der atmosphärischen Auswirkungen von CO_2. (2) *Handlungswissen* umfasst sowohl Wissen über mögliche Handlungsoptionen als auch konkrete Handlungsausführungen. Ein typisches Beispiel ist Wissen bezüglich der richtigen Art der Mülltrennung. (3) *Wirksamkeitswissen* bezieht sich auf die Kenntnis des Umweltschutzpotentials unterschiedlicher Verhaltensweisen. So haben beispielsweise der

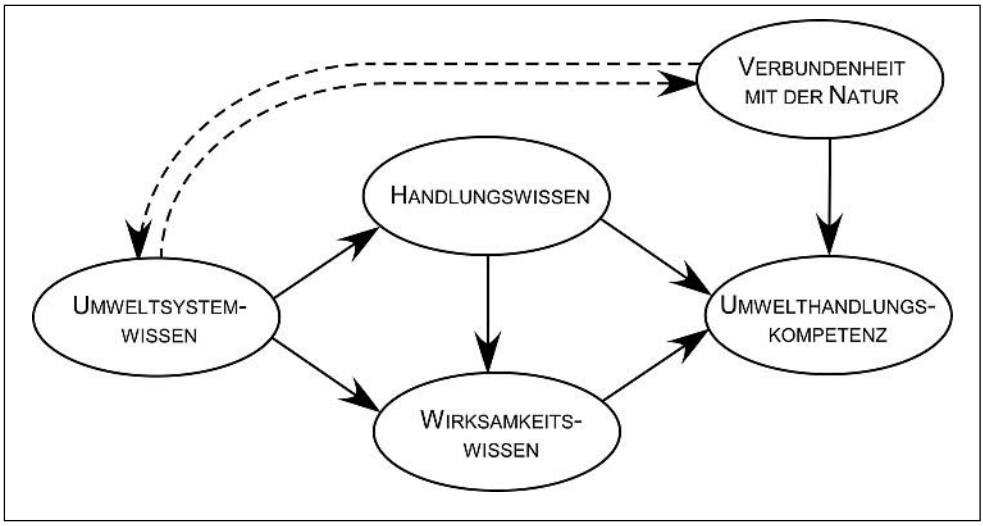


Abb.1: Wirkgefüge der Umweltkompetenz
Anmerkung: Durchgezogene Pfeile beziehen sich auf empirisch (korrelativ) bestätigte Effekte, durchbrochene Pfeile beziehen sich auf theoretisch antizipierte Effekte.

Kauf eines verbrauchsreduzierten Fahrzeuges und die freiwillige Mobilitätseinschränkung unterschiedliche Benzinsparpotentiale.

Auf der Basis eines mehrdimensional erweiterten Modells der Rasch-Familie konnten Frick und Kollegen die Subdimensionen Umweltsystem-, Handlungs- und Wirksamkeitswissen faktisch unterscheiden und in ihrer Struktur beschreiben (vgl. Frick/Kaiser/Wilson 2004). Dabei wirkt Systemwissen nicht direkt auf die Umwelthandlungskompetenz ein (dargestellt anhand durchgezogener Linien in Abbildung 1), liefert jedoch offenbar Gründe für die Suche nach angemessenen Handlungsweisen sowie nach Informationen über die Auswirkungen dieser Handlungen. Handlungswissen wiederum ist die Grundlage für die Aneignung von Wirksamkeitswissen und beeinflusst gleichzeitig die Umwelthandlungskompetenz. Wirksamkeitswissen wirkt seinerseits auf die Handlungskompetenz, beeinflusst die anderen Wissensformen jedoch nicht unmittelbar. Wie sich diese Struktur entwickelt, ist bislang noch ungeklärt. Querschnittstudien lassen eine zunehmende Integration der drei Umweltwissensarten und damit den Zerfall der Wissensstruktur mit zunehmendem Wissen vermuten (vgl. Frick/Kaiser/Wilson 2004; Kaiser/Frick 2002).

1.3 Naturverbundenheit

Über Wissen als notwendige Bedingung hinaus bedarf es auch einer motivationalen Komponente, um Umwelthandlungskompetenz aufzubauen. In den letzten Jahren zeichnet sich ab, dass eine essentielle Komponente „Verbundenheit mit der Natur“ ist. Zwar wird diese Eigenschaft unterschiedlich konzeptualisiert, einmal kognitiv im Rahmen

des Selbstkonzepts (vgl. „Umweltidentität“, Clayton 2003; „Natur als Teil des Selbstkonzepts“, Schultz 2001, 2002a), ein andermal emotional als „Verbundenheit mit der Natur“ (vgl. Mayer/Frantz 2004). Jedoch zeugt die umfangreiche gemeinsame Varianz davon, dass es sich bei diesen Konstrukten vermutlich um ein und dasselbe psychologische Phänomen handelt, das sich auch als Einstellung gegenüber der Natur konzeptualisieren lässt (nicht zu verwechseln mit Umwelteinstellung, bei welcher das Einstellungsobjekt nicht die Natur selbst, sondern der Umweltschutz ist; vgl. auch Brügger/Kaiser/Roczen 2009).

Auf welcher Grundlage sich Verbundenheit mit der Natur entwickelt, ist bisher noch kaum systematisch untersucht. Eine Reihe neuerer Untersuchungen legt die Vermutung nahe, dass eine positive Natureinstellung und damit Naturverbundenheit über Konditionierungsprozesse, d.h. durch die Assoziation positiver Erlebnisse mit der Natur, entsteht (vgl. Kaiser/Roczen/Bogner 2008; Schultz 2001) und darüber auch gefördert werden kann (vgl. z.B. Hartig/Kaiser/Strumse 2007).

Ungeklärt ist in diesem Zusammenhang auch, ob und auf welchen Wegen sich eine positive Einstellung der Natur gegenüber auf andere ökologiespezifische Fähigkeiten und Eigenschaften auszuwirken vermag. Wir gehen einerseits davon aus, dass Naturverbundenheit, zusätzlich zu dem bereits empirisch nachgewiesenen Zusammenhang mit Umwelthandlungskompetenz (vgl. z.B. Brügger/Kaiser/Roczen 2009), auch zur Aneignung von Wissen über die Natur und deren Funktionieren (zur Aneignung von Umweltsystemwissen also) motiviert. Es wäre aber andererseits auch denkbar, dass mit zunehmendem Umweltsystemwissen mehr Zeit in der Natur verbracht wird, was dann seinerseits eine noch intensivere Naturverbundenheit nach sich ziehen könnte. (Diese beiden theoretisch plausiblen, empirisch jedoch bislang nicht bestätigten Einflüsse sind in Abbildung 1 als durchbrochene Linien dargestellt).

2. Fragestellungen und Forschungsdesign

Einem Kompetenzansatz in der Umweltbildung folgend untersuchen wir, auf welche Weise diejenigen Fähigkeiten und Merkmale, die notwendige Voraussetzungen und Motivatoren von Umwelthandeln darstellen, sich sowohl gegenseitig als auch die Umwelthandlungskompetenz beeinflussen. Darüber hinaus ist von Interesse, wie sich eine Kompetenzstruktur im Jugendalter spontan entwickelt und wie Umweltkompetenz systematisch gefördert werden kann. Wir stellen zunächst unsere Forschungsfragen dar, anschließend gehen wir auf das daraus abgeleitete Forschungsdesign zu ihrer Beantwortung ein.

2.1 Fragestellungen

Unsere Forschungsfragen sind die Folgenden: (1) Lässt sich das Wirkgefüge der Umweltkompetenz bestehend aus Umweltwissen, Naturverbundenheit und Umwelthand-

lungskompetenz, wie theoretisch postuliert, modellieren? (2) Wie sieht der Entwicklungsverlauf der Umweltkompetenzstruktur aus? (3) Wie lässt sich Umweltkompetenz gezielt fördern?

2.2 Forschungsdesign

Zur Beantwortung dieser Forschungsfragen wurde eine erste Studie zur Entwicklung eines Messinstrumentes zur Erfassung von Naturverbundenheit durchgeführt, ein Umweltbildungspanel zusammengestellt und in diesem Rahmen bereits die Daten zu einer zweiten Studie erhoben. Darüber hinaus haben wir erste Lernmodule entwickelt. Wir beschreiben im Folgenden, wie wir dabei im Einzelnen methodisch vorgegangen sind.

Modellierung von Umweltkompetenz. Zur Modellierung der Umweltkompetenzstruktur haben wir eine große Datenerhebung an bayrischen Schulen vorgenommen. Folgende Messinstrumente sind dabei zum Einsatz gekommen: Das Instrument zur Messung von Naturverbundenheit (vgl. Brügger/Kaiser/Roczen 2009), die Skalen zur Erfassung von Umweltsystem-, Handlungs- und Wirksamkeitswissen (vgl. Frick/Kaiser/Wilson 2004) sowie das Maß zur Erhebung von Umwelthandlungskompetenz bei Jugendlichen (vgl. Kaiser/Oerke/Bogner 2007). Wir haben 82 Klassen in 3 Gymnasien und 4 Realschulen (jeweils der 6. bis 8. Jahrgangsstufen) vor Ort befragt ($N = 1922$). Die einzelnen Skalen sollen mit Hilfe des Raschmodells kalibriert werden. Anschließend wird die Kompetenzstruktur anhand von Mehrebenen-Strukturgleichungsmodellen analysiert.

Entwicklung von Umweltkompetenz. Aus der erhobenen Stichprobe soll ein Umweltbildungspanel entstehen. Das heißt, wir beabsichtigen, im Abstand von etwa einem Jahr in denselben Schulen erneut Daten von sechsten bis achten Klassen zu erheben. Dabei soll ein Kern von ursprünglichen Sechstklässlern über alle Schulstufen hinweg im Längsschnitt verfolgt werden. Zusätzlich zu der Analyse der spontanen Kompetenzenzentwicklung möchten wir auch den durch spezielle Interventionen geförderten Verlauf der Kompetenzstruktur nachzeichnen.

Förderung von Umweltkompetenz. Wir werden verschiedene unterrichtsbegleitende und ergänzende Interventionsmaßnahmen zur Förderung von Naturverbundenheit und Umweltwissen einsetzen, um auf diese Weise die von uns postulierte Kompetenzstruktur zu validieren. Mit anderen Worten: Wir werden überprüfen, ob sich über eine Verbesserung von Naturverbundenheit und Umweltwissen die Umwelthandlungskompetenz fördern lässt. Am Lehrstuhl für Didaktik der Biologie der Universität Bayreuth sind bereits einige Lernmodule mit Schwerpunkt auf Wissensvermittlung entwickelt worden. Es handelt sich dabei um Lerneinheiten für außerschulisch durchzuführende Projekttage für 6. Jahrgangsstufen der Realschule und des Gymnasiums. Theoretisch bauen diese Wissensinterventionen nicht nur auf allgemeinen Erkenntnissen der Biologiedidaktik auf, sondern basieren auch auf unseren theoretisch hergeleiteten Vermutungen zur strukturellen Entwicklung von Umweltwissen. Die entsprechenden Module dienen der Vermittlung aller drei Wissensarten, wobei besonderen Wert auf deren Integration gelegt wird. So gehen die Lernmodule immer von Systemwissen aus (d.h. Wissen über Zusammenhänge in Ökosystemen oder Naturzerstörung), welches dann in einem weite-

ren Schritt mit den Handlungsoptionen der/des Einzelnen (Handlungswissen) und den ökologischen Auswirkungen dieser Handlungen auf das Ökosystem (Wirksamkeitswissen) verknüpft wird.

3. Bisher durchgeführte Arbeiten – Entwicklung eines Instruments zur Erfassung von Naturverbundenheit

Während die Instrumente zur Erfassung der drei Wissensarten und der Handlungskompetenz bereits vorlagen, waren die bestehenden Instrumente zur Erfassung von Naturverbundenheit für den Einsatz bei Jugendlichen konzeptionell und messtechnisch nicht befriedigend. Aus diesem Grund war die erste empirische Arbeit auf dem Weg zur Modellierung von Umweltkompetenz die Entwicklung eines reliablen und validen Messinstruments für Naturverbundenheit, das auch für den Einsatz bei Jugendlichen geeignet ist (vgl. Brügger/Kaiser/Roczen 2009). Anhand einer Gelegenheitsstichprobe von $N = 1309$ Proband/innen wurde eine Skala entwickelt, bei welcher, im Gegensatz zu den meisten bestehende Instrumenten, das Ausmaß der eigenen Naturverbundenheit *indirekt* erfasst wird. Während bestehende Instrumente Naturverbundenheit bzw. Umweltidentität *direkt* über Selbsteinschätzungen von Aussagen wie „Dass ich ein Teil des Ökosystems bin, macht einen zentralen Teil dessen aus, wer ich bin" (vgl. Clayton 2003) oder „Ich fühle mich als Teil des Netzwerks des Lebens" (vgl. Mayer/Frantz 2004) erheben, wird bei unserem neuen Instrument die Naturverbundenheit einer Person indirekt aus Verhaltensberichten und einfachen Bewertungsaussagen erschlossen, von denen angenommen wird, dass sie indikativ für eine mehr oder weniger stark positiv ausgeprägte Einstellung der Natur gegenüber sind. Itembeispiele sind „Ich beobachte oder höre bewusst Vögeln zu" oder „Ich verspüre ein Bedürfnis, draußen in der Natur zu sein". Da die Einschätzung solcher Aussagen intellektuell weniger anspruchsvoll ist als die direkte Beurteilung des Ausmaßes der eigenen Naturverbundenheit bzw. der Umweltidentität, sollte ein solches Messinstrument auch angemessener für den Einsatz bei Kindern und Jugendlichen sein. Um die konvergente Validität zu überprüfen – d.h., um zu prüfen, ob trotz konzeptioneller Veränderung und veränderter Erhebungsart noch stets „Verbundenheit mit der Natur" gemessen wird – haben wir drei weitere Maße zur Erfassung von Naturverbundenheit und Umweltidentität mit erhoben (vgl. Environmental Identity: Clayton 2003; Connectedness to Nature Scale: Mayer/Frantz 2004; Inclusion of Nature in Self: Schultz 2001). Zur Bestimmung der diskriminanten Validität fügten wir zudem ein Instrument zur Erfassung von Umwelteinstellung hinzu (vgl. New Ecological Paradigm: Dunlap u.a. 2000).

Die Reliabilität des neu entwickelten Instruments war mit $rel = .89$ sehr gut und auch die Befunde zur konvergenten und diskriminanten Validität erwiesen sich als äußerst zufriedenstellend. Unsere neue Skala (Naturverbundenheit, NV) korrelierte hoch mit allen anderen Maßen für Verbundenheit mit der Natur ($r > .65$),[2] jedoch nur mäßig ($r = .39$) mit Umwelteinstellung (NEP). Dabei zeigte sich, dass sich unser Maß für Verbunden-

2 Bei den berichteten Werten handelt es sich um messfehlerkorrigierte Korrelationen.

heit mit der Natur deutlicher als die meisten der bereits bestehenden Verfahren von Umwelteinstellung unterscheiden lässt (siehe Tabelle 1).

Auch die Analysen zur prädiktiven Validität des neuen Instruments fielen positiv aus. Als alleiniger Prädiktor vermochte unser neues Naturverbundenheitsmaß 23% der Varianz der Umwelthandlungskompetenz aufzuklären. Und selbst in einer multiplen Regressionsanalyse klärte unser Messinstrument über Alter, Geschlecht, soziale Erwünschtheit, Umwelteinstellung und alle übrigen Naturverbundenheitsmaße hinaus einen eigenständigen kleinen, aber signifikanten Varianzanteil (1,7%) an der Handlungskompetenz auf.

	N	NV	EID	CNS	INS	NEP
Naturverbundenheit (NV)	1239	*.89*	**.79**	**.71**	**.65**	.39
Environmental Identity (EID)	1064	**.72***	*.93*	**.78**	**.68**	**.57**
Connectedness to Nature Scale (CNS)	1121	**.60***	**.67***	*.80*	**.66**	**.62**
Inclusion of nature in self (INS)	1182	**.56***	**.60***	**.54***	*.84*	.31
New Ecological Paradigm (NEP)	1128	.34*	**.51***	**.51***	.26*	*.84*

Tab. 1: Korrelationsmatrix zur Überprüfung der Validität des neu entwickelten Messinstrumentes zur Erfassung von Naturverbundenheit
Anmerkung: Bei den dargestellten Werten handelt es sich um unkorrigierte (unterhalb der Diagonalen) und um messfehlerkorrigierte (oberhalb der Diagonalen) Pearson-Korrelationen. Die Werte in der Diagonalen repräsentieren Reliabilitätsschätzungen. **Fettgedruckte** Werte stellen starke Effekte dar ($r > .50$). * steht für $p < .001$; allgemein akzeptierte Signifikanztests stehen lediglich für unkorrigierte Korrelationskoeffizienten zur Verfügung.

4. Diskussion

Auf dem Weg zu einem Kompetenzmodell für die Umweltbildung haben wir ein Messinstrument zur Erfassung von Verbundenheit mit der Natur entwickelt sowie ein Umweltbildungspanel zusammengestellt. Mit den an unserem Panel erhobenen Daten wird die Kompetenzstruktur, deren Entwicklungsverlauf und ihre Förderbarkeit untersucht. Im Folgenden möchten wir diskutieren, welche Erkenntnisgewinne sich in theoretischer, methodischer sowie auch praktischer Hinsicht bisher bereits ergaben bzw. zu erwarten sein werden.

Das von uns neu entwickelte Messinstrument folgt einem zu den bestehenden Instrumenten alternativen Ansatz und leitet die individuelle Naturverbundenheit einer Person mehrheitlich aus einfach einzuschätzenden Verhaltens- und Bewertungsaussagen ab. Damit steht uns nun ein Messinstrument zur Verfügung, das zuverlässig und valide ist und sich auch für den Einsatz bei Kindern und Jugendlichen eignet.

Unser Umweltkompetenzmodell liefert mit seinen ökologiespezifischen Fähigkeiten konkretere Ansatzpunkte für gezielte Fördermaßnahmen als dies herkömmliche Mo-

delle der Umweltbildung vermögen, die auf allgemeinen Fähigkeiten aufbauen (vgl. z.B. Kyburz-Graber 2004). Zudem ist davon auszugehen, dass eine erfolgreiche Förderung von ökologiespezifischen Fähigkeiten verhaltenswirksamer sein wird als die Förderung allgemeiner Fähigkeiten. Während die Modellierung von systematisch geförderten und spontanen Entwicklungsverläufen der Kompetenzstruktur einen Beitrag zum besseren theoretischen Verständnis von Umweltkompetenz leisten wird (vgl. Kaiser/Roczen/Bogner 2008), sollen die neu entwickelten (und hoffentlich erfolgreich wirksamen) Lernmodule den Schulen neue Wege bei der Umweltbildung weisen.

Insgesamt erhoffen wir uns, durch unsere Forschung zu einem fundierteren Verständnis von Umweltkompetenz, ihrer Struktur, Entwicklung sowie ihrer Förderbarkeit beizutragen. Zur Erfassung der ökologiespezifischen Fähigkeiten konnten wir bereits einen Beitrag leisten. Den praktischen Ertrag sehen wir langfristig in einer evidenzbasierten und damit unideologischen Förderung des Umwelthandelns von Kindern und Jugendlichen.

Literatur

Brügger, A./Kaiser, F.G./Roczen, N. (eingereicht): One to bind them all: Connectedness to nature, inclusion of nature, environmental identity, implicit association with nature.

Clayton, S. (2003): Environmental identity: A conceptual and operational definition. In: Clayton, S./Opotow, S. (Hrsg.): Identity and the natural environment. Cambridge, MA: MIT press, S. 45–65.

De Haan, G. (2006): The „BLK 21 program" in Germany: A „Gestaltungskompetenz"-based model for education for sustainable development. In: Environmental Education Research 12, S. 19–32.

Dunlap, R.E./Van Liere, K.D./Mertig, A.G./Jones, R.E. (2000): Measuring endorsement of the new ecological paradigm: A revised NEP scale. In: Journal of Social Issues 56, S. 425–442.

Frick, J./Kaiser, F.G./Wilson, M. (2004): Environmental knowledge and conservation behavior: Exploring prevalence and structure in a representative sample. In: Personality and Individual Differences 37, S. 1597–1613.

Gardner, G.T./Stern, P.C. (^{2}2002): Environmental problems and human behavior. Boston, MA: Pearson.

Hartig, T./Kaiser, F.G./Strumse, E. (2007): Psychological restoration in nature as a source of motivation for ecological behaviour. In: Environmental Conservation 34, S. 291–299.

Hines, J.M./Hungerford, H.R./Tomera, A.N. (1986/87): Analysis and synthesis of research on responsible environmental behavior: A meta-analysis. In: Journal of Environmental Education 18, H. 2, S. 1–8.

Kaiser, F.G./Frick, J. (2002): Entwicklung eines Messinstrumentes zur Erfassung von Umweltwissen auf der Basis des MRCML-Modells. In: Diagnostica 48, S. 181–189.

Kaiser, F.G./Oerke, B./Bogner, F.X. (2007): Behavior-based environmental attitude: Development of an instrument for adolescents. In: Journal of Environmental Psychology 27, S. 242–251.

Kaiser, F.G./Roczen, N./Bogner, F.X. (2008): Competence formation in environmental education: Advancing ecology-specific rather than general abilities. In: Umweltpsychologie 12, H. 2, S. 56–70.

Kaiser, F.G./Wilson, M. (2004): Goal-directed conservation behavior: The specific composition of a general performance. In: Personality and Individual Differences 36, S. 1531–1544.

Kyburz-Graber, R. (2004): Welches Wissen, welche Bildung? Aktuelle Entwicklungen in der Umweltbildung. In: Beiträge zur Lehrerbildung 20, S. 83–94.
Mayer, F.S./Frantz, C.M. (2004): The connectedness with nature scale: A measure of individuals' feeling in community with nature. In: Journal of Environmental Psychology 24, S. 503–515.
McClelland, D.C. (1973): Testing for competence rather than intelligence. In: American Psychologist 28, S. 1–14.
OECD, Directorate for Education, Employment, Labour and Social Affairs, Education Committee (2003): Definition and selection of competencies (DeSeCo). Theoretical and conceptual foundations (Summary of the final report "Key Competencies for a Successful Life and a Well-Functioning Society"). Neuchâtel, Switzerland: DeSeCo Secretariat.
Schultz, P.W. (2001): The structure of environmental concern: Concern for self, other people, and the biosphere. In: Journal of Environmental Psychology 21, S. 327–339.
Schultz, P.W. (2002a): Inclusion with nature: The psychology of human-nature relations. In: Schmuck, P./Schultz, W.P. (Hrsg.): Psychology of sustainable development. Boston: Kluwer, S. 61–78.
Schultz, P.W. (2002b): Knowledge, information, and household recycling: Examining the knowledge-deficit model of behavior change. In: Dietz, Th./Stern, P.C. (Hrsg.): New tools for environmental protection: Education, information, and voluntary measures. Washington, DC: National Academy Press, S. 67–82.
Weinert, F.E. (2001): Concept of competence: A conceptual clarification. In: Rychen, D.S./Salganik, L.H. (Hrsg.): Defining and selecting key competencies. Seattle, WA: Hogrefe & Huber, S. 45–65.

Anschrift der Autorin/der Autoren

Dipl. Psych. Nina Roczen, Otto-von-Guericke-Universität Magdeburg,
Sozial- und Persönlichkeitspsychologie, Postfach 4120, D-39016 Magdeburg
E-Mail: nina.roczen@gast.uni-magdeburg.de

Prof. Dr. Florian G. Kaiser, Otto-von-Guericke-Universität Magdeburg,
Sozial- und Persönlichkeitspsychologie, Postfach 4120, D-39016 Magdeburg
E-Mail: florian.kaiser@ovgu.de

Prof. Dr. Franz X. Bogner, Universität Bayreuth, Lehrstuhl Didaktik der Biologie (Z-MNU),
Universitätsstraße 30 (NW-1), D-95447 Bayreuth
E-Mail: franz.bogner@uni-bayreuth.de

Ilka Parchmann

Kompetenzmodellierung in den Naturwissenschaften

Vielfalt ist wertvoll, aber nicht ohne ein gemeinsames Fundament

Review

1. Einleitung

Das Schwerpunktprogramm (SPP) „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen" impliziert hohe Erwartungen an die Ergebnisse und die Nutzbarkeit der verorteten Forschungsarbeiten, einerseits für die Weiterentwicklung der Bildungsforschung, andererseits als forschungsbasiertes Fundament für Entwicklungsprozesse im Bildungssystem. Die damit explizit verbundene Forderung der beiden Koordinatoren nach einem „integrativen Forschungsprogramm" (Klieme/Leutner 2006, S. 878) bildet den Leitgedanken der nachfolgenden Ausführungen, die zunächst kurz den Beitrag der vier beteiligten Projekte beleuchten und anschließend eine Gesamtbetrachtung aktueller Arbeiten zur Kompetenzmodellierung in den Naturwissenschaften vornehmen.

2. Kompetenzmodellierung in den Naturwissenschaften – Welchen spezifischen Beitrag leisten die Projekte des SPP?

Im SPP werden Kompetenzen als „kontextspezifische kognitive Leistungsdispositionen, die sich funktional auf Situationen und Anforderungen in bestimmten Domänen beziehen" charakterisiert (Klieme/Leutner 2006, S. 879). Die von den Autoren hier zitierte Definition nach White (1959) betont ebenfalls die Wechselwirkung zwischen Individuum und Kontext, hier wird Kompetenz als: „effective interaction (of the individual) with the environment" charakterisiert (ebd., S. 879). In welcher Weise erfassen die betrachteten Projekte zu den Naturwissenschaften diese Konstrukte?

Alle vier Projekte untersuchen auf hohem Niveau mit unterschiedlichen Methodenrepertoires kognitive Leistungsausprägungen in verschiedenen Situationen. Besonders wertvoll für die Weiterentwicklung der fachdidaktischen Forschung erscheint die Vielfalt der eingesetzten Verfahren, die neben klassischen Testinstrumenten auch qualitative Interpretationen beinhalten. Ebenfalls positiv hervorzuheben sind die unterschiedlichen Blickwinkel, mit denen die Projekte das Kompetenzkonstrukt beleuchten, auch wenn sich gerade dadurch die Forderung nach einer zukünftigen Zusammenführung (s.u.) ergibt. So charakterisiert das Projekt *Science-P* die unterschiedlichen Vorstellungen von Kindern im Grundschulalter und verbindet damit die Forschungsrichtung der Schülervorstellungsuntersuchung mit der der Kompetenzmodellierung. Das Physik-

projekt greift das Modell „ESNaS“, das zur Überprüfung der Nationalen Bildungsstandards in den Naturwissenschaften eingesetzt wird, auf, erweitert dies jedoch um die bisher national wenig, international aber explizit betrachtete Dimension der inhaltlichen Sachstruktur (vgl. Atlas der American Association for the Advancement of Science). Eine Verbindung von kognitiven und affektiven bzw. metakognitiven Komponenten gelingt in den Projekten zur Umwelt- und Gesundheitskompetenz, sodass auch diese eine interessante und wertvolle Verknüpfung bisheriger Forschungstraditionen darstellen. Das eingangs zitierte Wechselspiel zwischen Individuen und Kontexten bzw. Lernumgebungen wird bisher wenig explizit betont, lediglich das Projekt *Gesundheitskompetenz* stellt in Aussicht, „Situationen“ zu klassifizieren. Dies erscheint vor dem Hintergrund, dass nicht nur im Beitrag von Csapó wiederkehrend auf die Problematik des Transfers und damit auch der Übertragbarkeit von Kompetenzaussagen hingewiesen wird, dringend erforderlich.

3. Versuch einer Einordnung in ein Gesamtvorhaben

Die Einordnung der vier exemplarischen Projekte in einen Forschungszusammenhang verlangt zunächst die Formulierung einer gemeinsamen Basis hinsichtlich des Grundbegriffsverständnisses (Wie grenzen sich etwa „Kompetenz“, „Wissen“, „Grundbildung“ und „Expertise“ voneinander ab? Vgl. Csapó in diesem Heft), hinsichtlich der Komponenten des Kompetenzbegriffs (Wann und wie erfolgt die auch von Klieme und Leutner geforderte Zusammenführung der Erfassung kognitiver und motivationaler Aspekte? Vgl. Oelkers 2007) und hinsichtlich der Dimensionierung und Graduierung von Kompetenzmodellen (vgl. Tabelle 1 und Parchmann 2008).

Csapó postuliert in seinem Aufsatz drei Dimensionen von Lernzielen (die internale oder psychologische, die disziplinäre und die sozio-kulturelle Dimension), die nach seinen Ausführungen nur zusammenwirkend zu einer Kompetenzentwicklung beitragen können. Lassen sich diese in Einklang bringen mit den Dimensionen der aktuell postulierten Kompetenzmodelle für die Domänen der Naturwissenschaften? Ein Blick in die internationale Forschungsliteratur zeigt hier vorrangig eines: Vielfalt! (vgl. Waddington/Nentwig/Schanze 2007; für das deutsche Bildungssystem vgl. z.B. Hammann 2004; Eggert/Bögeholz 2006; Neumann u.a. 2007; Bernholt u.a. 2009). Die Hintergründe dieser postulierten und z.T. empirisch untersuchten, bisher jedoch nur in Ansätzen bestätigten Modelle sind dabei sehr vielfältig: Während in den großen Studien TIMSS und PISA die Stufen post hoc beschrieben wurden, erwarten andere Modelle Niveaus auf der Basis psychologischer Theorien und Modelle (Bsp. ESNaS, MHC-C). Dimensionen begründen sich z.T. auf der Grundlage von Bildungskonzepten (bspw. *scientific literacy* in PISA), aber auch auf typischen Komponenten naturwissenschaftlicher Denk- und Arbeitsweisen (Bsp. TIMSS oder *Science-P*) oder auf empirischen Befunden quantitativer oder qualitativer Untersuchungen sowie Expertenbefragungen (siehe bspw. das Göttinger Modell zur Bewertungskompetenz oder das Projekt *Gesundheitskompetenz*).

Modell/Projekt	**Dimensionen/Komponenten und Stufen/ Niveaus/Ausprägungen**

TIMSS III (Klieme et al. 2000)	***Dimensionen:***
	● Sachgebiete
	● Fähigkeiten:
	  ● Verstehen von Einzelinformationen
	  ● Verstehen komplexer Informationen
	  ● Konzeptionalisieren und Anwenden
	  ● Experimentieren und Beherrschung von Verfahren
	Stufen (post hoc konkretisiert):
	● Naturwissenschaftliches Alltagswissen
	● Erklärung einfacher alltagsnaher Phänomene
	● Anwendung elementarer naturwissenschaftlicher Modellvorstellungen
	● Verfügung über grundlegende naturwissenschaftliche Fachkenntnisse

PISA 2006 (Prenzel et al. 2007)	***Dimensionen:*** ● Naturwissenschaftliche Fragestellungen erkennen ● Naturwissenschaftliche Phänomene erklären ● Naturwissenschaftliche Evidenz nutzen ***Stufen*** (post hoc konkretisiert): ● 6 Stufen
Projekt „Science-P/Grundschule“ (Hardy et al., in diesem Heft)	***Bezugnahme auf Komponenten naturwissenschaftlicher Kompetenzen*** nach Duit, Häußler & Prenzel 2001: ● Inhaltlich/konzeptuelle Komponente ● Komponente naturwissenschaftlicher Methoden und Denkweisen ● Komponente des Wissenschaftsverständnisses („Nature of Science“) ● Komponente des gesellschaftlichen Bezugs ***Ausgewählte Dimensionen:*** ● Naturwissenschaftliches Wissen ● Wissen über die Naturwissenschaften ***Niveaus*** (postuliert auf Basis empirischer Literatur zu Schülervorstellungen und Lerntheorien): ● Naive Vorstellungen ● Zwischenvorstellungen ● Wissenschaftliche Vorstellungen ● Integration von Vorstellungen

Modell/Projekt	Dimensionen/Komponenten und Stufen/ Niveaus/Ausprägungen
Länderübergreifende Bildungsstandards (NBS) der Naturwissenschaften (KMK 2005)	***Dimensionen:*** ● Fachwissen (strukturiert nach Basiskonzepten) ● Erkenntnisgewinnung ● Kommunikation ● Bewerten ***Ausprägungen*** (vorgegeben) ● Orientiert an den drei Anforderungsbereichen der einheitlichen Prüfungsanforderungen für die Abiturprüfung und formuliert für die vier Kompetenzbereiche
ESNaS und Projekt „Physikalische Kompetenz“ (Neumann et al. 2007; Kauertz 2008; Viering et al. in diesem Heft)	***Dimensionen des ESNaS-Modells:*** ● Leitidee (vgl. Basiskonzepte der NBS) ● Kognitive Aktivität ● Komplexität *Erweiterung/Differenzierung im SPP:* ● Konzeptentwicklung ● Aufgabenkomplexität als Differenz aus Lösungskomplexität und Textkomplexität ***Niveaus*** (postuliert und empirisch untersucht für die Dimension „Komplexität“): ● Ein Fakt ● Zwei Fakten ● Ein Zusammenhang ● Zwei Zusammenhänge ● Übergeordnetes Konzept/mehrere Fakten und Zusammenhänge
Bremen-Oldenburger Kompetenzmodell (Schecker & Parchmann 2006; Einhaus 2007) und Model of Hierarchichal Complexity (MHC-C; Bernholt et al. 2008, 2009a und b)	***Dimensionen:*** ● Inhaltsbereich/Basiskonzept (vgl. NBS) ● Prozesse (vgl. Kompetenzbereiche der NBS) ● Kontexte ● Hierarchische Komplexität ***Stufen*** (postuliert und empirisch untersucht für die Dimension „hierarchische Komplexität“): ● Unreflektiertes Erfahrungswissen ● Fakten ● Prozessbeschreibungen ● Einfache, lineare Kausalität ● Multivariate Interdependenz
Projekt „Umweltkompetenz“ (Roczen et al. in diesem Heft)	***Dimensionen:*** ● Umwelthandlungskompetenz ● Umweltwissen (Umweltsystemwissen, Handlungswissen, Wirksamkeitswissen) ● Naturverbundenheit ***Stufen/Ausprägungen:*** nicht formuliert

Modell/Projekt	**Dimensionen/Komponenten und Stufen/ Niveaus/Ausprägungen**
Projekt „Gesundheitskompetenz“ (Soellner et al. in diesem Heft)	***Komponenten*** (verkürzt): ● Fähigkeit zu Selbstregulation ● Fähigkeit zur Wahrnehmung eigener Bedürfnisse ● Fähigkeit zur Verantwortungsübernahme ● Grundfertigkeiten (Bsp. Texte lesen) ● Fähigkeit, Informationen angemessen zu interpretieren und zu nutzen ● Fähigkeit, Informationen zu beschaffen ● Systemwissen ● Fähigkeit zur Kommunikation und zur Kooperation ● Persönlichkeitseigenschaften (Hinweis auf Clusterung im Originalbeitrag) ***Stufen bzw. aufeinander aufbauende Formen eines zuvor publizierten Stufenmodells*** nach Nutbeam 2000: ● Funktionale Gesundheitskompetenz ● Kommunikative Gesundheitskompetenz ● Kritische Gesundheitskompetenz

Tab. 1: Strukturen von Kompetenzmodellen in Domänen der Naturwissenschaften
Hinweis: Die uneinheitlichen Begrifflichkeiten (Bsp. Dimensionen – Komponenten) ergeben sich aus den von den Autor/innen verwendeten Darstellungen, die zur Vermeidung verfälschender Aussagen übernommen wurden. Die angeführten empirischen Untersuchungen haben die Konstrukte nicht in allen Facetten bestätigt.

Gemeinsam scheint allen Modellen das Ausweisen einer Dimension oder Komponente „Wissen“ bzw. „Inhalt“ sowie eine den Stufungen zugrunde liegende Annahme von wachsender Vernetzung und Komplexität der inhaltlichen Zusammenhänge und Perspektiven zu sein. Ersteres steht in Übereinstimmung mit zahlreichen empirischen Untersuchungen, die einen starken Einfluss des Vorwissens auf das Lernen bzw. die Entwicklung von Expertise nachweisen. Die Annahme von Komplexität als Schwierigkeit erzeugendem Faktor und von einer mit wachsender Kompetenzausprägung zunehmend strukturierten Vernetzung geht konform mit lernpsychologischen Modellen sowie mit der Sachstruktur von Inhalten (vgl. Case 1992; Commons u.a. 1998; Dawson-Tunik 2006; Fischer/Bidell 2006; Steiner 2006 u.v.a.m). Welche weiteren gemeinsamen Dimensionen und Grundannahmen über Niveaus und Ausprägungen lassen sich domänenübergreifend ausweisen? Hier sollten die Beteiligten zum Abschluss des Programms sicher Auskunft geben können.

Es ist selbstverständlich, dass für das explorative Formulieren und Untersuchen von Kompetenzen zunächst „high impact factors“ (vgl. Hardy u.a. in diesem Heft und Bern-

holt/Parchmann/Commons 2009) fokussiert werden. Diese sollten aber eingebettet sein in ein Gesamtkonstrukt, das auf einer Metaebene über Fachdomänen hinweg *gemeinsame Strukturmerkmale* (bspw. die Dimension „Wissen" oder Prozessstrukturen von Bewerten) und dazu *domänenspezifische Dimensionen* (bspw. die Basiskonzeptstrukturen oder spezifisches Umweltwissen) ausweist. Auch die *Interaktionen verschiedener Parameter* wie Wissen und Motivation zur Wissensaneignung (vgl. Projekt zur Umweltkompetenz) oder die Auseinandersetzung eines Individuums mit verschiedenen Lernumgebungen (Kontextabhängigkeit) müssen langfristig in ein Gesamtmodell eingebracht werden,[1] um domänenübergreifend Fortschritte zu erlangen und um Fehlinterpretationen zu vermeiden.

4. Fazit und Ausblick

Die Beiträge zu den Domänen der Naturwissenschaften zeigen alle ein fundiertes und in sich stimmiges Vorgehen zur Modellierung von Kompetenzen; eine kohärente Entwicklung in Richtung eines gemeinsamen Grundmodells von Kompetenz auf der Metaebene *zuzüglich* ausgewiesener Domänen- und Kontextspezifitäten ist bisher jedoch nicht zu erkennen. Dies mag aufgrund der Anfangsphase des SPP auch nicht zu erwarten sein, muss jedoch an dieser Stelle bereits mitgedacht werden, damit ein späteres Zusammenführen und vergleichendes Bewerten von Ergebnissen möglich wird. Empirische, psychometrische Forschung im Bereich der Kompetenzmodellierung kann zweifelsohne einen entscheidenden Beitrag für die Weiterentwicklung der allgemeinen und fachdidaktischen Bildungsforschung und eine fundierte Basis für Prozesse im Bildungssystem darstellen, sofern es gelingt, die gewonnenen Ergebnisse über die einzelnen Projektfragen hinaus zu einem kohärenten Gefüge zusammenzuführen. Fragen wie „Wie ist der Stand der Forschung auf dem Gebiet der Kompetenzmodellierung?" lassen sich kaum mit „Aus der Physik wissen wir etwas über Komplexität, aus der Umweltbildung etwas über Naturbewusstsein, aus der Grundschule etwas über das Wissenschaftsverständnis von Kindern." beantworten, sondern sollten längerfristig zu Aussagen führen, wie sie in den Naturwissenschaften üblich sind: „Heutige Modelle und experimentelle Befunde zeigen für die Struktur von Atomen [analog für ein Grundmodell „Kompetenz"] allgemein folgende Parameter: …; für die besonderen Eigenschaften des Goldes [analog für Kompetenzausprägungen in ausgewählten Domänen] lassen sich damit folgende Vorhersagen und Erklärungen treffen: …" Das SPP kann und sollte m. E. zu diesem Ziel einen wichtigen Beitrag leisten.

1 Diese Forderung stellt nicht die Domänen- und Kontextbezogenheit von Kompetenzentwicklungen in Frage, sondern verlangt im Gegenteil nach deren expliziter Ausweisung im Rahmen eines Grundmodells von Kompetenzstruktur.

Literatur

American Association for the Advancement of Science: Atlas of Scientific Literacy. http://www.project2061.org/publications/atlas/default.htm [21.09.2009].

Bernholt, S./Parchmann, I. (2008): Lösungsstrategien bei der Bearbeitung von Aufgaben. In: Höttecke, D. (Hrsg.): Kompetenzen, Kompetenzmodelle, Kompetenzentwicklung. Tagungsband der Gesellschaft für Didaktik der Chemie und Physik. Berlin: Lit, S. 215–217.

Bernholt, S./Parchmann, I./Commons, M.L. (2009): Kompetenzmodellierung zwischen Forschung und Unterrichtspraxis. In: Zeitschrift für Didaktik der Naturwissenschaften 15, S. 217–243.

Bernholt, S./Walpuski, M./Sumfleth, E./Parchmann, I. (2009): Kompetenzentwicklung im Chemieunterricht – mit welchen Modellen lassen sich Kompetenzen und Aufgaben differenzieren? In: Naturwissenschaften im Unterricht – Chemie, Themenheft „Differenzieren", S. 78–85.

Case, R. (1992): Neo-Pagetian theories of intellectual development. In: Beilin, H./Pufall, P.B. (Hrsg.): Piaget's theory: Prospects and possibilities. Hillsdale, NJ: Lawrence Erlbaum, S. 61–104.

Commons, M.L./Trudeau, E.J./Stein, S.A./Richards, F.A./Krause, S.R. (1998): Hierarchical Complexity of Tasks Shows the Existence of Developmental Stages. In: Developmental Review 18, H. 3, S. 237–278.

Csapó, B. (2010): Goals of Learning and the Organization of Knowledge. In: Klieme, E./Leutner, D./Kenk, M.: Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. In: 56. Beiheft der Zeitschrift für Pädagogik, Weinheim/Basel: Beltz, S. 12–27.

Dawson-Tunik, T.L. (2006): Stage-like patterns in the development of conceptions of energy. In: Liu, X./Boone, W. (Hrsg.): Applications of Rasch measurement in science education. Maple Grove, MN: JAM Press, S. 111–136.

Duit, R./Häußler, P./Prenzel, M. (2001): Schulleistungen im Bereich der naturwissenschaftlichen Bildung. In: Weinert, F.E. (Hrsg.): Leistungsmessungen in Schulen. Weinheim/Basel: Beltz, S. 169–185.

Eggert, S./Bögeholz, S. (2006): Göttinger Modell der Bewertungskompetenz – Schwerpunkt Prozessdimension „Bewerten, Entscheiden und Reflektieren" im Kontext Nachhaltiger Entwicklung. In: Zeitschrift für Didaktik der Naturwissenschaften 12, S. 199–217.

Einhaus, E. (2007): Schülerkompetenzen im Bereich Wärmelehre – Entwicklung eines Testinstruments zur Überprüfung und Weiterentwicklung eines normativen Modells fachbezogener Kompetenzen. Studien zum Physik- und Chemielernen, Band 63, Berlin: Logos.

Fischer, K.W./Bidell, T.R. (^{6}2006): Dynamic development of action, thought, and emotion. In: Damon, W./Lerner, R.M. (Hrsg.): Theoretical models of human development. Handbook of child psychology. Bd. 1. New York: Wiley, S. 313–399.

Hammann, M. (2004): Kompetenzentwicklungsmodelle: Merkmale und ihre Bedeutung – dargestellt anhand von Kompetenzen beim Experimentieren. In: Der mathematische und naturwissenschaftliche Unterricht 57, H. 4, S. 196–203.

Hardy, I./Kleickmann, T./Koerber, S./Mayer, D./Möller, K./Pollmeier, J./Schwippert, K./Sodian, B. (2010): Die Modellierung naturwissenschaftlicher Kompetenz im Grundschulalter. In: Klieme, E./Leutner, D./Kenk, M.: Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. 56. Beiheft der Zeitschrift für Pädagogik, Weinheim/Basel: Beltz, S. 114–124.

Kauertz, A. (2008): Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben. In: Studien zum Physik- und Chemielernen, Bd. 79. Berlin: Logos.

Klieme, E./Baumert, J./Köller, O./Bos, W. (2000): Mathematische und naturwissenschaftliche Grundbildung: Konzeptuelle Grundlagen und die Erfassung und Skalierung von Kompetenzen. In: Baumert, J./Bos, W./Lehmann, R. (Hrsg.): Dritte Internationale Mathematik- und Na-

turwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Bd. 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit. Opladen: Leske + Budrich, S. 85–133.

Klieme, E./Leutner, D. (2006): Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. In: Zeitschrift für Pädagogik 53, H. 6, S. 876–903.

KMK, Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.) (2005): Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss. München: Luchterhand.

Neumann, K./Kauertz, A./Lau, A./Notarp, H./Fischer, H.E. (2007): Die Modellierung physikalischer Kompetenz und ihrer Entwicklung. In: Zeitschrift für Didaktik der Naturwissenschaften 13, S. 103–123.

Nutbeam, D. (2000): Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st Century. In: Health Promotion International 15, S. 259–267.

Oelkers, J. (2007): Bildungsstandards am Gymnasium – Ein neues Problem? In: Labbude, P. (Hrsg.): Bildungsstandards am Gymnasium – Korsett oder Katalysator? Bern: hep, S. 27–36.

Parchmann, I. (2008): Bildungsstandards und Kompetenzmodelle – Katalysatoren für fachdidaktische Forschung, Lehrerbildung und Unterrichtsentwicklung? In: Höttecke, D. (Hrsg.): Kompetenzen, Kompetenzmodelle, Kompetenzentwicklung. Tagungsband der Gesellschaft für Didaktik der Chemie und Physik, Berlin: Lit, S. 5–13.

Prenzel, M./Artelt, C./Baumert, J. (Hrsg.) (2007): PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie. Münster: Waxmann.

Roczen, N./Kaiser, F./Bogner, F. (2010): Umweltkompetenz – Modellierung, Entwicklung und Förderung. In: Klieme, E./Leutner, D./Kenk, M.: Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. 56. Beiheft der Zeitschrift für Pädagogik, Weinheim/Basel: Beltz, S. 125–133.

Schecker, H./Parchmann, I. (2006): Modellierung naturwissenschaftlicher Kompetenz. In: Zeitschrift für Didaktik der Naturwissenschaften 12, S. 45–66.

Soellner, R./Huber, S./Lenartz, L./Rudinger, G. (2010): Facetten der Gesundheitskompetenz – eine Expertenbefragung. In: Klieme, E./Leutner, D./Kenk, M.: Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. 56. Beiheft der Zeitschrift für Pädagogik, Weinheim/Basel: Beltz, S. 104–113.

Steiner, G. (2006): Lernen und Wissenserwerb. In: Krapp, A./Weidenmann, B. (Hrsg.): Pädagogische Psychologie. Weinheim/Basel: Beltz, S. 137–202.

Viering, T./Fischer, H.E./Neumann, K. (2010): Die Entwicklung physikalischer Kompetenz in der Sekundarstufe I. In: Klieme, E./Leutner, D./Kenk, M.: Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. 56. Beiheft der Zeitschrift für Pädagogik, Weinheim/Basel: Beltz, S. 92–103.

Waddington, D./Nentwig, P./Schanze, S. (Hrsg.) (2007): Standards in Science Education. Münster: Waxmann.

Anschrift der Autorin

Ilka Parchmann, Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik an der Universität Kiel, Olshausenstraße 62, D-24098 Kiel
E-Mail: parchmann@ipn.uni-kiel.de

Wolfgang Schnotz/Holger Horz/Nele McElvany/Sascha Schroeder/
Mark Ullrich/Jürgen Baumert/Axinja Hachfeld/Tobias Richter

Das BITE-Projekt: Integrative Verarbeitung von Bildern und Texten in der Sekundarstufe I

Projekt BITE[1]

1. Ziel des Projekts

Nach dem Leseunterricht in den ersten Jahren in der Grundschule sollen Schüler/innen die erworbene Lesefähigkeit zum Erwerb neuen Wissens einsetzen. D.h.: Auf das Erlernen des Lesens folgt das Lesen, um zu lernen. In der Sekundarstufe verändern sich die Leseanforderungen insofern, als in vielen Fächern neben schriftlichen Texten auch verschiedene Arten von instruktionalen Bildern (Visualisierungen, Graphiken, Diagramme oder thematische Karten) zur Vermittlung von Wissen eingesetzt werden, die nicht mehr nur eine Dekorations- oder Situierungsfunktion haben, sondern als eine eigenständige, den Text ergänzende Informationsquelle fungieren. Text- und Bildinformationen müssen hier aufeinander bezogen und integrativ verarbeitet werden. Die Entwicklung dieser Kompetenz zur integrativen Verarbeitung von Text- und Bildinformationen bei Schüler/innen dürfte analog zur Kompetenzentwicklung in anderen Bereichen maßgeblich vom Unterricht ihrer Lehrkräfte beeinflusst werden.

Die integrative Verarbeitung von Texten und Bildern wird allerdings trotz ihrer allgegenwärtigen Bedeutung in den allgemeinbildenden Schulen bislang oft nicht systematisch gelehrt. Während realistische Bilder, wie z.B. Gemälde und Fotografien ein Erkennen des Abgebildeten anhand von Schemata der alltäglichen Wahrnehmung ermöglichen (vgl. Weidenmann 1994), benötigen logische Bilder wie z.B. Kreis-, Balken- oder Liniendiagramme und thematische Karten, die in Lehrmaterialien in der Regel eine deskriptive und/oder erklärende Funktion haben, ein spezielles Vorwissen über die betreffenden Darstellungskonventionen (vgl. Pinker 1990).

Ziel des BITE-Projekts ist die Überprüfung von Kompetenzmodellen zur Bild-Text-Integration auf Schüler- und Lehrerebene. Zum einen sollen Struktur und Entwicklung

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: SCHN 665/3-1 und BA 1461/7-1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

der Kompetenz zur integrativen Text-Bild-Verarbeitung bei Schüler/innen der Klassenstufen 5 bis 8 analysiert werden. Zum anderen soll untersucht werden, welche Kompetenzen zum Umgang mit Texten und Bildern im Unterricht Lehrkräfte unterschiedlicher Schulfächer besitzen und inwieweit diese Kompetenzen von Ausbildung und Berufserfahrung abhängig sind.

2. Theoretischer Ansatz

Kompetenzen können sowohl aus allgemein-kognitionswissenschaftlicher Sicht als auch aus differentieller Sicht analysiert werden, wobei jeweils unterschiedliche Strukturbegriffe zum Tragen kommen. Aus allgemein-kognitionswissenschaftlicher Sicht geht es zum einen um die Frage, welche Komponenten innerhalb der Architektur des kognitiven Systems bei der Ausübung einer Kompetenz beteiligt sind. Das Ergebnis einer solchen Analyse bezeichnen wir als Systemstrukturmodell. Ein Beispiel für ein Systemstrukturmodell wäre das Mehrspeichermodell des Gedächtnisses von Atkinson und Shiffrin (1968). Zum anderen geht es um die Frage, welche Prozesse einer Kompetenz zugrunde liegen und wie diese in bestimmter strukturierter Weise aufeinanderfolgen. Das Ergebnis einer solchen Analyse bezeichnen wir als Prozessstrukturmodell. Prozessstrukturmodelle können sich z.B. aus einer sog. rationalen Aufgabenanalysen ergeben, in denen die zur Bewältigung einer Anforderung erforderlichen Prozesse in eine sachlogisch notwendige Abfolge gebracht werden, wobei sich infolge von Inklusionsbeziehungen zugleich eine Lernhierarchie ergibt (vgl. Gagné 1968; Resnick/Wang/Kaplan 1973). Aus differenzieller Sicht steht die Frage im Vordergrund, wie sich die vorhandenen individuellen Unterschiede in der Ausprägung einer Kompetenz abbilden lassen. Hier geht es um die Struktur des zur Darstellung interindividueller Unterschiede erforderlichen Raumes, um die sog. Dimensionalität der betreffenden Kompetenz. Das Ergebnis einer solchen Analyse bezeichnen wir als metrisches Strukturmodell. Die Mehrdimensionalität des metrischen Strukturmodells ist ein hinreichender, jedoch nicht notwendiger Indikator für die qualitative Verschiedenheit der zugrunde liegenden Prozesse und beteiligten Komponenten des kognitiven Systems (vgl. Schnotz 1979): Strukturell und prozessual verschiedene Anforderungen im Rahmen einer Kompetenz bleiben auch dann verschieden, wenn die betreffenden Leistungen hoch korrelieren und sich die metrische Struktur der interindividuellen Unterschiede insofern als eindimensional erweist.

Im BITE-Projekt werden Prozessstrukturmodelle und metrische Strukturmodelle der Kompetenz zur integrativen Text-Bild-Verarbeitung entwickelt und überprüft. Ausgangspunkt dabei ist das von Schnotz und Bannert (2003) entwickelte integrative Modell des Text- und Bildverstehens (vgl. Schnotz 2005). Dieses basiert auf einer kategorialen Unterscheidung zwischen deskriptionalen und depiktionalen Repräsentationen. Texte, mentale sprachliche Oberflächenstrukturen und Propositionen gelten als deskriptionale Repräsentationen; Bilder, visuelle Vorstellungen und mentale Modelle gelten als depiktionale Repräsentationen. Dem Modell zufolge beinhaltet die integrative Verarbei-

tung von Bildern und Texten Strukturabbildungsprozesse. Da die abzubildenden Strukturen teilweise ineinander verschachtelt sind, lassen sich kognitive Hierarchieebenen der Verarbeitung unterscheiden, da die Abbildung hierarchieniedrigerer Strukturen Voraussetzung für die Abbildung hierarchiehöherer Strukturen ist (vgl. Wainer 1992). Dies soll im Folgenden an einem Beispiel veranschaulicht werden.

Abbildung 1 zeigt eine Text-Bild-Kombination aus dem Biologieunterricht, die den Aufbau eines Insektenbeins erläutert.[2] Eine sehr einfache Anforderung der Bild-Text-Integration wäre, dass die bzw. der Lernende bestimmen muss, wie man das letzte Glied eines Insektenbeines bezeichnet. Wir klassifizieren ein solches Ablesen von Detailinformation als eine Anforderung der kognitiven Hierarchiebene 1. Eine komplexere Anforderung wäre zu bestimmen, ob der Schenkel des Schwimmbeines kürzer ist als der Schenkel des Sprungbeins. Wir bezeichnen ein solches Ablesen einfacher Relationen als eine Anforderung der kognitiven Hierarchieebene 2. Eine noch komplexere Anforderung wäre die Beantwortung der Frage, ob das Laufbein eine längere Schiene als das Schwimmbein, aber eine kürzere Schiene als das Sprungbein hat. Wir bezeichnen ein solches Ablesen komplexer Relationen als eine Anforderung der kognitiven Hierarchieebene 3. Die verschiedenen Hierarchieebenen kennzeichnen somit logisch aufeinander

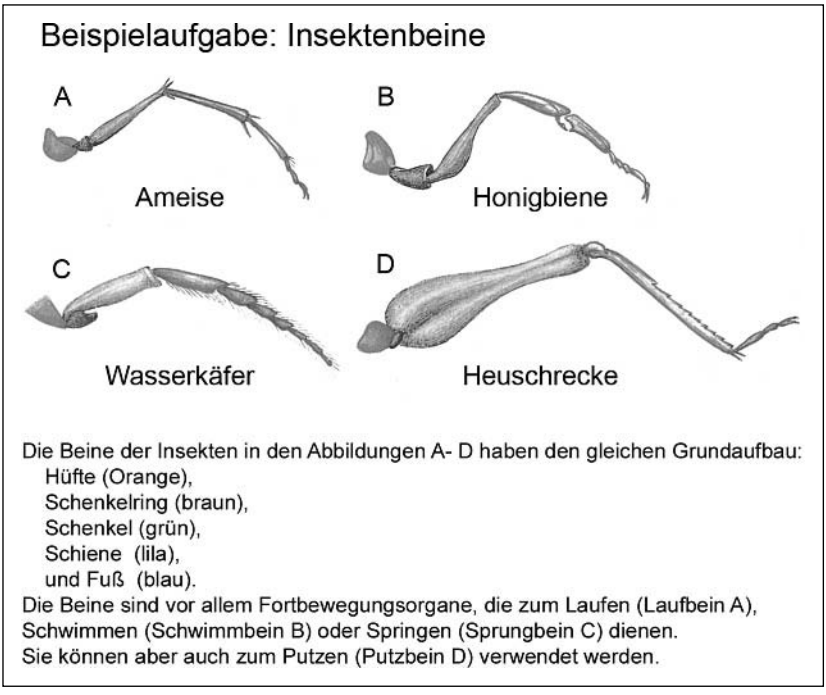


Abb. 1

2 Im Original sind im Bild die verschiedenen Teile des Beins nicht durch Farbnamen beschriftet, sondern in entsprechender Farbe gezeichnet. Im Text steht hinter jedem Namen des Bein-Teils nicht der Farbname, sondern eine entsprechende farbige Markierung.

aufbauende Komplexitätsstufen jeweils aufgabenspezifischer Strukturabbildungsprozesse innerhalb des depiktionalen Repräsentationszweiges, die von deskriptionalen Repräsentationen geleitet werden. Es sei noch einmal betont, dass es sich bei den Items verschiedener kognitiver Hierarchieebenen um Anforderungstypen hinsichtlich der notwendigen Strukturabbildung und nicht um Niveaustufen auf einer Dimension eines metrischen Modells handelt.

Die Kompetenz von Schüler/innen zur Bild-Text-Integration dürfte maßgeblich durch das Unterrichtshandeln von Lehrkräften beeinflusst werden. Die Förderung dieser Kompetenz verlangt von Lehrer/innen, bei Schüler/innen lernrelevante kognitive Prozesse anzuregen. Der Expertise der Lehrkräfte wird eine direkte Funktion für das unterrichtliche Handeln und die Lehr-Lern-Prozesse zugeschrieben (vgl. Krauss u.a. 2008). Als wesentliche Aspekte der Lehrerkompetenz gelten dabei Wissen, Einstellungen, Motivation und Selbstregulation (vgl. Kunter/Baumert 2006; Shulman 1987). Daher soll im Rahmen des Projekts überprüft werden, inwiefern die entsprechenden Lehrerkompetenzen im Bereich Bild-Text-Integration vermittelt über die Unterrichtsqualität Einfluss auf die Leistung und Motivation der Schüler/innen haben. Einen zentralen Aspekt der Lehrerkompetenz stellt die diagnostische Kompetenz bei der Einschätzung von Schülerfähigkeiten oder Aufgabenanforderungen, -potenzial und -schwierigkeiten dar (vgl. McElvany u.a. im Druck; Spinath 2005). Fehlende Akkuratheit der Einschätzung wie z.B. durch Überschätzung der Schülerleistung, das Nicht-Erkennen von schwachen Leser/innen oder Probleme bei der Unterscheidung von leichten und schweren Aufgaben können die Planung und Durchführung des Unterrichts nachhaltig negativ beeinflussen. Dementsprechend kann diagnostische Kompetenz im Bereich der Bild-Text-Integration als eine Voraussetzung zur gezielten Auswahl und Gestaltung von Texten mit Bildern und für eine Adaption des Unterrichts entsprechend unterschiedlicher Leistungsniveaus angesehen werden.

3. Forschungsfragen

Im BITE-Projekt stehen folgende Forschungsfragen im Vordergrund:

(a) Aus welchen systemstrukturellen und prozessualen Komponenten besteht die Kompetenz zur integrativen Text-Bild-Verarbeitung und wieweit schlagen sich diese in einer entsprechenden Dimensionalität nieder?
(b) Wieweit lassen sich die verschiedenen Komponenten voneinander abgrenzen und wieweit unterscheiden sie sich von der allgemeinen textbezogenen Lesekompetenz sowie den allgemeinen kognitiven Fähigkeiten?
(c) Welche Unterschiede bestehen zwischen Schüler/innen verschiedener Schularten und verschiedener Jahrgangsstufen?
(d) Über welche Kompetenzen verfügen Lehrkräfte unterschiedlicher Fächer und Schulformen zum Umgang mit Texten und Bildern im Unterricht und sind Ausbildung bzw. Berufserfahrung Moderatoren dieser Kompetenzen?

(e) Inwiefern nehmen die Kompetenzen der Lehrkräfte vermittelt über die Unterrichtsqualität Einfluss auf Leistung und Motivation der Schüler/innen beim Umgang mit Texten und Bildern?

4. Methodisches Vorgehen und Forschungsdesign

Zur Beantwortung der o.g. Forschungsfragen wurden bisher eine Pilotierungsstudie und die erste Erhebung einer Längsschnittstudie durchgeführt. Die Pilotierungsstudie diente zum einen als Grundlage für die Entwicklung der Messinstrumente und sollte zum anderen erste Hinweise auf den Entwicklungsstand der Kompetenz zur Bild-Text-Integration innerhalb der Klassenstufen 5 bis 8 in den verschiedenen Schularten geben. Die Längsschnittstudie soll über drei Messzeitpunkte anhand von zwei Kohorten (Kohorte A: Klassenstufe 5 bis 7, Kohorte B: Klassenstufe 6 bis 8) Aufschluss über die Entwicklung der Kompetenzen zur Bild-Text-Integration liefern. Im Folgenden werden nur die Ergebnisse der Pilotierungsstudie berichtet.

4.1 Aufgabenkonstruktion und Itemanalyse auf Schülerebene

Ausgehend von einer Analyse nahezu aller in der Bundesrepublik Deutschland verwendeten Schulbücher der Fächer Biologie und Geographie der Klassenstufen 5 bis 8 wurden 48 Aufgaben zur Bild-Text-Integration entwickelt. Jede Aufgabe bestand aus einem Aufgabenstamm aus einem kurzen Text (38 bis 160 Worte), 1–3 Bildern (Karten, schematischen Visualisierungen und Diagrammen) sowie 6 Multiple-Choice-Testitems mit jeweils 4 Antwortalternativen, von denen jeweils eine richtig war. Je zwei dieser Items stellten Anforderungen der Hierarchieebene 1, zwei stellten Anforderungen der Hierarchieebene 2 und zwei stellten Anforderungen der Hierarchieebene 3. Zur Beantwortung der Items mussten die Proband/innen Text- und Bildinformation aufeinander beziehen, da weder der Text alleine noch das Bild alleine die korrekte Beantwortung der Items ermöglichte. Die 48 Aufgaben mit insgesamt 48x6=288 Items wurden nach einem Youden-Design 60 Testheften zugeordnet, welche systematisch rotiert insgesamt 1060 Schüler/innen der Klassenstufen 5 bis 8 von zufällig ausgewählten Gymnasien, Realschulen und Hauptschulen des Landes Rheinland-Pfalz vorgegeben wurden. Pro Schulart-Klassenstufen-Kombination nahmen jeweils 4 zufällig ausgewählte Klassen an der Untersuchung teil, wobei pro Schule jeweils nur eine Klasse involviert war. Die von den Schüler/innen bearbeiteten Testitems wurden einer Itemanalyse aufgrund eines ein-parametrischen logistischen Modells (Rasch-Modell) unterzogen. Zur Auswahl modellkonformer Items für die Hauptuntersuchung wurden residuenbasierte Item-Fit-Statistiken verwendet. Zusätzlich wurden DIF-Analysen (Geschlecht, Klassenstufe und Schultyp) durchgeführt, um weitere problematische Items zu identifizieren und auszusortieren.

Es wurde erwartet, dass innerhalb jeder Aufgabe Strukturabbildungsprozesse höherer Ebene mehr Schwierigkeit bereiten als Strukturabbildungsprozesse niedrigerer

Ebene. Zur Überprüfung dieser Hypothese wurde für die bei der Itemanalyse ausgewählten Items die Korrelation zwischen der theoretisch angenommenen kognitiven Hierarchieebene je Aufgabe mit dem Rankplatz der empirischen Itemschwierigkeit innerhalb der jeweiligen Aufgabe korreliert. Die entsprechende Kontingenztafel ist in Tabelle 1 angegeben. Für den Zusammenhang beider Variablen ergab mit einem Kendalls Tau von .55 ein mittlerer Zusammenhang. Dabei ist zu berücksichtigen, dass infolge des Multimatrix-Designs vor allem bei den leichten Items bereits geringe Unterschiede in der Häufigkeit von Falschantworten zu einem Wechsel in der Rangposition der Schwierigkeiten führten. Außerdem mussten die Items zur Sicherung der stochastischen Unabhängigkeit so konstruiert werden, dass zwischen ihnen keine direkten semantischen Abhängigkeiten bestanden, weshalb die Items jeweils unterschiedlichen Hierarchien angehörten. Vor diesem Hintergrund ist der vorliegende Zusammenhang zwischen theoretischer Hierarchieebene und empirischer Schwierigkeit als befriedigend anzusehen. Die verschiedenen kognitiven Hierarchieebenen lassen sich insofern hinreichend voneinander abgrenzen.

	Empirische Rangstufe 1	**Empirische Rangstufe 2**	**Empirische Rangstufe 3**
Hierarchieebene 1	53	25	5
Hierarchieebene 2	23	37	26
Hierarchieebene 3	5	18	56

Anmerkung: In den Zellen der Tabelle ist die Häufigkeit angegeben, mit der die betreffende Kombination von theoretischer Hierarchieebene und Rangstufe der empirischen Schwierigkeit vorkommt. Bei einem perfekten Zusammenhang zwischen Hierarchieebene und empirischer Rangstufe wären nur die Diagonalzellen (von links oben nach rechts unten) besetzt.

Tab. 1: Zusammenhang zwischen theoretisch angenommener Hierarchieebene der Strukturabbildung und Rangstufe der empirischen Itemschwierigkeit pro Aufgabe über alle Items

Um die Frage nach der Dimensionalität der Kompetenz zur Bild-Text-Integration zu beantworten, wurden drei alternative hypothetische metrische Strukturmodelle überprüft: ein eindimensionales, ein zweidimensionales und ein dreidimensionales Modell. Im eindimensionalen Modell wurde angenommen, dass die Items der verschiedenen Hierarchieebenen so hoch korrelieren, dass eine Dimension zur Beschreibung der individuellen Unterschiede ausreicht.[3] Im zweidimensionalen Modell wurde angenommen, dass die Strukturabbildungsanforderungen bei Items der Hierarchieebene 1 (Ablesen von Detailinformation) sich von den Strukturabbildungsanforderungen bei Items der höheren Hierarchieebenen 2 und 3 (Ablesen von Relationen) soweit unterscheiden, dass

3 Kognitiv sehr unterschiedliche Anforderungen sind beispielsweise auch bei den PISA-Lesetestaufgaben und -items gegeben, die gleichwohl in ein eindimensionales metrisches Strukturmodell integriert werden können.

beide Itemgruppen jeweils eigenständige Dimensionen konstituieren. Im dreidimensionalen Modell wurde angenommen, dass die Strukturabbildungsanforderungen bei Items der kognitiven Hierarchieebene 1, bei Items der Hierarchieebene 2 und bei Items der Hierarchieebene 3 qualitativ soweit verschieden sind, dass die drei Itemgruppen jeweils eigenständige Dimensionen konstituieren.

Die Ergebnisse der Modellfit-Prüfungen sind in Tabelle 2 aufgeführt. Es zeigt sich, dass der BIC-Wert für das zweidimensionale Modell am geringsten, die Modellanpassung somit am höchsten ist. Demnach werden die vorliegenden Daten der Pilotierungsuntersuchung am besten durch ein metrisches Strukturmodell beschrieben, in dem einerseits eine Dimension für das Ablesen von Detailinformation und andererseits eine Dimension für das Ablesen von Relationen unterschieden werden. Die manifeste Korrelation zwischen beiden Dimensionen beträgt $r = .760$, die latente Korrelation $r = .950$. Für pragmatische Zwecke kann angesichts der Höhe der latenten Korrelation auch mit einer eindimensionalen Lösung operiert werden.

	–2 ln L	**Parameter**	**BIC**
Eindimensional	44841,7	289	46854,9
Zweidimensional	44819,3	291	46846,4
Dreidimensional	44815,6	294	46863,7

Anmerkung: Um die Güte der verschiedenen Modelle miteinander zu vergleichen, wurde das Bayes Information Criterion (BIC) ausgewählt, bei dem die logarithmierte Likelihood mit der Anzahl der verwendeten Parameter gewichtet wird. Je niedriger der BIC-Wert, umso besser ist die Passung des Modells.

Tab. 2: Ergebnisse der Modellanpassungsprüfung für das ein-, das zwei- und das dreidimensionale metrische Strukturmodell

4.2 Skalenkonstruktion und Fragebogenerhebung auf Lehrerebene

Für die Befragungen der Lehrer/innen wurden Fragebogen zur Erhebung der Kompetenzen von Lehrkräften (Wissen, Einstellungen, Motivation, diagnostische Fähigkeiten) zur Thematik der Bild-Text-Integration entwickelt. Dabei betrafen die pädagogischen Einstellungen die Überzeugungen der Lehrkräfte hinsichtlich der Wichtigkeit von Abbildungen für den Unterricht, die Bedeutung von Übung im Umgang mit Bildern und Texten, die Bedeutung der Vermittlung von Strategien der Bild-Text-Integration sowie der Selbständigkeit im Umgang mit Bildern und Texten. Die Skalen bestanden aus jeweils 3–4 Items und besaßen eine interne Konsistenz zwischen .67 und .85 (Cronbachs α). Befragt wurden die Biologie- (33 Beantwortungen), Geographie- (33 Beantwortungen) und Deutschlehrer/innen (42 Beantwortungen) der untersuchten Klassen.

Die Lehrer/innen, welche die Fragebogen beantwortet hatten, unterrichteten insgesamt 856 Schüler/innen der untersuchten Schülerstichprobe. Um Hinweise auf Art und

Qualität des Unterrichts hinsichtlich des Umgangs mit Texten und Bildern zu erhalten, wurden die Schüler/innen unter anderem nach ihrer Motivation beim Lernen mit Texten und integrierten Bildern im Unterricht befragt. Die interne Konsistenz der aus 4 Items bestehenden Skala betrug .79 (Cronbachs α). Innerhalb der untersuchten Klassen bewertet jeweils ein Drittel der Schüler/innen eine der drei o.g. Lehrkräfte.

5. Ergebnisse der Pilotierungsstudie

5.1 Ergebnisse auf der Schülerebene

Die Mittelwerte der Schülerleistungen in den einzelnen Klassenstufen und Schularten sind in Abbildung 2 grafisch dargestellt.[4] Da es sich hier um Daten einer Querschnittanalyse von verschiedenen Kohorten handelt, handelt es sich hier nicht um tatsächliche Entwicklungsverläufe. Allerdings können erhebliche Leistungsunterschiede zwischen den verschiedenen Schularten und den verschiedenen Klassenstufen festgestellt werden. Eine Kovarianzanalyse mit den Faktoren „Klassenstufe" und „Schulart" sowie der Kovariaten „Kognitive Fähigkeiten" erbrachte für alle drei Variablen signifikante Unterschiede hin-

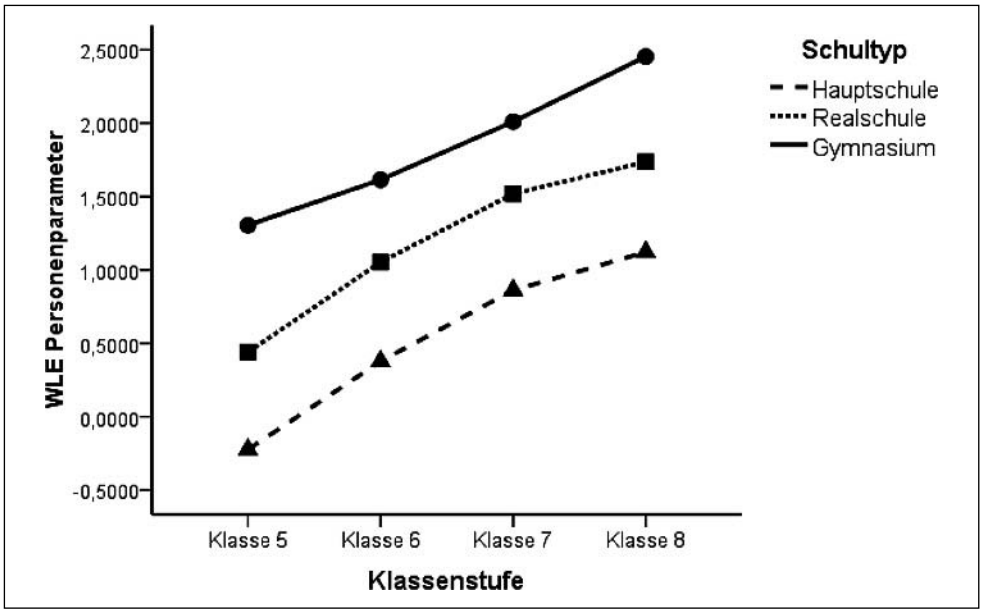


Abb. 2: Mittelwerte der rasch-skalierten Schülerleistungen zur Bild-Text-Integration in den Klassenstufen 5 bis 8 in der Hauptschule, der Realschule und im Gymnasium

4 Bei den hier berichteten Kompetenzunterschieden wird der Einfachheit halber auf das eindimensionale metrische Modell zurückgegriffen.

sichtlich der Kompetenz zur Bild-Text-Integration mit insgesamt relativ hohen Anteilen an aufgeklärter Varianz. Die Effektstärken betrugen $\eta^2 = 26{,}4\%$ für den Faktor „Klassenstufe“, $\eta^2 = 25{,}4\%$ für den Faktor „Schulart“ sowie $\eta^2 = 15{,}8\%$ für die Kovariate „Kognitive Fähigkeiten“. Es fand sich keine nennenswerte Interaktion zwischen Klassenstufe und Schulart. D.h.: Die Kompetenzunterschiede zwischen aufeinanderfolgenden Klassenstufe waren in allen drei Schularten im Wesentlichen die gleichen. Bemerkenswert ist, dass die mittleren Leistungen von Hauptschüler/innen in der Klassenstufe 8 mit 0,87 ($SD = 0{,}72$) im Vergleich zu den mittleren Leistungen von Gymnasiast/innen der Klassenstufe 5 mit 1,58 ($SD = 0{,}96$) hochsignifikant geringer sind ($t(156{,}9) = 5{,}47$; $p < .001$).

5.2 Ergebnisse auf der Lehrerebene

Es zeigte sich, dass die Schüler/innen einen Unterricht präferierten, der weitgehend störungsfrei abläuft und genügend Zeit zur Klärung von Verständnisproblemen lässt. Ein solcher Unterricht wurde eher von Lehrkräften gegeben, die die Bedeutung der Strategievermittlung bei der Befragung zu ihren Einstellungen betonten. Allerdings wurde auch deutlich, dass die geäußerten Überzeugungen der Lehrkräfte nicht immer mit ihrem tatsächlichen Unterrichtshandeln konform waren: So integrierten Lehrkräfte, welche Abbildungen in den Befragungen für besonders wichtig hielten, diese nach Schüleraussage weniger in ihren Unterricht. Auch ergaben sich Hinweise darauf, dass die Schwierigkeit von Bildern häufig unterschätzt wird und dass diese eher als Mittel zur Illustration angesehen werden.

6. Diskussion

Den dargestellten Ergebnissen zufolge hat die Beschulung von der 5. bis zur 8. Klassenstufe nur relativ moderate Effekte auf die Entwicklung der Kompetenz zur Bild-Text-Integration. Die mittlere Kompetenz von Hauptschüler/innen in der 8. Klassenstufe scheint immer noch unter der mittleren Kompetenz von Gymnasiast/innen in der 5. Klassenstufe zu liegen. Kognitive Fähigkeiten scheinen hier eher eine untergeordnete Rolle zu spielen. Die Kompetenzentwicklung in den verschiedenen Schularten scheint innerhalb klar getrennter Bereiche stattzufinden, ohne dass dies durch unterschiedliche kognitive Fähigkeitsniveaus hinreichend erklärbar wäre. Die Ergebnisse bieten zu der Vermutung Anlass, dass eine systematische Förderung der Kompetenz zur Bild-Text-Integration im Unterricht eher selten stattfindet und die beobachtbare Kompetenzsteigerung eher ein Begleitphänomen anderer Lehr-Lernprozesse ist.

Während der Förderung der Lesefähigkeit im Schulalltag zu Recht viel Aufmerksamkeit gewidmet wird, erhält das Lesen von Bildern, Diagrammen und anderen Visualisierungen in den meisten Lehrplänen offenbar vergleichsweise wenig Beachtung. Tatsächlich erfordert eine erfolgreiche visuelle Wissenskommunikation mit Hilfe von Bildern komplexe Fähigkeiten (vgl. Bertin 1981; Wainer 1992). Die integrative Ver-

arbeitung von Bildern und Texten stellt insofern noch höhere Anforderungen, da hier ein strategischer Umgang mit den wechselseitigen „constraints“ eine wichtige Rolle spielt.

Auf der Seite der Lehrkräfte erscheint es sinnvoll, das Bewusstsein für die Komplexität von Anforderungen der Bild-Text-Integration zu verstärken und die Sensibilität für die dabei zu überwindenden Schwierigkeiten zu fördern. Nicht nur für Schüler/innen, sondern auch für Lehrer/innen gilt, dass die integrierte semantische Verarbeitung von Bildern und Texten spezifische Verarbeitungsstrategien erfordert, die als Kulturtechnik erworben, verstanden und eingeübt werden müssen. Dies scheint im Unterricht eher sporadisch zu geschehen. Insgesamt gesehen ist die Entwicklung der Kompetenz zur Bild-Text-Integration zwar eine immer wichtigere Voraussetzung für eine erfolgreiche Bewältigung von Alltagsanforderungen und für Bildungsprozesse. Die Herausbildung dieser Kompetenz scheint bislang allerdings eher ein Nebenprodukt des schulischen Lehrens und Lernens als das Ergebnis systematischer Unterrichtsbemühungen zu sein.

Literatur

Atkinson, R.C./Shiffrin, R.M. ([2]1968): Human memory: A proposed system and its control processes. In: Spence, K.W./Spence, J.T. (Hrsg.): The psychology of learning and motivation. London: Academic Press, S. 89–195.

Bertin, J. (1981): Graphics and graphic-information-processing. Berlin: Walter de Gruyter.

Gagné, R.M. (1968): Learning hierarchies. In: Educational Psychologist 6, S. 3–6.

Krauss, S./Brunner, M./Kunter, M./Blum, W./Jourdan, A./Neubrand, M./Baumert, J. (2008): Pedagogical content knowledge and content knowledge of secondary mathematics teachers. In: Journal of Educational Psychology 100, H. 3, S. 716–725.

Kunter, M./Baumert, J. (2006): Linking TIMSS to research on learning and instruction: A re-analysis of the German TIMSS and TIMSS Video data. Chapter 21. In: Howie, S.J./Plomp, T. (Hrsg.): Contexts of learning mathematics and science: Lessons learned from TIMSS. London u.a.: Routledge, S. 335–351.

McElvany, N./Schroeder, S./Richter, T./Hachfeld, A./Baumert, J./Schnotz, W./Horz, H./Ullrich, M. (im Druck): Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung von Schülerfähigkeiten und Aufgabenschwierigkeiten bei Lernmedien mit instruktionalen Bildern. In: Zeitschrift für Pädagogische Psychologie.

Pinker, S. (1990): A theory of graph comprehension. In: Freedle, R. (Hrsg.): Artificial intelligence and the future of testing. Hillsdale: Erlbaum, S. 73–126.

Resnick, L.B./Wang, M.C./Kaplan, J. (1973): Task analysis in curriculum design. A hierarchically sequenced introductory methematics curriculum. In: Journal of Applied Behavior Analysis 6, S. 679–710.

Schnotz, W. (1979): Lerndiagnose als Handlungsanalyse. Weinheim: Beltz.

Schnotz, W. (2005): An Integrated Model of Text and Picture Comprehension. In: Mayer, R.E. (Hrsg.): Cambridge Handbook of Multimedia Learning. Cambridge: Cambridge University Press, S. 49–69.

Schnotz, W./Bannert, M. (2003): Construction and interference in learning from multiple representations. In: Learning and Instruction 13, S. 141–156.

Shulman, L. (1987): Knowledge and Teaching: Foundations of the New Reform. In: Harvard Educational Review 57, S. 1–22.

Spinath, B. (2005): Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. In: Zeitschrift für Pädagogische Psychologie 19, S. 85–95.
Wainer, H. (1992): Understanding graphs and tables. In: Educational Researcher 21, H. 1, S. 14–23.
Weidenmann, B. (Hrsg.) (1994): Wissenserwerb mit Bildern. Bern: Hans Huber.

Anschrift der Autor/innen

Prof. Dr. Wolfgang Schnotz, Universität Koblenz-Landau, Arbeitseinheit Allgemeine und Pädagogische Psychologie, Thomas-Nast-Str. 44, D-76829 Landau
E-Mail: schnotz@uni-landau.de

Prof. Dr. Holger Horz, Fachhochschule Nordwestschweiz, Hochschule für Angewandte Psychologie, Institut für Kooperationsforschung und -entwicklung, Riggenbachstrasse 16, CH-4600 Olten
E-Mail: holger.horz@fhnw.ch

Prof. Dr. Nele McElvany, Institut für Schulentwicklung (IFS), Technische Universität Dortmund, Vogelpothsweg 78, D-44227 Dortmund
E-Mail: mcelvany@ifs.tu-dortmund.de

Dr. Sascha Schroeder, Max-Planck-Institut für Bildungsforschung, Lentzeallee 94, D-14195 Berlin
E-Mail: sascha.schroeder@mpib-berlin.mpg.de

Dipl.-Psych. Mark Ullrich, Universität Koblenz-Landau, Arbeitseinheit Allgemeine und Pädagogische Psychologie, Thomas-Nast-Str. 44, D-76829 Landau
E-Mail: ullrichm@uni-landau.de

Prof. Dr. h.c. mult. Jürgen Baumert, Max-Planck-Institut für Bildungsforschung, Lentzeallee 94, D-14195 Berlin
E-Mail: sekbaumert@mpib-berlin.mpg.de

Dipl.-Psych. Axinja Hachfeld, Max-Planck-Institut für Bildungsforschung, Lentzeallee 94, D-14195 Berlin
E-Mail: hachfeld@mpib-berlin.mpg.de

Dr. Tobias Richter, Universität zu Köln, Lehrstuhl Allgemeine Psychologie II, Bernhard-Feilchenfeld-Str. 11, D-50969 Köln
E-Mail: tobias.richter@uni-koeln.de

Tobias Dörfler/Stefanie Golke/Cordula Artelt

Dynamisches Testen der Lesekompetenz

Theoretische Grundlagen, Konzeption und Testentwicklung

Projekt Dynamisches Testen[1]

1. Einleitung

Dynamische Tests gelten insbesondere im Bildungssektor als Alternative zu herkömmlichen Tests. Für den Bereich der Lesekompetenz fehlen bislang jedoch gesicherte Erkenntnisse, die die diagnostische Güte eines dynamischen Lesekompetenztests belegen können. Der hier vorliegende Beitrag bietet einen theoretischen Überbau für die Entwicklung eines solchen dynamischen Lesekompetenztests und informiert zudem (nicht immer chronologisch) über grundlegende Schritte der Testentwicklung sowie erste Ergebnisse der Testkonstruktion.

2. Theoretischer Rahmen

2.1 Feedback und (meta-) kognitive Hilfen als Elemente dynamischer Tests

Kognitive Fähigkeiten sind nicht direkt beobachtbar, sondern müssen aus Verhaltensindikatoren erschlossen werden. Hierzu werden Personen in der Regel mit einer Reihe von Testaufgaben konfrontiert, deren Bearbeitung unter Berücksichtigung eines Messfehlers Aufschluss über die Ausprägung ihrer zugrunde liegenden Fähigkeit(en) geben soll. Beachtenswert ist, dass sich auf diese Weise nur derjenige Anteil der Leistungsfähigkeit abbilden lässt, der bereits in ausreichendem Maße im Verhalten der Proband/innen realisiert wurde. Aussagen über mögliche Leistungsreserven können dadurch nicht getroffen werden. Ist neben der Diagnostik der aktuellen Leistung zusätzlich die Abschätzung von individuellen Leistungsreserven bzw. einer Lernfähigkeit im untersuchten Bereich von Interesse, kommen dynamische Tests zum Einsatz (vgl. Guthke/Beckmann/Wiedl 2003).

Die Kernidee zur Erfassung der Lernfähigkeit geht auf Wygotski (1964) zurück, der konstatiert, dass eine zuverlässige Diagnose nur unter Beachtung zweier sogenannter Entwicklungszonen erfolgen kann. Nach seiner Auffassung muss für die Bewertung der Leistungsfähigkeit einer Person neben dem aktuellen Entwicklungsstand („Zone der ak-

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: AR 301/7-2) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

tuellen Entwicklung“) zusätzlich ihre Entwicklungspotenz („Zone der nächsten Entwicklung“) erfasst werden; also die Fähigkeit eines Individuums, unter förderlichen Bedingungen (z.B. durch Hilfe von Erwachsenen, fähigeren Peers oder computergestützten tutoriellen Systemen) Leistungsverbesserungen zu erzielen. Diese Entwicklungspotenzen können auch durch einen simulierten Lernvorgang im Testprozess – der durch gezielte Feedbacks und spezifische Lern- und Denkhilfen von außen gesteuert wird – „sichtbar“ gemacht werden. Das Ausmaß, in dem ein/e Proband/in von den angebotenen Feedbacks und Hilfen profitiert, gilt als Indikator der intellektuellen Lernfähigkeit.

Feedback kann in diesem Zusammenhang als Informationsangebot an eine Person bezüglich ihrer Leistung oder Verstehensprozesse beschrieben werden, mit dem Ziel, regulierend auf zukünftige Prozesse einzuwirken (vgl. Hattie/Timperley 2007). Es bietet mit einer zumeist mittleren Effektstärke eine erfolgversprechende Möglichkeit zur Leistungsverbesserung (vgl. Bangert-Drowns u.a. 1991; Kluger/DeNisi 1996). Die Qualität der rückgemeldeten Information reicht nach Narciss (2006) von einfachen bis hin zu sogenannten elaborierten Feedback-Arten.

Elaborierte Feedback-Arten, die beispielsweise fehlerspezifische Korrekturhinweise oder Hinweise auf kognitive bzw. metakognitive Strategien bieten, sind – unabhängig vom untersuchten Kompetenzbereich – zur Leistungsverbesserung besonders geeignet (Shute 2008). Während kognitive Strategien bei einfachen und komplexen Anforderungen gleichermaßen bedeutsam sind, gewinnen metakognitive Strategien mit zunehmender Komplexität der Aufgabe an Bedeutung. Als Schlüssel hierfür gelten übergeordnete Strategien der Planung, Überwachung und Bewertung und die darauf basierende Regulation des eigenen Lernvorgangs. Zentrale Momente metakognitiver Aktivitäten sind zum einen die Reflexion über den eigenen Lernprozess und zum anderen sind es die durch diese Reflexionen ausgelösten strategischen Aktivitäten. Die sichere Anwendung lösungsrelevanter kognitiver und metakognitiver Strategien ist beim verstehenden Lesen von besonderer Relevanz (vgl. Artelt/Baumert/Julius-McElvany 2003).

2.2 Lesekompetenz/Textverstehen als Anwendungsfeld dynamischer Tests

Das Konzept des Dynamischen Testens ist im Bereich der Intelligenzdiagnostik am weitesten verbreitet, es kann jedoch nach Guthke, Beckmann und Wiedl (2003) auch auf andere Kompetenzbereiche übertragen werden. Lesen umfasst kognitive Anforderungen, für die ähnliche Annahmen getroffen werden können wie für die verstärkt im Rahmen dynamischer Tests untersuchten Facetten der Intelligenz: Das Verstehen von Texten ist wie das Lösen von Aufgaben in einem Intelligenztest eine komplexe kognitive Anforderung, die durch Hilfen und Rückmeldungen unterstützt werden kann.

Textverstehen ist ein partiell automatisch ablaufender, aktiver Interaktionsprozess zwischen Leser/in und Text. Dabei werden, teilweise unter Nutzung des vorhandenen Vorwissens, verschiedene kognitive Repräsentationen des Textes aufgebaut, die nicht unabhängig voneinander sind (vgl. Kintsch 1998): Bei der wörtlichen Repräsentation stehen Oberflächenmerkmale des Textes im Vordergrund, die ihrerseits eine wesentliche

Voraussetzung für die semantische und syntaktische Analyse der Textinhalte bilden. Wird diese (textbasierte) propositionale Repräsentation des Geschriebenen durch zusätzliche (textferne) Propositionen angereichert, entsteht eine situative Repräsentation des Textes, ein sogenanntes Situationsmodell. Die Anreicherung mit Propositionen, die keine direkte Entsprechung im Text haben, geschieht unterdessen über inferenzielle und vorwissensabhängige Prozesse (vgl. Richter/Christmann 2002).

Inferenzen werden allerdings bereits auf hierarchieniedrigen Repräsentationsebenen beim Textverstehen gebildet. Hierzu zählen beispielsweise lokale Kohärenzbildungen im Sinne von identischen, synonymen oder abstrakten Wortwiederholungen. Zu den hierarchiehohen Verarbeitungsprozessen beim Lesen wird z.B. der Prozess der globalen Kohärenzbildung gezählt, der u.a. durch thematische und elaborative Inferenzen gekennzeichnet ist (vgl. Graesser/Singer/Trabasso 1997). Die erfolgreiche Generierung von Inferenzen ist laut Oakhill und Garnham (1988) ein anspruchsvoller – und im Falle hierarchiehoher Prozesse auch kapazitätsintensiver – intellektueller Prozess, der eine wesentliche Bedingung für verstehendes Lesen darstellt: Gute Leser/innen unterscheiden sich von schlechten Lesern/Leserinnen u.a. auch durch die Anzahl der beim Lesen gezogenen Inferenzen (vgl. ebd.; Oakhill/Cain 2004).

Ein tieferes Verstehen eines Textes, der nicht allein auf leicht verfügbarem Weltwissen beruht, bedarf zusätzlich der intentionalen und strategischen Steuerung des Lesevorgangs (vgl. Coté/Goldman 1999). Der Aufbau einer kohärenten Textrepräsentation hängt – besonders bei längeren und schwierigen Texten – stark davon ab, ob und in welchem Maße kognitive und metakognitive Strategien effektiv verwendet werden. Hierzu zählt etwa die kontinuierliche Überprüfung des Verständnisses des gelesenen Textes und die entsprechende Regulation des Leseprozesses, z.B. durch das erneute Lesen einzelner Textstellen oder den gezielten Abruf von Informationen aus dem Langzeitgedächtnis (vgl. Kintsch 1998). Die nicht prinzipiell automatisch ablaufenden Prozesse beim verstehenden Lesen werden unter dem Strategiebegriff subsumiert. Die Unterstützung der Leserin/des Lesers bei der strategischen Bearbeitung eines Textes ist daher Ziel zahlreicher Trainingsprogramme (vgl. z.B. Souvignier/Mokhlesgerami 2006).

Bei der Konstruktion eines dynamischen Lesekompetenztests sind die Erkenntnisse aus der Feedback- und Trainingsliteratur ebenso zu berücksichtigen wie die Besonderheiten der Entwicklung von Testaufgaben in diesem Bereich. Der folgende Abschnitt liefert eine Zusammenfassung zur theoretischen Konzeption eines solchen dynamischen Lesekompetenztests.

3. Konzeption eines dynamischen Lesekompetenztests

Bei einem dynamischen Test der Lesekompetenz müssen Testaufgaben zwei spezifische Anforderungen erfüllen: Einerseits müssen sie Prozesse des Textverstehens abbilden, um eine valide Diagnose der Textverstehensleistung zu gewährleisten; andererseits müssen die Aufgaben zusätzlich genügend Möglichkeiten zur Implementierung von passenden Feedbacks und Hilfen bieten.

Um beiden Anforderungsbereichen im hier vorgestellten Projekt gerecht zu werden, wurden Texte und Aufgaben teils adaptiert, teils neu konstruiert. Die Konstruktion und Auswahl von Aufgaben zur Messung des Textverstehens erfolgte angelehnt an die o.g. Unterscheidung nach hierarchieniedrigen und -hohen Anforderungen und dabei insbesondere an inferenziellen Prozessen (vgl. Kintsch 1998), die im Folgenden näher erläutert werden. Im Test werden Indikatoren lokaler Kohärenzbildung (Herstellung eines Sinnzusammenhangs zwischen aufeinanderfolgenden Propositionen oder Sätzen), Indikatoren globaler Kohärenzbildung (Herstellung eines Sinnzusammenhangs zwischen größeren Textteilen auf höherer Abstraktionsebene) und Indikatoren aktiv-konstruktiver Prozesse zur Bildung einer inhaltsspezifischen und anschaulichen Repräsentation des im Text beschriebenen Sachverhalts (Situationsmodell) abgebildet.

Bezogen auf diese Textverstehensprozesse orientierte sich die Konstruktion der Feedbacks und (meta-)kognitiven Hilfen an zwei Forschungsrichtungen. Einerseits wurden Erkenntnisse aus dem Bereich der Feedbackforschung, die sich mit Auswirkungen gezielter Rückmeldungen auf die Testleistung beschäftigt (vgl. u.a. Bangert-Drowns u.a. 1991; Hattie/Timperley 2007; Kluger/DeNisi 1996; Shute 2008), genutzt. Feedback scheint demnach dann am wirksamsten zu sein, wenn es spezifische Fehler korrigiert bzw. verständnisfördernd wirkt. Andererseits musste die spezifische Domäne – Lesekompetenz/Textverstehen – bei der Konstruktion von Rückmeldungen beachtet werden. Um eine Verständnisförderung beim Lesen durch Feedbacks bzw. Hilfen zu erreichen, sollten jene inferenziellen und strategischen Prozesse unterstützt werden, die für die Bildung einer kohärenten Textrepräsentation bezogen auf die jeweiligen Aufgabentypen bedeutsam sind (vgl. z.B. Graesser/Singer/Trabasso 1997). Die notwendigen Erkenntnisse zur Wirksamkeit unterschiedlicher Feedbackarten für die konstruierten Textverstehensanforderungen konnten bereits in einer experimentellen Untersuchung gewonnen werden (vgl. Golke/Dörfler/Artelt 2009).

3.1 Testentwicklung

Die bislang geleistete Testentwicklung diente primär dem Ziel, für das o.g. Experiment einen angemessenen Lesekompetenztest zur Verfügung zu stellen. Da diese vorläufige Testkonstruktion auch zur Überprüfung der Aufgabenqualität im Sinne einer Pilotstudie herangezogen werden kann, wird auf die Ergebnisse der ersten Phase der Testentwicklung nachfolgend genauer eingegangen.

Nach der Auswahl von Textmaterial und der Auswahl und Konstruktion entsprechender Items (s.u.) wurden zunächst kognitive Interviews mit Schüler/innen der 6. Klassenstufe durchgeführt, mit dem Ziel, Texte und Aufgaben hinsichtlich ihrer Angemessenheit für diesen Altersbereich zu überprüfen und mögliche Verständnisschwierigkeiten zu eruieren. Insgesamt wurden dabei 20 Schüler/innen mit Texten und Aufgaben konfrontiert und nach der Wirksamkeit eingesetzter Feedbacks befragt. Die Interviews dauerten etwa 1 Stunde. Hierbei wurden Fragen zum allgemeinen Verständnis des Textes und der Aufgaben gestellt und die Schüler/innen aufgefordert, laut bei der Bearbei-

Abb. 1: Beispielhafte Darstellung einer Lesekompetenzaufgabe

tung der Materialien zu denken. Die Durchführung der kognitiven Interviews erfolgte in Anlehnung an Lewis und Reiman (1993) sowie Willis (2004).

Für die Testentwicklung wurden drei expositorische und zwei narrative Texte mit insgesamt 37 Aufgaben verwendet. Es wurde ein computerbasiertes Multiple-Choice-Format mit fünf Antwortalternativen umgesetzt. Das Testformat wird in Abbildung 1 schematisch dargestellt. Während der Bearbeitung der Aufgaben war der jeweilige Text stets verfügbar. Vor- und Zurückblättern zwischen Texten und Aufgaben war nicht möglich.

3.1.1 Stichprobe

Die neu entwickelten Texte und Aufgaben wurden insgesamt 566 bayerischen Schüler/innen und Schülern vorgelegt. Die Schüler/innen verteilten sich hinreichend gleichmä-

ßig auf die drei Schulformen Haupt-, Realschule und Gymnasium. Beide Geschlechter waren zu gleichen Anteilen in der Stichprobe vertreten.

Skalierung

In den folgenden Abschnitten werden erste Ergebnisse der Testentwicklung berichtet. Hierbei werden insbesondere teststatistische Kennwerte dargestellt, die über die Güte des Testprototypens informieren sollen.

Itemparameter. Aus der bisherigen Testentwicklung resultierten 37 Aufgaben. Bei der Analyse der Items wird angenommen, dass alle Anforderungen eine zugrunde liegende Fähigkeitsdimension (Lesekompetenz) abbilden. Zur Modellierung der Antwortwahrscheinlichkeiten wird daher ein eindimensionales Raschmodell mit der Software ConQuest (vgl. Wu/Adams/Wilson 1997) geschätzt. Zur Bewertung der Items werden Itemfit-Kennwerte (Weighted Mean Square Residualwerte (MNSQ), *T*-Werte), Itemtrennschärfeparameter und Itemschwierigkeiten herangezogen. Zudem können erwartete und beobachtete itemcharakteristische Kurven (ICC) verglichen werden, deren gleichförmiger Verlauf einen guten Indikator für die Modellpassung in allen untersuchten Fähigkeitsbereichen liefert. Die Reliabilität der Gesamtskala liegt bei $\alpha = .76$.

Die Items liefern insgesamt zufriedenstellende itemstatistische Kennwerte (Tabelle 1). Lediglich vier Items zeigen einen leicht auffälligen gewichteten MNSQ-Wert. Sowohl die Verläufe der ICCs der vier Items als auch die herkömmlichen Itemschwierigkeits- und Trennschärfeparameter sind jedoch unauffällig. Drei weitere Aufgaben zeigen bei unbedenklichen Fitindizes unterdurchschnittlich niedrige Trennschärfen.

Anhand der gemeinsamen Verteilung der Personen- und Aufgabenparameter kann festgestellt werden, dass zahlreiche Aufgaben für die untersuchten Sechstklässler/innen tendenziell zu schwierig waren. Dieser unter herkömmlichen Testbedingungen kritische Umstand ist jedoch für weitere Projektschritte notwendig: Um Feedbacks in den Testprozess zu implementieren, sind tendenziell zu schwierige Items besser geeignet als Items mit optimaler Schwierigkeitspassung. Ziel der Rückmeldungen im Zuge dynamischer Tests ist es schließlich, Leistungsreserven abzubilden. Dazu ist es notwendig, intellektuell fordernde Aufgaben anzubieten, um den Testpersonen die Möglichkeit zur Leistungssteigerung einzuräumen.

Differenzielle Itemfunktionen. Die bislang entwickelten Aufgaben sind mit der Annahme eines eindimensionalen Lesekompetenztests vereinbar. Als weitere Voraussetzung für die Güte des konstruierten Tests wurde geprüft, ob die Aufgaben wie intendiert alle in gleicher Weise Fähigkeitsunterschiede zwischen Personen abbilden, oder ob in Abhängigkeit vom Geschlecht oder der Schulform über die Unterschiede im Gesamttest hinausgehende Vor- oder Nachteile bei einzelnen Aufgaben existieren (sog. Differenzielle Itemfunktionen, DIF). Die DIF-Analyse nach Geschlecht ergab, dass es bei einem Item deutliche und über die im Gesamttest zu findenden Unterschiede (s.u.) zwischen Jungen und Mädchen hinausgehende Vorteile zugunsten der Mädchen gab (−0.53 logits), was auf ein aus diagnostischer Perspektive schlechtes, da unfaires Item hindeutet. Die DIF-Analysen hinsichtlich der erfassten Schulformen ergeben beim Vergleich der beiden Extreme Hauptschule und Gymnasium bei sechs der 37 Aufgaben einen über die

Item	Estimate	MSNQ	*T*	p_i	r_{it}
1	0.61	1.03	0.7	27.48	0.26
2	–0.76	1.01	0.5	57.02	0.32
3	0.73	1.03	0.5	25.40	0.25
4	0.44	1.10	2.4	30.73	0.13
5	0.09	0.98	–0.7	38.08	0.38
6	0.29	1.08	2.1	33.81	0.18
7	–0.05	0.94	–2.3	41.13	0.44
8	1.25	1.10	1.3	17.20	0.04
9	–0.24	0.96	–1.4	45.21	0.40
10	–0.22	0.95	–2.0	44.86	0.43
11	0.62	0.96	–0.8	27.30	0.38
12	0.23	1.09	2.5	35.11	0.18
13	0.57	1.04	1.0	28.19	0.23
14	–0.64	0.95	–1.9	54.43	0.43
15	–0.49	0.98	–0.9	50.89	0.39
16	0.55	1.06	1.5	28.72	0.19
17	–0.66	0.94	–2.5	54.96	0.46
18	–0.06	1.09	3.1	41.31	0.18
19	0.59	1.10	2.2	27.84	0.12
20	–0.63	0.94	–2.5	54.08	0.43
21	–0.67	0.88	–4.8	55.14	0.53
22	–0.37	1.03	1.1	48.23	0.29
23	0.09	0.99	–0.4	37.94	0.35
24	–0.52	0.97	–1.4	51.77	0.39
25	–0.49	0.93	–3.0	51.06	0.47
26	0.24	1.03	1.0	34.75	0.26
27	0.29	1.05	1.4	33.69	0.22
28	–0.33	1.01	0.4	47.34	0.31
29	0.07	0.98	–0.8	38.48	0.38
30	0.26	1.01	0.2	34.51	0.31
31	0.19	0.99	–0.3	35.93	0.35
32	–0.28	0.98	–0.9	46.19	0.38
33	0.31	1.01	0.4	33.45	0.29
34	–0.86	0.97	–0.9	59.40	0.38
35	–0.02	0.94	–2.1	40.50	0.45
36	–0.56	0.91	–3.7	52.75	0.50
37	0.44	1.03	0.7	30.73	0.28

Anmerkungen: Estimate = Itemschwierigkeit (aus Raschmodell), MNSQ = weighted mean square, *T* = Wert aus *T*-Verteilung, p_i = Itemschwierigkeit (nach Klassischer Testtheorie), r_{it} = Itemtrennschärfe

Tab. 1: Itemkennwerte

im Gesamttest vorhandenen Unterschiede hinausgehenden relativen Vorteil für Gymnasialschüler/innen und bei fünf Aufgaben einen Vorteil der Hauptschüler/innen (Differenz von mehr als 0.5 logits). In den Test werden perspektivisch nur solche Aufgaben aufgenommen, die keine oder nur geringe differenziellen Itemfunktionen aufweisen.

Personenparameter. Im Folgenden werden die im Gesamttest gemessenen Mittelwertunterschiede in den Personenfähigkeiten zwischen Schüler/innen als auch zwischen den drei Schulformen berichtet. Diese Analysen werden auf Basis derjenigen Items ohne differenzielle Itemfunktionen berechnet. Ziel ist es zu überprüfen, ob der Test die theoretisch anzunehmenden und/oder in der Literatur berichteten Unterschiede für den Bereich der Lesekompetenz abbildet. So sollten signifikante Schulformunterschiede sowie höhere Testleistungen bei den Mädchen bestehen.

Hierzu wurden für 566 Schüler/innen aller Schulformen zunächst individuelle Personenparameter (Weighted Likelihood Estimates, WLEs) mit der Software ConQuest erzeugt. Um die erhaltenen Werte besser interpretieren zu können, wurden die WLEs anschließend auf eine *T*-Wert-Skala ($M = 50$, $SD = 10$) transformiert. Die entsprechenden deskriptiven Daten für die Substichproben sind in Tabelle 2 aufgeführt.

Faktoren	**Faktorstufen**	***N***	***M***	***SD***
Geschlecht	männlich	266	52.35	7.55
	weiblich	300	51.59	6.93
Schulform	Hauptschule	205	47.71	5.80
	Realschule	189	52.85	6.31
	Gymnasium	172	55.99	7.08

Anmerkungen: *N* = Stichprobenumfang, *M* = arithmetisches Mittel, *SD* = Standardabweichung

Tab. 2: Deskriptive Daten für Lesekompetenz nach Geschlecht und Schulform

Zur Überprüfung des Einflusses des Geschlechts bzw. der Schulform (und dem Interaktionsterm beider Faktoren) auf die individuelle Lesekompetenzausprägung wurde in einem ersten Schritt ein latentes Regressionsmodell in ConQuest geschätzt. Diese messfehlerkorrigierte Analyse zeigt folgendes Bild: Es findet sich ein signifikanter Einfluss des Faktors Schulform auf die Höhe der Lesekompetenz; eine Geschlechtsspezifik ist dagegen nicht nachweisbar. Die Interaktion der Faktoren Geschlecht und Schulform bleibt ebenfalls ohne statistische Bedeutung (Tabelle 3). Mädchen und Jungen zeigen vergleichbare Leistungen in ihrer Lesekompetenz. Die nach Schulform differenzierten Ergebnisse verdeutlichen die erwarteten Schulformunterschiede.

Da die latente Regressionsanalyse lediglich einen Globaltest zur Signifikanzabschätzung liefert und keinen multiplen Mittelwertvergleich post hoc anbietet, wurde in

	Latente Regression mit ConQuest			Univariate Varianzanalyse mit SPSS		
Faktoren	*B*	*SE*	*T*	*F*	*p*	η^2
Geschlecht	–.083	.109	–.761	2.296	.130	.004
Schulform	.294	.081	3.630	82.027	.000	.227
Geschlecht × Schulform	.084	.052	1.615	0.922	.398	.003

Anmerkungen: abhängige Variable = WLE (transformiert auf *T*-Wert-Skala); *B* = unstandardisierter Regressionskoeffizient, *SE* = Standardfehler, *T* = Wert aus *T*-Verteilung, *F* = Wert aus *F*-Verteilung, *p* = Signifikanz, η^2 = Effektgröße Eta-Quadrat

Tab. 3: Mittelwertunterschiede nach Geschlecht, Schulform und Geschlecht × Schulform in latenter Regressionsanalyse und univariater Varianzanalyse

einem zweiten Schritt eine univariate Varianzanalyse in SPSS berechnet. Hier zeigt sich ein vergleichbares Bild (Tabelle 3) wie bei der latenten Modellierung. Zusätzlich lassen sich die Schulformunterschiede zugunsten höherer Schulformen statistisch absichern. Demnach liegen die Ergebnisse der Gymnasialschüler/innen nahezu eine Standardabweichung über den Leistungen der Hauptschüler/innen.

Die zuvor berichteten Analysen dienen als Ausgangsbasis für weitere Testentwicklungsschritte. Für den dynamischen Test werden jedoch ca. 110 Items benötigt, sodass die jetzige Anzahl an Items zu verdreifachen sein wird. Bereits vorhandene Aufgaben werden überarbeitet und weiter genutzt.

4. Diskussion und Ausblick

Zur experimentellen Überprüfung der Wirksamkeit unterschiedlicher Hilfen und Feedbacks im Rahmen eines computerbasierten Lesekompetenztests wurden bereits Texte und Aufgaben konstruiert, die für zukünftige Projektarbeiten weiterhin genutzt werden sollen. Eine Itemanalyse konnte zeigen, dass die bestehenden Aufgaben einem eindimensionalen Lesekompetenzmodell genügen. DIF-Analysen erbrachten nur bei einem Item eine unzulässige Bevorteilung der Mädchen; in Bezug auf die Schulform wiesen allerdings einige Items DIF auf. Diese Items wurden für nachfolgende Berechnungen zur Überprüfung von Mittelwertunterschieden nicht verwendet. In der Gesamttestleistung treten keine geschlechtsspezifischen Unterschiede auf, was bei sonst verstärkt nachgewiesener Geschlechtsspezifik im sprachlichen Bereich (vgl. Beck/Klieme 2007) im Zuge der Testkonstruktion weiter analysiert werden muss. Erwartungsgemäß gelingt hingegen die Differenzierung der Leseleistung zwischen Schüler/innen aller drei Schulformen. Zusammenfassend kann festgehalten werden, dass der pilotierte Test zu diagnostischen Zwecken herangezogen werden kann, jedoch weiter optimiert werden muss. Zudem kann derzeit nicht der Anspruch auf deutschlandweit gültige Aussagen erfüllt

werden, da die berichteten Ergebnisse aus einer nicht repräsentativen bayerischen Schulpopulation stammen.

Übergeordnetes Ziel der kommenden Projektjahre wird die Etablierung eines dynamischen Lesekompetenztests sein. Dieser Test soll einen breiten Einsatz in unterschiedlichen Fähigkeitsbereichen finden. Geplant ist es, den Test bundesweit in den Klassenstufen 5 bis 8 an Grund-, Haupt-, Realschule und Gymnasium einzusetzen.

Nachdem die Erprobung der entwickelten Textverstehensindikatoren den Grundstein für die weitere Testentwicklung legte, werden die bestehenden Materialien optimiert und für den Einsatz an Fünft- bis Achtklässler/innen hinsichtlich der gestellten Anforderungen in den Randbereichen der Itemschwierigkeiten erweitert. Basierend auf den Ergebnissen des Experiments zur Wirksamkeit von Fördermaßnahmen werden auch für die neuen Texte und Aufgaben Feedbacks und Lern-/Denkhilfen erstellt. Die Validierung einer so erzeugten dynamischen Version des Lesekompetenztests im Vergleich zur herkömmlichen Statustestapplikation (ohne Feedbacks und Hilfen) bildet einen Hauptstrang der zukünftigen Arbeiten. Ein zweiter Hauptstrang in der Testentwicklung ergibt sich aus der Erarbeitung und Evaluierung einer adaptiv dynamischen Variante des gleichen Lesekompetenztests.

Neben der technischen Umsetzung bietet die psychometrische Modellierung der Testdaten zahlreiche Herausforderungen: Die Schwierigkeit bei der Skalierung der Testdaten entsteht durch die mögliche Veränderung der Lesekompetenz mit Hilfe der angebotenen Feedbacks und Hilfen. Die Abbildung dieses Veränderungsprozesses und die Erfassung einer Lernfähigkeit stellen zentrale Herausforderungen beim Einsatz eines dynamischen Tests dar und werden weitere Betätigungsfelder im Zuge der Testentwicklung darstellen (vgl. Dörfler/Golke/Artelt 2009).

Inwiefern sich ein dynamischer Lesekompetenztest als Instrument zur Erfassung von Förderbedarf eignet, ist Gegenstand nachgelagerter Untersuchungen mit dem noch zu validierenden Test. Von Interesse ist hierbei z.B., ob sich Personen identifizieren lassen, die in besonderer Weise von den angebotenen Feedbacks oder Hilfen profitieren (sog. Gainer). Diese Testpersonen sprächen demnach gut auf Fördermaßnahmen im Bereich des Textverstehens an, woraus sich spezifisches Förderpotenzial ableiten ließe. Somit wäre durch die Etablierung des dynamischen Lesekompetenztests ein Schritt zur Vereinbarkeit von Diagnostik und Intervention getan.

Literatur

Artelt, C./Baumert, J./Julius-McElvany, N. (2003): Selbstreguliertes Lernen: Motivation und Strategien in den Ländern der Bundesrepublik Deutschland. In: Baumert, J./Artelt, C./Klieme, E./Neubrand, N./Prenzel, M./Schiefele, U./Schneider, W./Tillmann, K.J./Weiß, M. (Hrsg.): PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland. Opladen: Leske + Budrich.

Bangert-Drowns, R.L./Kulik, C./Kulik, J.A./Morgan, M.T. (1991): The instructional effect of feedback in test-like events. In: Review of Educational Research 61, S. 213–238.

Beck, B./Klieme, E. (Hrsg.): (2007). Sprachliche Kompetenzen. Konzepte und Messung. Weinheim: Beltz.

Coté, N./Goldman, S.R. (1999): Building representations of informational text: Evidence from children's think-aloud protocols. In: van Oostendorp, H./Goldman, S.R. (Hrsg.): The construction of mental representations during reading. Mahwah, NJ: Erlbaum, S. 169–193.
Dörfler, T./Golke, S./Artelt, C. (2009): Dynamic Assessment and its Potential for the Assessment of Reading Competence. In: Studies in Educational Evaluation 35, S. 77–82.
Golke, S./Dörfler, T./Artelt, C. (2009): Effects of Accuracy Feedback during a Text Comprehension Test. In: Educational and Child Psychology 26, S. 30–39.
Graesser, A.C./Singer, M./Trabasso, T. (1997): Constructing inferences during narrative text comprehension. In: Psychological Review 101, S. 371–395.
Guthke, J./Beckmann, J.F./Wiedl, K.H. (2003): Dynamik im dynamischen Testen. In: Psychologische Rundschau 54, S. 225–232.
Hattie, J./Timperley, H. (2007): The power of feedback. In: Review of educational research 77, S. 81–112.
Kintsch, W. (1998): Comprehension: A paradigm for cognition. Cambridge: Cambridge University Press.
Kluger, A.N./DeNisi, A. (1996): Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. In: Psychological Bulletin 119, S. 254–284.
Lewis, C./Reiman, J. (1993): Task-centered user interface design: A practical introduction. Boulder, CO: University of Colorado.
Narciss, S. (2006): Informatives tutorielles Feedback. Münster: Waxmann.
Oakhill, J.V./Cain, K. (2004): The Development of Comprehension Skills. In: Nunes, T./Bryant, P. (Hrsg.): Handbook of Childrens Literacy. Dordrecht: Kluwer Academic Publishers, S. 155–180.
Oakhill, J.V./Garnham, A. (1988): Becoming a skilled reader. Oxford: Blackwell.
Richter, T./Christmann, U. (2002): Lesekompetenz: Prozessebenen und interindividuelle Unterschiede. In: Groeben, N./Hurrelmann, B. (Hrsg.): Lesekompetenz: Bedingungen, Dimensionen, Funktionen. Weinheim: Juventa, S. 25–58.
Shute, V.J. (2008): Focus on Formative Feedback. In: Review of Educational Research 78, S. 153–189.
Souvignier, E./Mokhlesgerami, J. (2006): Using self-regulation as a framework for implementing strategy instruction to foster reading comprehension. In: Learning and Instruction 16, S. 57–71.
Wygotski, L.S. (1964): Denken und Sprechen. Berlin: Akademie.
Willis, G.B. (2004): Cognitive Interviewing: A Tool for Improving Questionnaire Design. Thousand Oaks: Sage Publications.
Wu, M./Adams, R./Wilson, M. (1997): ConQuest: Generalized item response modelling software. Melbourne: Australian Council for Educational Research.

Anschrift des Autors/der Autorinnen

Dr. Tobias Dörfler, Otto-Friedrich-Universität Bamberg, Markusplatz 3, D-96045 Bamberg
E-Mail: tobias.doerfler@uni-bamberg.de

Dipl.-Psych. Stefanie Golke, Otto-Friedrich-Universität Bamberg, Markusplatz 3,
D-96045 Bamberg
E-Mail: stefanie.golke@uni-bamberg.de

Prof. Dr. Cordula Artelt, Otto-Friedrich-Universität Bamberg, Markusplatz 3,
D-96045 Bamberg
E-Mail: cordula.artelt@uni-bamberg.de

Thorsten Roick/Petra Stanat/Oliver Dickhäuser/Volker Frederking/ Christel Meier/Lydia Steinhauer

Strukturelle und kriteriale Validität der literarästhetischen Urteilskompetenz

Projekt literarästhetische Urteilskompetenz[1]

1. Theoretische Modellierung der literarästhetischen Urteilskompetenz

1.1 Literarästhetische Urteilskompetenz als Herausforderung für die empirische Bildungsforschung

Die kompetenztheoretische Modellierung und empirische Validierung literarästhetischer Lehr-Lern-Prozesse gehört zu den schwierigsten und dringlichsten Aufgaben sprachbezogener Bildungsforschung. Die Schwierigkeit erklärt sich aus der besonderen Beschaffenheit des zu erfassenden Gegenstandes. Denn künstlerische Texte sind prinzipiell durch Polyvalenz bzw. Mehrdeutigkeit (vgl. Eco 1962) gekennzeichnet. Die dadurch bedingte „Ambiguität der künstlerischen Botschaft" (ebd., S. 11) stellt die Formulierung eindeutiger Aussagen und damit eine Operationalisierung richtiger Lösungen vor besondere Herausforderungen (vgl. Frederking u.a. 2008). Die Dringlichkeit ergibt sich zum einen daraus, dass es einer theoretisch anschlussfähigen und empirisch tragfähigen Modellierung literarästhetischer Kompetenz bedarf, um Bildungsstandards, Lernstandserhebungen und die kompetenzorientierte Erforschung von Lehr-Lernprozessen im Hinblick auf literarästhetisches Verstehen auf ein wissenschaftlich solides Fundament zu stellen. Zum anderen gilt es zu verhindern, dass mit der ästhetischen Bildung mittelfristig ein Kernbereich des Deutschunterrichts an den Rand des Unterrichtsgeschehens gedrängt wird, nur weil er empirisch vergleichsweise schwierig zu erfassen ist (vgl. Frederking 2008).

Auch empirische Befunde sprechen dafür, dass es sinnvoll ist, literarästhetische Urteilskompetenz genauer zu bestimmen. So weisen Reanalysen der PISA-2000-Daten darauf hin, dass literarästhetische Kompetenz nicht mit einer auf informatorische Verstehensprozesse fokussierenden Lesekompetenz identisch ist, sondern eine eigene Kompetenzdimension darstellt (vgl. Artelt/Schlagmüller 2004). Die meisten Studien (vgl. z.B. Beck/Klieme 2007; Groeben/Hurrelmann 2002), die sich mit der vertieften Untersuchung von Lesekompetenzen befassen, legen den Fokus jedoch auf Leseprozesse im Zusammenhang mit Sach- und Gebrauchstexten, während literarästhetische Aspekte nur in Ansätzen betrachtet werden. Im Unterschied zur PISA-Studie (vgl. Artelt u.a.

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: DI 929/3-1, FRE 2640/1-1, STA 626/4-1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen" (SPP 1293).

2001) wird dabei zwar gelegentlich eine größere Anzahl von literarischen Texten einbezogen (vgl. z.B. Willenberg 2007), eine fundierte Konzeption literarästhetischer (Urteils-)Kompetenz ist bislang jedoch weder entwickelt, noch empirisch erhoben worden (vgl. Frederking 2008).

1.2 Modell der literarästhetischen Urteilskompetenz

Als literaturtheoretisches Fundament für die Modellierung literarästhetischer Urteilskompetenz fungiert die ästhetische Semiotik Umberto Ecos (vgl. z.B. 1962, 1972, 1990, 1992). Denn anders als kognitionspsychologische Ansätze (vgl. z.B. Kintsch 1994; Krommer 2003) erlaubt diese eine hinreichende Erfassung der Komplexität literarästhetischer Texte. Und anders als rezeptionsästhetische, konstruktivistische oder dekonstruktivistische literaturwissenschaftliche Ansätze, die die Leser/innen ins Zentrum des Interesses stellen (vgl. z.B. Jauß 1982; Scheffer 1992; de Man 1993), ermöglicht Ecos stark an der Textintention orientierte Ästhetik eine Lösung für das im Zusammenhang mit literarästhetischen Verstehensprozessen existierende Grundproblem: Wie lassen sich eindeutige und damit operationalisierbare Aussagen über einen literarischen Text formulieren, obschon dieser mehrdeutig ist? Wegweisend ist in dieser Hinsicht Ecos Unterscheidung von drei Konstituenten im literarischen Produktions- und Rezeptionsprozess – Autor/in, Text und Leser/in – sowie von drei damit korrespondierenden Intentionsbereichen – *intentio auctoris*, *intentio operis* und *intentio lectoris*. Sowohl die Intention der Autorin/des Autors, die *intentio auctoris*, als auch die Intention der Leserin/des Lesers, die *intentio lectoris*, lassen sich nur in eingeschränkter Form in ihrer Beziehung zum Text objektivieren und erfassen. Eco spricht von „der unergründlichen Intention des Autors und der anfechtbaren Intention des Lesers“ (1992, S. 87). Demgegenüber bietet die als Ebene der *transparenten Textintention* (ebd., S. 87) verstandene *intentio operis* verlässliche Ansatzpunkte für die Unterscheidung von richtigen und falschen Aussagen. Entscheidendes Kriterium ist dabei, inwieweit sich Deutungen am Text belegen lassen (Eco 1990, S. 51).

Auf der Ebene der *intentio operis* lassen sich nun in Anknüpfung an Eco zwei Dimensionen literarästhetischer Urteilskompetenz (LUK) postulieren: die Fähigkeit zu semantischem und die Fähigkeit zu idiolektalem literarästhetischem Urteilen (vgl. Frederking u.a. 2008). Auf der Grundlage anderer, über Eco hinausgehender literaturwissenschaftlicher Bezugstheorien lässt sich noch eine dritte Teildimension extrapolieren: die Fähigkeit zu kontextuellem literarästhetischem Urteilen (vgl. ebd.). Begründungszusammenhänge für alle drei Teildimensionen seien nachfolgend jeweils im Grundansatz erläutert.

Semantische LUK bezieht sich auf die Fähigkeit zur Erschließung zentraler Textinhalte und zur Generierung einer kohärenten Textdeutung. Dies stellt an die Leser/innen besondere Anforderungen. Denn einerseits ist jeder literarische Text prinzipiell „… offen … für eine virtuell unendliche Reihe möglicher Lesarten, deren jede das Werk gemäß einer persönlichen Perspektive, Geschmacksrichtung, Ausführung neu belebt“ (Eco 1962, S. 57). Andererseits sind die Deutungsspielräume des rezipierenden Sub-

jekts durch den in der *intentio operis* angelegten und trotz „semantischer Pluralität" (ebd., S. 87) bestimmbaren „kohärenten Textsinn" (Eco 1990, S. 43) eingeschränkt. Dieser Sachverhalt bildet den Ansatzpunkt für die kompetenztheoretische Modellierung und empirische Erhebung semantischen literarästhetischen Urteilens.

Unter *idiolektaler LUK* wird die Fähigkeit verstanden, den „ästhetischen Idiolekt" (Eco 1972, S. 151) zu erfassen, d.h. die Strukturmerkmale, aufgrund derer ein Text „… diese (oder andere) semantische Interpretationen hervorbringen kann" (Eco 1990, S. 43). Das Verstehen eines literarischen Textes hat nach Eco sowohl die Identifizierung und Dechiffrierung formaler Auffälligkeiten als auch das Durchdringen des Zusammenhangs zwischen Textinhalt und Textform zur Voraussetzung. Die Fähigkeit, die sprachlichen Besonderheiten eines literarischen Textes wahrzunehmen und ihre Bedeutung für die Initiierung semantischer Verstehensprozesse zu erkennen, wird als Teildimension literarästhetischen Urteilens verstanden.

Die *kontextuelle LUK* lässt sich als Fähigkeit verstehen, auch textextern präsentierte historische, entstehungs-, motiv- oder mentalitätsgeschichtliche sowie epoche- oder gattungsspezifische Zusatzinformationen für die Interpretation eines literarischen Textes fruchtbar machen zu können. Auf die Bedeutung derartiger inter- bzw. transtextueller Bezüge ist aus unterschiedlichen literaturwissenschaftlichen und -theoretischen Perspektiven hingewiesen worden (vgl. z.B. zur Fiktionalitätsproblematik Currie 1990, zur Autorschaft Jannidis u.a. 1999, zu Paratexten Genette 1987, zu Gattungsfragen Voßkamp 1992 oder zum Epochenproblem Rosenberg 1992).

Dass die Verbindung textinterner und textexterner Aspekte im literarästhetischen Verstehensprozess tatsächlich eine eigene Teilkompetenz darstellt, die andere Fähigkeiten als semantisches oder idiolektales literarästhetisches Urteilen voraussetzt, ist auf der Grundlage der angeführten literaturtheoretischen Positionen eine naheliegende Annahme. Plausibel ist aber auch, dass es sich bei kontextuellen Urteilsprozessen lediglich um komplexere Ausprägungen der beiden anderen Teildimensionen handelt, weil alle Kontextinformationen entweder auf der semantischen oder auf der idiolektalen Ebene angesiedelt sind. Damit lässt sich literaturtheoretisch sowohl ein zweidimensionales Modell literarästhetischer Urteilskompetenz begründen, das nur aus semantischer und idiolektaler Teildimension besteht, als auch ein dreidimensionales, das zusätzlich noch die Fähigkeit zu kontextuellen literarästhetischen Urteilen umfasst.

1.3 Fragestellungen

Im Zusammenhang mit dem Modell literarästhetischer Urteilskompetenz werden zwei Fragestellungen aufgeworfen, die im Wesentlichen die Validität des Konstrukts betreffen. Die erste Fragestellung zielt auf die Überprüfung des auf der Grundlage der semiotischen Ästhetik Ecos sowie anderer literaturwissenschaftlicher Theorien entwickelten zwei- bzw. dreidimensionalen Modells der LUK. Empirisch zu untersuchen ist, ob sich LUK tatsächlich im Sinne der drei postulierten Teildimensionen (semantisch, idiolektal und kontextuell) modellieren lässt oder ob das zweidimensionale Modell eine größere Passung zu den

ermittelten Daten aufweist. In jedem Fall sollten sowohl das zwei- als auch das dreidimensionale LUK-Modell einer eindimensionalen Betrachtung des Konstrukts überlegen sein.

Die zweite Fragestellung zielt auf die Abgrenzbarkeit der LUK von allgemeiner Lesekompetenz sowie von schulischen Leistungen im Sinne der Prüfung konvergenter und diskriminanter Validität. Da die LUK und die Lesekompetenz gleichermaßen über kontinuierliche Texte erfasst werden, ist zwar mit substanziellen Zusammenhängen zwischen LUK und Lesekompetenz zu rechnen, beide Konstrukte sollten aber empirisch ausreichend voneinander abgrenzbar sein. Hinsichtlich der Zusammenhänge mit Schulnoten ist zu erwarten, dass sowohl Lesekompetenz als auch LUK substanziell mit Schulleistungen korrelieren, LUK aber Zusammenhänge zu Deutschleistungen aufweist, die über die Korrelation mit basaler Lesekompetenz hinausgehen. Zur Betrachtung der diskriminanten Validität werden die Schulnoten in den Fächern Mathematik, Englisch sowie Kunst herangezogen.

2. Methode der Untersuchung

2.1 Variablen

Auf der Grundlage der theoretischen Konstruktmodellierung und von Annahmen über Anforderungsniveaus wurden zunächst Aufgaben-Units zu 21 literarischen Texten konstruiert, von denen 16 umfassend pilotiert wurden. Im Rahmen der Pilotierung wurden neun Aufgaben-Units mit insgesamt 62 Testitems ausgewählt (zu Details der Pilotierungsstudie und zu Item-Beispielen vgl. Frederking u.a. 2009). 44 der Testitems ließen sich a priori einer der zwei theoretisch postulierten Kompetenzdimensionen (semantisch, idiolektal) zuordnen, die restlichen 18 Items bezogen sich auf semantisch-kontextuelle oder idiolektal-kontextuelle Urteile. Als Aufgabenformate wurden sowohl offene (47%) als auch geschlossene Formate (53%) verwendet. Die LUK-Aufgaben wurden für ein Multi-Matrix-Erhebungsdesign in insgesamt neun Booklets zu je vier Units organisiert, wobei die Position der Units, die Verschränkbarkeit der Booklets untereinander sowie die Länge und Gattungszugehörigkeit der Stammtexte berücksichtigt wurde.

Die allgemeine Lesekompetenz der Schüler/innen wurde mit Aufgaben-Units eines Leseverständnistests erfasst (vgl. Institut für Qualitätsentwicklung 2007),[2] dem das Lesekompetenzmodell der PISA-Studien zugrunde lag (vgl. dazu Artelt u.a. 2001). Aus dem Leseverständnistest wurden vier Aufgaben-Units mit kontinuierlichen Texten zu Sachthemen und insgesamt 18 Items ($\alpha = .75$) ausgewählt und in zwei Pseudoparallelformen eingesetzt. Dem Vorgehen bei Artelt u.a. (2001) folgend wurde über alle 18 Items hinweg ein Gesamtwert für die Lesekompetenz gebildet. Darüber hinaus wurden die Schüler/innen gebeten, für die Fächer Deutsch, Englisch, Mathematik und Kunst ihre letzten Zeugnisnoten anzugeben.

---

2 Wir bedanken uns herzlich bei Cordula Artelt und ihrer Arbeitsgruppe (Universität Bamberg) für die Bereitstellung des von ihnen entwickelten Lesekompetenztests.

2.2 Stichprobe

Bei der Schulauswahl wurde darauf geachtet, dass Schulen in städtischen und ländlichen Gebieten und Jugendliche aus unterschiedlichen Herkunftsmilieus in die Stichprobe einbezogen wurden. Insgesamt wurde eine Stichprobe von $N = 1187$ Schüler/innen (49% Mädchen, Alter: $M = 15.31$ Jahre, $SD = 0.73$) der neunten Jahrgangsstufe bayerischer Hauptschulen, Realschulen und Gymnasien untersucht, die im Rahmen einer 90-minütigen Sitzung die LUK-Testaufgaben und in einer weiteren 45-minütigen Sitzung die Lesekompetenzaufgaben bearbeiteten. Die Erhebungen fanden an zwei getrennten Testtagen statt und wurden von geschulten Testleiter/innen durchgeführt.

3. Ergebnisse der Untersuchung

3.1 Strukturprüfung

Zur Prüfung der strukturellen Validität der LUK liegen verwertbare Daten von $N = 1052$ Schüler/innen vor. Dem Multi-Matrix-Erhebungsdesign entsprechend bearbeiteten die Jugendlichen dabei jeweils nur Teilmengen von Testitems. Um die Ergebnisse der Schüler/innen dennoch miteinander vergleichen zu können, wurde auf der Grundlage der *Item-Response-Theorie* eine gemeinsame Skalierung der LUK-Aufgaben vorgenommen (vgl. z.B. Yen/Fitzpatrick 2006). Die Überprüfung der theoretisch angenommenen Struktur literarästhetischer Urteilskompetenz erfordert zudem den Einsatz mehrdimensionaler Modelle, die die Beziehung der Teildimensionen untereinander modellieren können (Basis bildet das *Multidimensional Random Coefficient Multinominal Logit Model*, vgl. Adams/Wilson/Wang 1997). Die Analysen zur Dimensionalität des Konstrukts wurden mit der Software ConQuest (vgl. Wu u.a. 2007) vorgenommen. Da für 44% der LUK-Testaufgaben ein abgestuftes *Scoring* vorgesehen war, erfolgte die Modellierung dieser Aufgaben über ein *Partial-credit-Modell*.

In Bezug auf die empirische Modellierbarkeit der postulierten dreidimensionalen Struktur der LUK weisen die Ergebnisse darauf hin, dass das dreidimensionale Modell (*Deviance* = 39028.14, *Parameter* = 95) einen statistisch signifikant besseren Fit an die Daten aufweist als ein einfaches Modell (*Deviance* = 39050.51, *Parameter* = 90), in dem LUK als eindimensionales Konstrukt abgebildet wird ($\chi^2(5) = 22.38$, $p < .01$). Eine Prüfung des zweidimensionalen Modells, bei dem die kontextuellen Items (so wie a priori postuliert) entweder dem Faktor semantische LUK oder idiolektale LUK zugeordnet wurden, zeigt, dass sich ein solches zweidimensionales Modell (*Deviance* = 39008.65, *Parameter* = 92) nicht nur besser an die Daten anpassen lässt als das eindimensionale Basismodell ($\chi^2(2) = 41.87$, $p < .01$), sondern auch dem dreidimensionalen LUK-Modell überlegen ist (geringere *Deviance*). Hinsichtlich der auf der Theorie-Ebene nicht abschließend zu klärenden Frage, ob LUK sich eher zweidimensional oder dreidimensional konzeptualisieren lässt, sprechen die Daten daher eher für ein Modell mit zwei Teil-

	Gesamtwert LUK	**semantische LUK[a]**	**idiolektale LUK[a]**
MW (SD)	−0.36 (1.03)	−0.32 (1.16)	−0.43 (1.11)
Gesamtwert LUK	–	.94	.87
semantisch-(kontextuell)e LUK	–	–	.68 [.92]

Anmerkungen: a) Teildimensionen des zweidimensionalen LUK-Modells; Mittelwerte, Streuungen und manifeste Korrelationswerte der Personenparameter (als Warm's weighted likelihood Schätzwerte, WLE); latente Korrelation der Teildimensionen in Eckklammern.

Tab. 1: Deskriptiva und Interkorrelationen der LUK-Testwerte

dimensionen: einem Faktor der semantisch(-kontextuell)en LUK und einem Faktor der idiolektal(-kontextuell)en LUK.

Wie Tabelle 1 zeigt, finden sich zwischen diesen beiden Teildimensionen des zweidimensionalen LUK-Modells moderate bis hohe Korrelationswerte auf manifester und latenter Ebene, die es nahe legen, auch einen Gesamtwert (Personenparameter des eindimensionalen Modells) für die literarästhetische Urteilskompetenz (LUK) zu bilden. Erwartungsgemäß finden sich hohe korrelative Zusammenhänge zwischen dem LUK-Gesamtwert und den beiden Teildimensionen.

3.2 Prüfung der konvergenten und diskriminanten Validität

Die konvergente und diskriminante Validität der LUK wurde anhand von korrelativen Zusammenhängen mit Lesekompetenz und selbst berichteten Zeugnisnoten überprüft. Durch eine wechselseitige Auspartialisierung soll bestimmt werden, in welcher Höhe Zusammenhänge mit der jeweiligen Zeugnisnote unter Kontrolle sämtlicher anderer Variablen erhalten bleiben. Tabelle 2 fasst die Ergebnisse der Korrelationsberechnungen für die Gesamtgruppe zusammen.

Wie die Ergebnisse zeigen, liegen substanzielle Korrelationen in mittlerer Höhe zwischen der Lesekompetenz und der LUK vor. Der Anteil der geteilten Varianz (maximal 36%, minimal 29%) spricht jedoch gleichzeitig dafür, dass Lesekompetenz und LUK empirisch trennbar sind und in Teilen distinkte Konstrukte repräsentieren.[3] Sowohl Lesekompetenz als auch LUK korrelieren in erwarteter Höhe mit den von den Schüler/innen berichteten Zeugnisnoten. Der LUK-Gesamtwert korreliert – ebenfalls wie erwartet – statistisch signifikant höher mit der Deutschnote als mit den Noten in den anderen

3 Der Anteil der gemeinsamen Varianz auf latenter Ebene (bei dem die Werte um die Reliabilität korrigiert werden) liegt zwischen 49% und 66%. Zusätzliche konfirmatorische Analysen, auf deren Darstellung aus Platzgründen hier verzichtet werden muss, zeigen, dass auf latenter Ebene zwischen Lesekompetenz und LUK getrennt werden kann und dass beide Faktoren latent zu $r = .78$ korrelieren.

	Lese-kompetenz		Gesamtwert LUK		semantische LUK[a]		idiolektale LUK[a]	
	r	r_{par}	*r*	r_{par}	*r*	r_{par}	*r*	r_{par}
Lesekompetenz	–	–	.60*	.57*	.57*	.54*	.54*	.50*
Deutschnote	−.24*	−.04	−.36*	−.17*	−.34*	−.18*	−.30*	−.09
Englischnote	−.16*	.01	−.27*	−.07	−.23*	−.02	−.28*	−.12*
Mathematiknote	−.10*	.05	−.22*	−.09	−.20*	−.08	−.20*	−.07
Kunstnote	−.15*	−.06	−.17*	−.02	−.16*	−.01	−.15*	−.02

Anmerkungen: * $p < .01$; $N = 744$; a) Teildimensionen des zweidimensionalen LUK-Modells; r_{par}: Partialkorrelation zwischen zwei Variablen unter Kontrolle aller anderen Variablen; das negative Vorzeichen der Korrelationen ist auf die Notenskala zurückzuführen, die besseren Leistungen niedrigere Werte zuweist.

Tab. 2: Korrelationen 0-ter Ordnung und Partialkorrelationen zwischen LUK, Lesekompetenz und Zeugnisnoten

Fächern (alle *p*-Werte < .01). Im Fach Deutsch, aber auch in den Fächern Englisch und Mathematik, finden sich signifikant höhere Korrelationen zwischen den Noten und der Leistung in den LUK-Aufgaben als zwischen den Noten und den Leistungen im Lesetest. Die Partialkorrelationen schließlich zeigen, dass substanzielle Zusammenhänge zwischen LUK und Deutschnote selbst unter Kontrolle der Lesekompetenz und der anderen Zensuren erhalten bleiben, nicht jedoch umgekehrt. Es ist also gemeinsame Varianz zwischen LUK und Schulnoten in Deutsch zu beobachten, die nicht durch Unterschiede in der Lesekompetenz der Schüler/innen oder deren sonstiges Leistungsniveau erklärt werden kann.

Für die beiden Teildimensionen des zweidimensionalen LUK-Modells zeigt sich ein ähnliches Bild. Nur numerisch sind die Zusammenhänge zwischen Lesekompetenz und semantischer LUK (Partialkorrelationen) etwas enger als zwischen Lesekompetenz und idiolektaler LUK ($t(741) = 1.22$, $p = .22$). Statistisch signifikant unterschiedliche Zusammenhänge zeigen sich zwischen den Zeugnisnoten in den Fächern Deutsch bzw. Englisch und den beiden LUK-Teildimensionen. Hier ergibt sich, dass der Zusammenhang zwischen semantischer LUK und Deutschnote enger ausfällt als der zwischen idiolektaler LUK und Deutschnote ($t(741) = 2.53$, $p = .01$), während die idiolektale LUK relativ betrachtet höher mit der Englischnote korreliert ($t(741) = 2.72$, $p = .01$).

4. Diskussion und Ausblick

In diesem Beitrag wurde der Frage nachgegangen, ob sich ein dreidimensionales Modell der literarästhetischen Urteilskompetenz (LUK), für das literaturtheoretisch gute Gründe sprechen, empirisch bestätigen lässt oder ob sich ein theoretisch ebenfalls plausibles zweidimensionales LUK-Modell angesichts der ermittelten Daten als überlegen erweist. Weiterhin wurde der Versuch unternommen, LUK gegen allgemeine Lesekom-

petenz abzugrenzen. In Bezug auf die strukturelle Validität lässt sich festhalten, dass das dreidimensionale Modell aus semantischer, idiolektaler und kontextueller LUK empirisch besser mit den Daten vereinbar ist als ein eindimensionales Modell. Eine noch bessere Anpassung an die Daten zeigt jedoch das zweidimensionale Modell, welches nur zwischen semantischer und idiolektaler Teildimension differenziert. Eine Erklärung könnte dieses Ergebnis in dem Sachverhalt finden, dass sich kontextuelle LUK-Aufgaben immer entweder auf Inhalts- oder aber auf Formaspekte beziehen. Sie lassen sich also als semantische und idiolektale LUK-Aufgaben verstehen, die um zusätzliche Informationen angereichert werden. Die semantischen und idiolektalen Prägungen wiegen dabei offenbar stärker als die Tatsache, dass zusätzliche textexterne Informationen auf einen Text bezogen werden müssen. In weiteren Analysen ist zu prüfen, inwieweit sich dieses zweidimensionale LUK-Modell auch gegenüber anderen, inhaltlich sinnvollen Modellen bewährt.

Hinsichtlich der kriterialen Validität weisen die Ergebnisse darauf hin, dass sich LUK von allgemeiner Lesekompetenz abgrenzen lässt und Zusammenhänge zu Schulleistungen aufweist, die über eine basale Lesekompetenz hinausgehen. Im Fach Deutsch, aber auch in anderen Fächern finden sich signifikant höhere Korrelationen zwischen den Zeugnisnoten und der LUK als zwischen den Noten und der Lesekompetenz. Die Partialkorrelationen zeigen schließlich, dass substanzielle Zusammenhänge zwischen LUK und Zeugnisnoten selbst unter Kontrolle der Lesekompetenz erhalten bleiben, nicht jedoch umgekehrt.

Betrachtet man die beiden Teildimensionen der LUK getrennt, so ist denkbar, dass die Fähigkeit zum Verständnis der formalen Spezifika eines Textes (idiolektale LUK) vermittelt über grammatische Kompetenz mit der Fremdsprachen-Kompetenz korreliert. Ein vertieftes Verständnis für syntaktische Strukturen ist sowohl für das Erfassen formaler Besonderheiten in muttersprachlichen literarischen Texten als auch für die gute Beherrschung einer von der Muttersprache divergierenden fremdsprachlichen Syntax nötig. Dies könnte eine Erklärung für die Tatsache bieten, dass idiolektale LUK enger mit der Leistung im Fach Englisch in Beziehung steht als die semantische LUK, denn semantische LUK geht weit über die primär lexikalische Verstehensebene hinaus, auf der sich der Englischunterricht 15-Jähriger normalerweise bewegt.

Zusammengenommen spricht das Befundmuster für die diskriminante Validität des LUK-Tests gegenüber einem Test zur basalen Lesekompetenz. In Bezug auf die Abgrenzung zur Lesekompetenz ist allerdings zu berücksichtigen, dass die hier eingesetzten Teile des Lesekompetenztests nur kontinuierliche Sachtexte umfassten. Weitere Studien müssen zeigen, inwieweit sich LUK auch dann von allgemeiner Lesekompetenz abgrenzen lässt, wenn ihre Operationalisierung auch literarische Texte beinhaltet. Im Rahmen von Subgruppenanalysen gilt es, weitere Belege für die differenzielle Validität der LUK-Teildimensionen zu finden.

Insgesamt können die dargestellten Befunde zur strukturellen und kriterialen Validität als erster Hinweis darauf gewertet werden, dass eine modelltheoretische und empirische Fundierung der literarästhetischen Urteilskompetenz gelungen ist. Eine valide und reliable Erfassung des Konstrukts bildet die Voraussetzung für weitere Forschungsbe-

mühungen in diesem Bereich. Das vorliegende Instrument bietet z.B. Möglichkeiten zu prüfen, durch welche Unterrichtsinhalte und -methoden literarästhetische Urteilskompetenz oder ihre Teilkomponenten bedingt sind und gefördert werden können.

Literatur

Adams, R.J./Wilson, M./Wang, W.-C. (1997): The multidimensional random coefficients multinomial logit model. In: Applied Psychological Measurement 21, S. 1–23.

Artelt, C./Schlagmüller, M. (2004): Der Umgang mit literarischen Texten als Teilkompetenz im Lesen? Dimensionsanalysen und Ländervergleiche. In: Schiefele, U./Artelt, C./Schneider, W./Stanat, P. (Hrsg.): Struktur, Entwicklung und Förderung von Lesekompetenz. Vertiefende Analysen im Rahmen von PISA 2000. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 169–196.

Artelt, C./Stanat, P./Schneider, W./Schiefele, U. (2001): Lesekompetenz: Testkonzeption und Ergebnisse. In: Baumert, J./Klieme, E./Neubrand, M./Prenzel, M./Schiefele, U./Schneider, W./Stanat, P./Tillmann, K.-J./Weiß, M. (Hrsg.): PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske + Budrich, S. 69–137.

Beck, B./Klieme, E. (2007): Sprachliche Kompetenzen: Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International). Weinheim: Beltz.

Currie, G. (1990): The nature of fiction. New York: Cambridge University Press.

de Man, P. (1993): Die Ideologie des Ästhetischen. Frankfurt a.M.: Suhrkamp.

Eco, U. (1962): Das offene Kunstwerk. Frankfurt a.M.: Suhrkamp.

Eco, U. (1972): Einführung in die Semiotik. München: Fink.

Eco, U. (1990): Die Grenzen der Interpretation. München: dtv.

Eco, U. (1992): Zwischen Autor und Text. Interpretation und Überinterpretation. München: dtv.

Frederking, V. (2008): Literarische bzw. (literar)ästhetische Kompetenz. Möglichkeiten und Probleme der empirischen Erhebung eines Kernbereichs des Deutschunterrichts. In: Frederking, V. (Hrsg.): Schwer messbare Kompetenzen. Baltmannsweiler: Schneider, S. 36–64.

Frederking, V./Meier, C./Roick, T./Steinhauer, L./Stanat, P./Dickhäuser, O. (2009): Literarästhetische Urteilskompetenz erfassen. In: Bertschi-Kaufmann, A./Rosebrock, C. (Hrsg.): Literalität. Bildungsaufgabe und Forschungsfeld. Weinheim: Beltz, S. 165–180.

Frederking, V./Meier, C./Stanat, P./Dickhäuser, O. (2008): Ein Modell literarästhetischer Urteilskompetenz. In: Didaktik Deutsch 25, S. 11–31.

Genette, G. (1987): Paratexte. Das Buch vom Beiwerk des Buches. Frankfurt a.M.: Suhrkamp.

Groeben, N./Hurrelmann, B. (2002): Lesekompetenz: Bedingungen, Dimensionen, Funktionen. Weinheim: Juventa.

Institut für Qualitätsentwicklung (2007): Lese(verständnis)test 7 – Hessen. Wiesbaden: Institut für Qualitätsentwicklung.

Jannidis, F./Lauer, G./Martinez, M./Winko S. (1999): Rückkehr des Autors. Tübingen: Niemeyer.

Jauß, H.R. (1982): Ästhetische Erfahrung und literarische Hermeneutik. Frankfurt a.M.: Suhrkamp.

Kintsch, W. (1994): Kognitionspsychologische Modelle des Textverstehens: Literarische Texte. In: Reusser, K./Reusser-Weyeneth, M. (Hrsg.): Verstehen. Psychologischer Prozeß und didaktische Aufgabe. Bern: Huber, S. 39–54.

Krommer, A. (2003): Fiktionen lesen. Ein philosophisch-didaktisches Plädoyer für eine ontologiefreie Theorie der Fiktionalität. In: Frederking, V. (Hrsg.): Lesen und Symbolverstehen in medialen Kontexten. Jahrbuch Medien im Deutschunterricht 2003, Band 2. München: KoPäd, S. 83–99.

Rosenberg, R. (1992): Epochen. In: Brackert, H./Stückrath, J. (Hrsg.): Literaturwissenschaft. Ein Grundkurs. Reinbek: Rowohlt, S. 269–280.
Scheffer, B. (1992): Interpretation und Lebensroman. Zu einer konstruktivistischen Literaturtheorie. Frankfurt a.M.: Suhrkamp.
Voßkamp, W. (1992): Gattungen. In: Brackert, H./Stückrath, J. (Hrsg.): Literaturwissenschaft. Ein Grundkurs. Reinbek: Rowohlt, S. 253–269.
Willenberg, H. (2007): Lesen. In: Beck, B./Klieme, E. (Hrsg.): Sprachliche Kompetenzen: Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International). Weinheim: Beltz, S. 107–117.
Wu, M.L./Adams, R.J./Wilson, M.R./Haldane, S.A. (2007): Acer ConQuest version 2.0: Generalised item response modelling software. Camberwell, Victoria: ACER Press.
Yen, W.M./Fitzpatrick, A.R. (^{4}2006): Item Response Theory. In: Brennan, R.L. (Hrsg.): Educational Measurement. Westport: Praeger Publishers, S. 111–153.

Anschrift der Autor/innen

Dr. Thorsten Roick, Freie Universität Berlin, Arbeitsbereich Empirische Bildungsforschung, Habelschwerdter Allee 45, D-14195 Berlin
E-Mail: thorsten.roick@fu-berlin.de

Prof. Dr. Petra Stanat, Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Unter den Linden 6, D-10099 Berlin
E-Mail: IQBoffice@IQB.hu-berlin.de

Prof. Dr. Oliver Dickhäuser, Universität Mannheim, Lehrstuhl für Pädagogische Psychologie, D-68131 Mannheim
E-Mail: oliver.dickhaeuser@uni-mannheim.de

Prof. Dr. Volker Frederking, Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl für Didaktik der deutschen Sprache und Literatur, Regensburger Straße 160, D-90478 Nürnberg
E-Mail: dr.volker.frederking@t-online.de

Dr. Christel Meier, Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl für Didaktik der deutschen Sprache und Literatur, Regensburger Straße 160, D-90478 Nürnberg
E-Mail: christel.meier@gmx.net

Lydia Steinhauer, c/o Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl für Didaktik der deutschen Sprache und Literatur, Regensburger Straße 160, D-90478 Nürnberg
E-Mail: lydia.steinhauer@googlemail.com

Hans Anand Pant/Simon P. Tiffin-Richards/Olaf Köller

Standard-Setting für Kompetenztests im Large-Scale-Assessment

Projekt Standardsetting[1]

1. Einleitung

Die 2003 von der Kultusministerkonferenz (vgl. KMK 2003) beschlossenen Bildungsstandards gelten verbindlich in allen Bundesländern und sollen den Aufbau eines auf Leistungsmessungen basierenden Systems der Rechenschaftslegung (Accountability) auf der Ebene der Länder ermöglichen. Das Erreichen der Bildungsstandards für die Fächer Englisch, Französisch und Deutsch wurde 2009 das erste Mal durch das Institut zur Qualitätsentwicklung im Bildungswesen (IQB) bundesweit überprüft. Im Large-Scale-Assessment gilt es als essentiell, dass standardbezogene Rückmeldeformate gegenüber verschiedenen gesellschaftlichen Akteuren einfach zu kommunizieren sind. Zu diesem Zweck werden an Stelle von Ergebnisdarstellungen, die sich auf Testrohwerte beziehen (z.B. Mittelwerte und Streuungsmaße) Rückmeldeformate präferiert, die die Verteilung von Schülerleistungen auf kategorial gestuften Kompetenzskalen abbilden (Beispiel: „Das Kompetenzniveau B1 einer selbständigen Sprachverwendung in einer Fremdsprache wird zum Zeitpunkt des Mittleren Schulabschlusses von 60% der Schüler/innen erreicht“).

Die Setzung von Schwellenwerten (Cut-Scores), durch die benachbarte Kategorien auf einer kontinuierlichen Testwertskala abgegrenzt werden, stellt daher ein wichtiges Transformationsmoment zwischen fachdidaktisch und psychometrisch fundierter Kompetenzmessung einerseits und politischer und administrativer Verwertbarkeit andererseits dar. Das prozedurale Vorgehen bei der Festlegung von Cut-Scores auf einer kontinuierlichen Leistungstestskala wird als *Standard-Setting* bezeichnet (vgl. ausführlich Cizek/Bunch 2007).

2. Konzepte und Verfahrensvarianten des Standard-Setting

Das vorliegende Projekt geht – am Beispiel der Kompetenzstufenmodelle für die rezeptiven Kompetenzen Leseverständnis und Hörverständnis in Englisch als erster Fremdsprache – der Frage nach, welche Standard-Setting-Varianten *valide* Cut-Scores bzw. sich daraus ergebende Kompetenzniveaueinteilungen generieren. In den folgenden Abschnitten werden die gängigsten Verfahren zum Standard-Setting, das zugrunde liegende Validitätskonzept sowie die Ziele des vorliegenden Projektes erläutert.

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: PA 1532/2-2) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

2.1 Methoden des Standard-Settings

Idealtypisch wird in einem Standard-Setting Verfahren ein Panel aus Expert/innen konstituiert, das in einem iterativen Verfahren aus Einzelurteilen und Gruppendiskussionen zur Festlegung von Cut-Scores kommt. Dabei werden den Panelteilnehmer/innen je nach angewandter Methode unterschiedliche Information über die empirischen Itemschwierigkeiten und die Folgen präsentiert, die ihre Cut-Score-Setzungen für die „reale" Verteilung der Schülerschaft auf die dadurch entstandenen Kompetenzstufen haben. In der Literatur werden die zahlreichen Verfahrensvarianten in testzentrierte vs. personenzentrierte Methoden klassifiziert.

Bei den *testzentrierten* Verfahren steht die Beurteilung der Testaufgaben bzw. Items durch die Panelteilnehmer/innen im Mittelpunkt. In den US-amerikanischen Large-Scale-Assessments der letzten zehn Jahre wurden zwei testzentrierte Standard-Setting-Verfahren am häufigsten angewendet: Varianten der Angoff-Methode (vgl. Angoff 1971) und die Bookmark-Methode (vgl. Mitzel u.a. 2001).

Im *Angoff-Verfahren* werden die Expert/innen aufgefordert, sich eine hypothetische Person vorzustellen, die an der Grenze zwischen zwei benachbarten Kompetenzstufen steht. Zu jedem Testitem des Kompetenztests ist dann von jedem Panelmitglied die Wahrscheinlichkeit anzugeben, mit der diese vorgestellte Person das Item löst. Im *modifizierten Angoff-Verfahren* wird zu jedem Testitem entschieden, ob die grenzkompetente Person das Item lösen kann oder nicht. Diese Ja/Nein-Einschätzungen werden mit 1/0 kodiert. Die Wahrscheinlichkeitsratings werden für beide Varianten des Angoff-Verfahrens pro Panelmitglied und über alle Mitglieder aggregiert, um den Cut-Score zu ermitteln.

Bei der *Bookmark-Methode* wird den Beurteiler/innen ein „Buch" vorgegeben, das alle Items aufsteigend nach ihrer empirischen Schwierigkeit geordnet enthält. Die Aufgabe der Panelist/innen ist es, im wiederholten Abgleich mit den Kompetenzstufendeskriptoren an denjenigen Stellen im Item-Buch eine Markierung zu setzen, an denen ein/e vorgestellte/r, für dieses Kompetenzniveau gerade kompetente/r Schüler/in mit einer spezifizierten Antwortwahrscheinlichkeit *(Response Probability [RP]*) das Item lösen kann. Die Urteile werden aggregiert und auf der Fähigkeitsskala lokalisiert, die sich aus der Rasch-Skalierung ergibt. Hierbei wird die a priori bestimmte Response Probability zugrunde gelegt, die oft mit 2/3 angesetzt wird (bei PISA: $RP = .62$, vgl. OECD 2009).[2]

Im Unterschied zu den testzentrierten Methoden verwenden *personenzentrierte* Standard-Setting Verfahren wie die *Contrasting-groups-Methode* (vgl. van Nijlen/Janssen 2008) Urteile der Panelmitglieder über *reale* Schüler/innen bzw. deren Leistungen. Sie klassifizieren die Lernenden anhand der Kompetenzstufendeskriptoren direkt auf den Kompetenzniveaus. Diese Verfahren eigenen sich vor allem dann, wenn die Beurteiler/innen die zu Beurteilenden bzw. deren Leistungen gut kennen (z.B. Lehrkräfte). Im Kontext von Large-Scale-Assessments dienen personenzentrierte Verfahren in erster Linie zur externen Validierung der testzentrierten Cut-Score-Entscheidungen.

2 Technisch wird dies erreicht, indem man zum ermittelten Schwierigkeitsparameter (Logit) eines Items eine Verschiebungskonstante von ln(RP/1-RP) hinzuaddiert.

Synopsen zum empirischen Bewährungsstand von Standard-Setting-Methoden (vgl. Hurtz/Auerbach 2003; Karantonis/Sireci 2006) zeigen, dass sich bisher keine der Verfahrensansätze als allgemein akzeptiert etablieren konnte. Derzeit existieren nur wenige Studien, die einen Verfahrensvergleich unter Verwendung desselben Testinstruments vornehmen (vgl. z.B. Buckendahl u.a. 2002; Green/Trimble/Lewis 2003; Yin/Schulz 2005). Diese Studien kommen insgesamt zu inkonsistenten Empfehlungen mit einer Tendenz zugunsten der Bookmark-Methode.

Zunächst soll im folgenden Abschnitt das zugrundegelegte Validitätskonzept kurz erläutert werden.

2.2 Validitätsaspekte bei Standard-Setting-Verfahren

Das Setzen von Cut-Scores durch Expertenurteile stellt *per se* einen bewertenden Vorgang dar. In der Standard-Setting-Literatur wird daher in der Regel nicht von einem *True-Cut-Score*-Konzept ausgegangen (vgl. Kane 2001; für eine Ausnahme siehe Reckase 2006). Van der Linden (1995, S. 110) fasst diesen Standpunkt wie folgt zusammen: „… the correct view is to see the standard-setting methods as methods to *set* true standards – not to reflect them". Validität wird im Bereich des Standard-Settings häufig im Sinne Samuel Messicks (1994) konzeptualisiert, d.h. sie wird nicht als Eigenschaft einer Kompetenzskala *per se* verstanden, sondern als Eigenschaft der Interpretationen und der

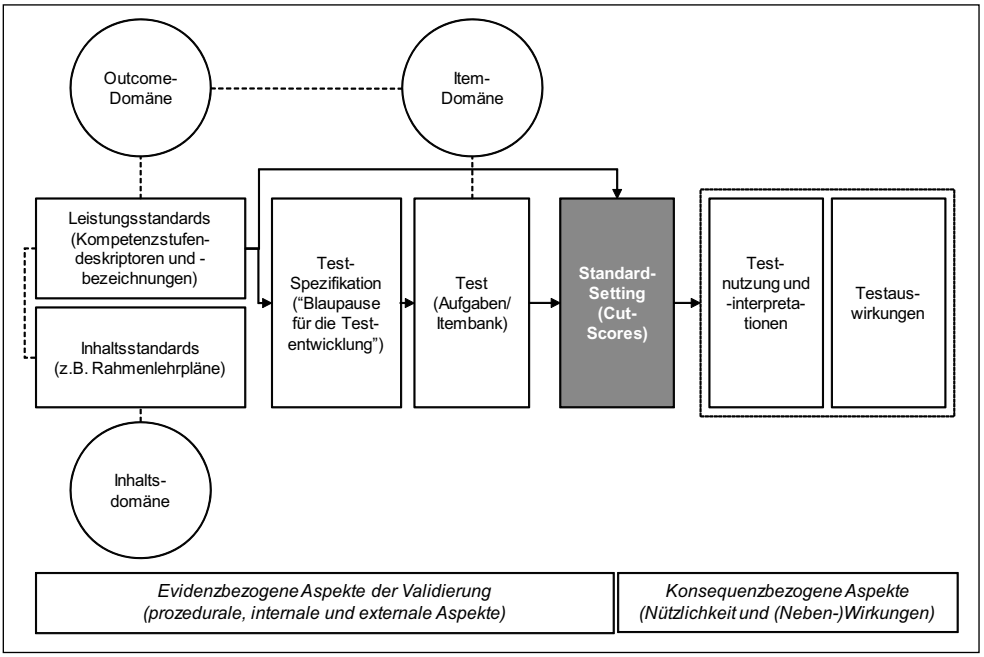


Abb. 1: Schematische Darstellung der Funktion des Standard-Settings im Large-Scale-Assessment

Verwendungen, die mit einer Kompetenzstufeneinteilung verbunden werden. Nach Messicks Verständnis ist es dazu notwendig, ein kohärentes Validitätsargument zu entwickeln, das empirische Befunde zu verschiedenen evidenzbezogenen Teilaspekten „klassischer" Validitätskonzepte, wie z.B. Inhalts-, Kriteriums- oder Konstruktvalidität, aber auch die *sozialen* Konsequenzen der Testanwendung diskursiv abwägt und integriert.

Generell sollte zwischen dem Validitätsargument für das System des Large-Scale-Assessments als Ganzes und dessen Subsystemen unterschieden werden, von denen das Standard-Setting eines darstellt (vgl. Abb. 1). Die Festlegung der Cut-Scores stellt allerdings ein besonders kritisches Verbindungsglied zwischen den evidenzbezogenen, empirisch gut untersuchbaren Aspekten des Gesamtsystems und den konsequenzbezogenen, eher normativen und praxisrelevanten Aspekten dar (vgl. Pant u.a. 2009).

2.3 Projektziele und eigener Forschungsbeitrag

Ein Validitätsargument im Sinne Messicks (1994) wird in diesem Projekt in vier Teilstudien entwickelt, die jeweils einen eigenen Validitätsaspekt fokussieren.

In der ersten Teilstudie zur Inhaltsvalidität wird die Passung zwischen den Kompetenzstufendeskriptoren der Bildungsstandards für die erste Fremdsprache Englisch sowie des Gemeinsamen Europäischen Referenzrahmens für Sprachen (GERS) einerseits und den am IQB entwickelten Testaufgaben andererseits evaluiert. Im zweiten Fokus wird die interne Validität der Cut-Scores untersucht. In quasi-experimentellen Designs wird analysiert, welchen Einfluss die fachliche Zusammensetzung von Expertenpanels und die gewählte Methode des Standard-Settings auf die resultierenden Cut-Scores haben. In einem dritten Schritt zur externen Validierung wird betrachtet, in welchem Maße eine Klassifizierung von Schüler/innen in Kompetenzniveaus durch ihre Lehrer/innen mit den testbasierten Klassifizierungen nach erfolgtem Standard-Setting übereinstimmen. Im vierten und letzten Schritt werden konsequenzbezogenen Aspekte der Validität betrachtet. Hierbei geht es vorrangig um die Frage, ob die aus dem Standard-Setting resultierenden Kompetenzstufenfestlegungen in der medialen Öffentlichkeit und von anderen Rezipient/innen der Kompetenzstufenrückmeldungen (Lehrkräfte, Bildungsadministration) so aufgenommen und interpretiert werden, dass sie mit den Intentionen der Bildungsstandards vereinbar und somit in Messicks Sinne valide sind.

3. Standard-Setting für das Kompetenzstufenmodell der Bildungsstandards Englisch

3.1 Das Kompetenzstufenmodell der Bildungsstandards und des Gemeinsamen Europäischen Referenzrahmens für Sprachen im Fach Englisch

Die in den KMK-Bildungsstandards formulierten Kompetenzmodelle für Fremdsprachenhören bzw. -lesen sind fast vollständig aus dem GERS übernommen worden (vgl. Europarat 2001; Figueras u.a. 2005).

Der GERS beschreibt u.a., welche Kompetenzen Fremdsprachenlernende aufweisen sollen, „… um eine Sprache für kommunikative Zwecke zu benutzen, und welche Kenntnisse und Fertigkeiten sie entwickeln müssen, um in der Lage zu sein, kommunikativ erfolgreich zu handeln“ (Europarat 2001, S. 14). Für die kommunikativen Aktivitäten werden die drei Basis-Niveaustufen A (elementare Sprachverwendung), B (selbständige Sprachverwendung) und C (kompetente Sprachverwendung) unterschieden, die in je zwei Unterniveaus aufgeteilt werden. Die Unterniveaus werden im GERS anhand von Kann-Beschreibungen konkretisiert (Beispiel für die Stufe B2 aus der Globalskala des GERS: „Kann die Hauptinhalte komplexer Texte zu konkreten und abstrakten Themen verstehen; versteht im eigenen Fachgebiet auch Fachdiskussionen“; Europarat 2001, S. 35). Die Niveaustufen beschreiben dabei sukzessiv und kumulativ zu erlernende Teilkompetenzen.

3.2 Eingesetzte Testmaterialien

Die auf der Basis der Bildungsstandarddokumente und des GERS entwickelten Testaufgaben wurden 2007 an $N = 2.932$ Schüler/innen der Jahrgangsstufen 8–10 aller Bildungsgänge in 15 Ländern pilotiert. Das Aufgabenmaterial und Details der Aufgabenentwicklung sind andernorts ausführlich beschrieben (vgl. Rupp u.a. 2008). Die Aufgaben wurden in Form eines Balanced Incomplete Block Designs (vgl. van der Linden/Veldkamp/Carlson 2004) so auf die verschiedenen Testhefte aufgeteilt, dass eine gemeinsame Skalierung aller Aufgaben möglich war. Die Testleistungen wurden mit dem Programm Acer ConQuest skaliert (vgl. Wu/Adams/Wilson 1998). Dabei wurde ein zweidimensionales Modell mit je einer Dimension für Leseverständnis und Hörverständnis geschätzt. Insgesamt wurde jedes Item von durchschnittlich $N = 330$ Personen bearbeitet. Aus diesem Itempool wurden insgesamt je 74 Items zum Lese- bzw. Hörverstehen für das Standard-Setting ausgewählt. Als Auswahlkriterien galten eine Gleichverteilung über die *a priori* eingeschätzten GERS-Niveaus (A1–C1), die annähernd gleiche Fächerung der Itemschwierigkeiten pro GERS-Niveau und „Repräsentativität“ von Itemformaten, Konstruktfacetten (z.B. Art des getesteten Leseverhaltens) und Textsorten. Die Personenparameterschätzer wurden auf eine Skala mit Mittelwert $M = 500$ und Standardabweichung $SD = 100$ (9. Jahrgangsstufe) transformiert.

3.3 Design und Durchführung der Standard-Setting-Studie

In der ersten Projektphase hatten $N = 45$ Expert/innen im Rahmen einer viertägigen Standard-Setting-Klausur die Aufgabe, die metrischen IRT-Kompetenzskalen für Lese- und für Hörverstehen durch das Setzen von jeweils vier Cut-Scores (A1/A2; A2/B1; B1/B2 und B2/C1) in die Kompetenzstufen des GERS zu unterteilen.

In einem quasi-experimentellen 2 × 2-Design wurde der Einfluss der beiden in der Forschungsliteratur herausgestellten Faktoren (A) *Panelzusammensetzung* (homogene

Panels ausschließlich mit Lehrkräften vs. heterogene Panels aus Lehrkräften, Vertreter/innen aus Fachdidaktik, Psychometrie und Bildungsadministration) und (B) *Standard-Setting-Methode* (klassische Bookmark-Methode vs. modifizierte Bookmark-Methode[3]) auf die Platzierung der Cut-Scores untersucht. Die Zuweisung der Lehrkräfte bzw. heterogenen Panelteilnehmer/innen auf die beiden Standard-Setting-Bedingungen erfolgte pro Untergruppe randomisiert.

Runde 1
1. Die Panelteilnehmer/innen erhalten 60 Min. Zeit, sich mit dem OIB vertraut zu machen, indem sie (a) alle 74 Items lesen bzw. anhören, (b) diskutieren, welche Kompetenzen, Fähigkeiten und Fertigkeiten zur Lösung jedes Items erforderlich sind, und (c) diskutieren, welche Itemmerkmale die empirische Schwierigkeitsreihung bewirkt haben.
2. Die Panelteilnehmer/innen setzen individuell erstmalig die vier Cut-Scores (75 Min.).
Runde 2
1. Die Teilnehmer/innen erhalten als Feedback die Cut-Scores der übrigen Mitglieder inkl. Mittelwerte und Streuung der Cut-Scores.
2. Anhand der Kompetenzniveaudeskriptoren des GERS diskutieren die Panelist/innen in Kleingruppen, welche Kompetenzen ein/e Schüler/in aufweisen muss, um ein bestimmtes GERS-Niveau zu erreichen.
3. Die Panelteilnehmer/innen setzen individuell zum zweiten Mal die vier Cut-Scores (75 Min.).
4. Die Teilnehmer/innen erhalten als Feedback erneut die Cut-Scores der übrigen Mitglieder inkl. Mittelwerte und Streuung der Cut-Scores. Die Ergebnisse werden in der Gesamtgruppe diskutiert.
Runde 3
1. Den Panelteilnehmer/innen werden Wirkungsdaten (Impact Data) präsentiert, d.h. die prozentuale Verteilung der Schüler/innen auf die Kompetenzstufen, wenn man die Cut-Scores der zweiten Runde zugrunde legte.
2. Die Panelteilnehmer/innen diskutieren die Wirkungsdaten in der Gesamtgruppe mit dem Ziel, möglichst eine Konvergenz der Einzelurteile zu erreichen.
3. Die Panelteilnehmer/innen setzen wiederum individuell die vier finalen Cut-Scores (75 Min.).

Anmerkung: OIB (Ordered Item Booklet) bezeichnet das nach aufsteigenden empirischen Schwierigkeiten geordnete „Buch" der Einzelitems.

Tab. 1: Ablauf einer Panelsitzung bei der „klassischen" Bookmark-Methode

3 In der zweiten Standard-Setting-Klausur (Juni 2009) wird bei diesem Experimentalfaktor die klassische Bookmark-Methode mit dem modifizierten Angoff-Verfahren verglichen.

Die klassische Bookmark-Methode wurde in Abschnitt 2 bereits dargestellt. Die modifizierte Bookmark-Variante (auch: *Criterion-Map*-Methode) wurde am Berkeley Evaluation & Assessment Research Center (BEAR) der University of California entwickelt (vgl. Wilson/Draney 2002, 2004). Im Unterschied zum klassischen Verfahren ermöglicht die Criterion-Map-Methode den Panelteilnehmer/innen eine computergestützte Visualisierung des Standard-Setting-Prozesses. So wird die Schwierigkeitsverteilung der Items visualisiert, ebenso die Relationen der jeweils gesetzten Cut-Scores zueinander und die auf jedem Kompetenzniveau befindlichen Items. Auch die aus den Cut-Scores resultierende Personenverteilung auf die Niveaus kann unmittelbar dargestellt werden. Der zweite Unterschied bei der modifizierten Variante besteht darin, dass die finalen Cut-Scores – nach intensiver Diskussion der Panelmitglieder – im *Konsensverfahren* bestimmt werden sollen, während dies bei der klassischen Bookmark-Methode durch Mittelung der Einzelurteile geschieht.

Alle Expert/innen unterzogen sich vor dem eigentlichen Workshop einer Familiarisierungsübung zum GERS. Hierbei wurde die Fähigkeit der Teilnehmer/innen zur korrekten Zuordnung von Kompetenzstufendeskriptoren zu den Kompetenzniveaus (A1 bis C2) eingeübt und überprüft, Fehlzuordnungen wurden diskutiert. In Tabelle 1 ist zur Illustration der Ablauf eines Bookmark-Panels skizziert. Ein Feedback über die Häufigkeitsverteilungen von Schüler/innen auf die Kompetenzniveaus, die sich aus den gesetzten Cut-Scores ergeben, soll den Panelmitgliedern erlauben, ggf. realitätsgerechte Adjustierungen ihrer Entscheidungen vorzunehmen.

3.4 Ergebnisse

Die vorgestellten Ergebnisse beziehen sich auf die erste von insgesamt drei Standard-Setting-Studien und sind daher als vorläufig zu betrachten. Von den vier Teilstudien (vgl. Abschnitt 2.1) liegen zunächst Befunde zur internen und zur externen Validierung des Standard-Settings vor.

Interne Validierung. Die im Abschnitt 3.3 beschriebenen Unterschiede im experimentellen Faktor ‚Standard-Setting-Methode' führen dazu, dass in der Bedingung „klassische Bookmark-Methode" *pro Panelmitglied* 24 Cut-Score-Informationen vorliegen, nämlich für 2 Fähigkeiten (Hören, Lesen) × 4 Cut-Score-Niveaus (A1/A2, A2/B1, B1/B2, B2/C1) × 3 Runden. In der Criterion-Map-Bedingung hingegen werden aufgrund des Konsensverfahrens keine individuellen Cut-Scores generiert. Insgesamt werden hier lediglich acht Cut-Score-Informationen *pro Gruppe* ermittelt, d.h. 2 Fähigkeiten (Hören, Lesen) × 4 Cut-Score-Niveaus (A1/A2, A2/B1, B1/B2, B2/C1). Eine gemeinsame statistische Auswertung des 2 × 2-Designs ist daher nicht angebracht.

Die Cut-Score-Daten der klassischen Bookmark-Panels wurden getrennt für Lese- und Hörverstehen in Varianzanalysen mit Panelkomposition als Zwischensubjektfaktor sowie für Runden und Cut-Score-Niveaus als Messwiederholungsfaktoren ausgewertet. Bei beiden rezeptiven Aktivitäten zeigt sich ein signifikanter Effekt der Panelkomposition, d.h. homogene Panels aus Lehrkräften setzen insgesamt niedrigere Cut-Scores als heterogen zusammengesetzte (Lesen: $F = 7.1$; $df_1 = 1$, $df_2 = 20$; $p < .05$; $Eta^2_{part.} = .26$;

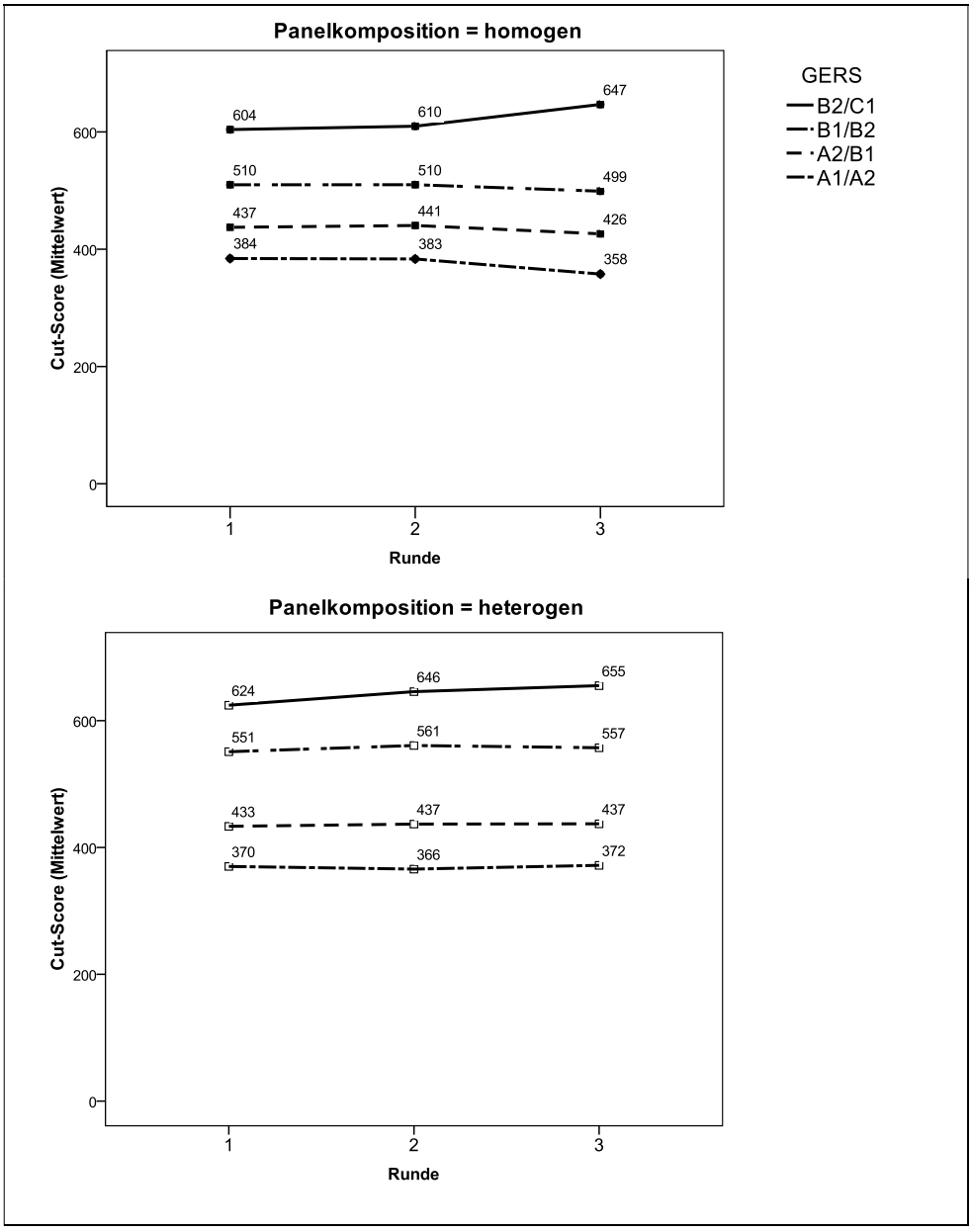


Abb. 2: Gemittelte Cut-Scores im Leseverstehen nach Durchgangsrunde, GERS-Niveau und Panelkomposition unter Verwendung der klassischen Bookmark-Methode

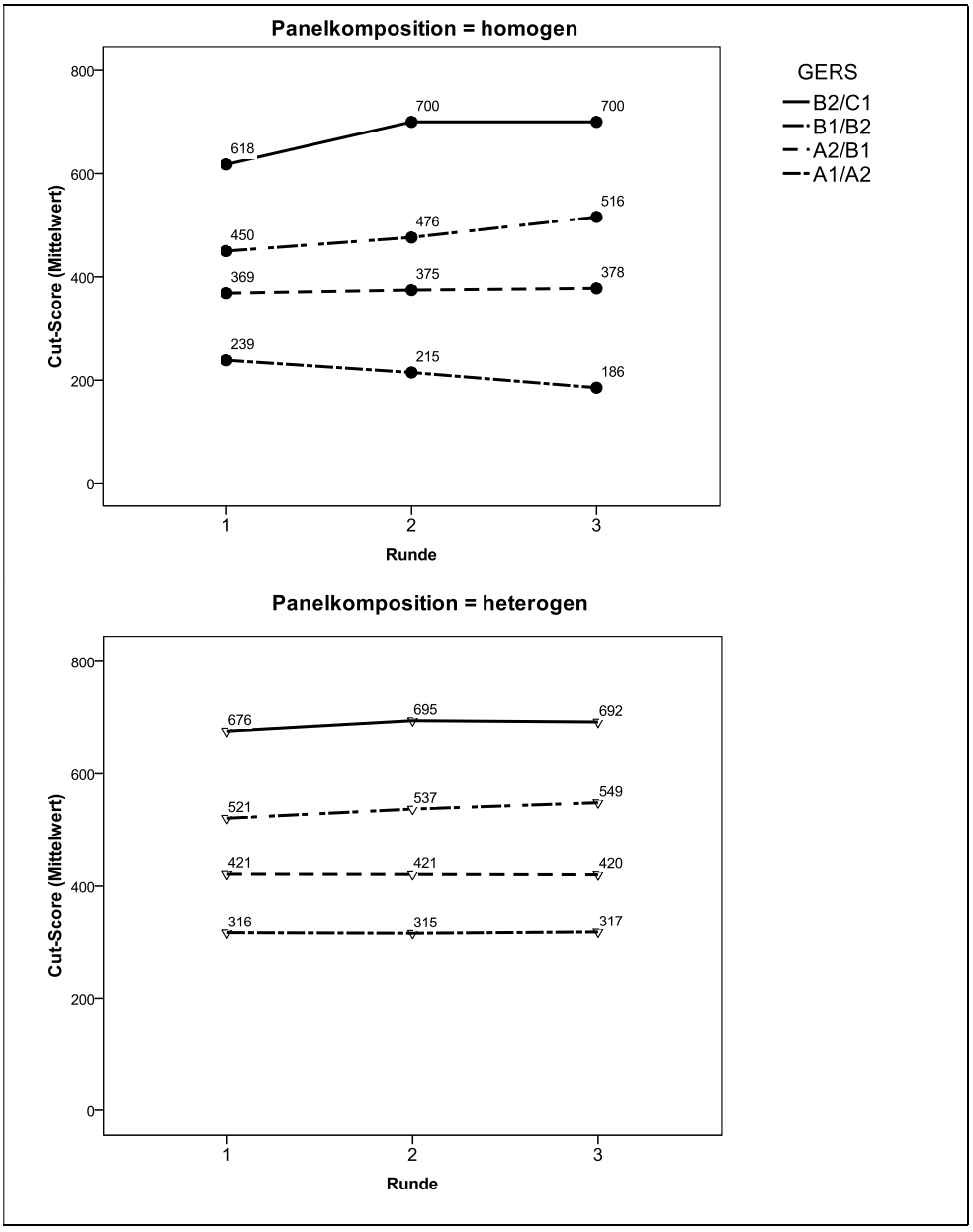


Abb. 3: Gemittelte Cut-Scores im Hörverstehen nach Durchgangsrunde, GERS-Niveau und Panelkomposition unter Verwendung der klassischen Bookmark-Methode

Hören: $F = 11.9$; $df_1 = 1$, $df_2 = 20$; $p < .01$; $Eta^2_{part.} = .37$). Dieser Haupteffekt ist allerdings im Lichte mehrerer signifikanter Wechselwirkungseffekte zu relativieren (siehe hierzu auch die Abb. 2 und 3).

So wirkt sich die Panelkomposition je nach betrachteten Niveaustufen des GERS, zwischen denen ein Schwellenwert zu platzieren war, signifikant unterschiedlich aus (Wechselwirkung Panelkomposition x GERS-Niveau). Bei der Lesekompetenz setzen Lehrkräftepanels erst bei der Grenzziehung zwischen den Niveaustufen B1 und B2 deutlich früher den Cut als Panels mit gemischtem Fachhintergrund ($F = 14.7$; $df_1 = 3$, $df_2 = 18$; $p < .001$; $Eta^2_{part.} = .71$). Bei den Aufgaben zum Hörverstehen tritt dieser Mildeeffekt der Lehrkräfte bereits beim untersten Schwellenwert (A1/A2) auf. Die sowohl beim Lesen wie auch beim Hören auftretenden substantiellen Dreifachwechselwirkungen (Panelkomposition × GERS-Niveau × Runde) zeigen weiter (siehe Abb.2 und 3), dass die niveauspezifischen Effekte der Panelkomposition vor allem in der finalen dritten Runde akzentuiert wurden. Bemerkenswert ist weiterhin, dass sowohl beim Lesen als auch beim Hören *im Mittel* die Cut-Scores über die Runden nicht signifikant unterschiedlich gesetzt wurden, solche Schwankungen jedoch auf einzelnen Niveaustufen sehr deutlich auftraten (Wechselwirkung Runde × GERS-Niveau; Lesen: $F = 4.8$; $df_1 = 6$, $df_2 = 15$; $p < .01$; $Eta^2_{part.} = .66$; Hören: $F = 10.9$; $df_1 = 6$, $df_2 = 15$; $p < .001$; $Eta^2_{part.} = .81$). Sie waren vor allem bei den Randkategorien (A1/A2 bzw. B2/C1) zu beobachten.

Für das gesamte 2 × 2-Design wurden pro Gruppe die finalen Cut-Scores deskriptiv betrachtet. Abbildungen 4 und 5 verdeutlichen, dass der beschriebene Panelkompositionseffekt nur bei der klassischen, nicht aber bei der modifizierten Bookmark-Variante erkennbar auftritt.

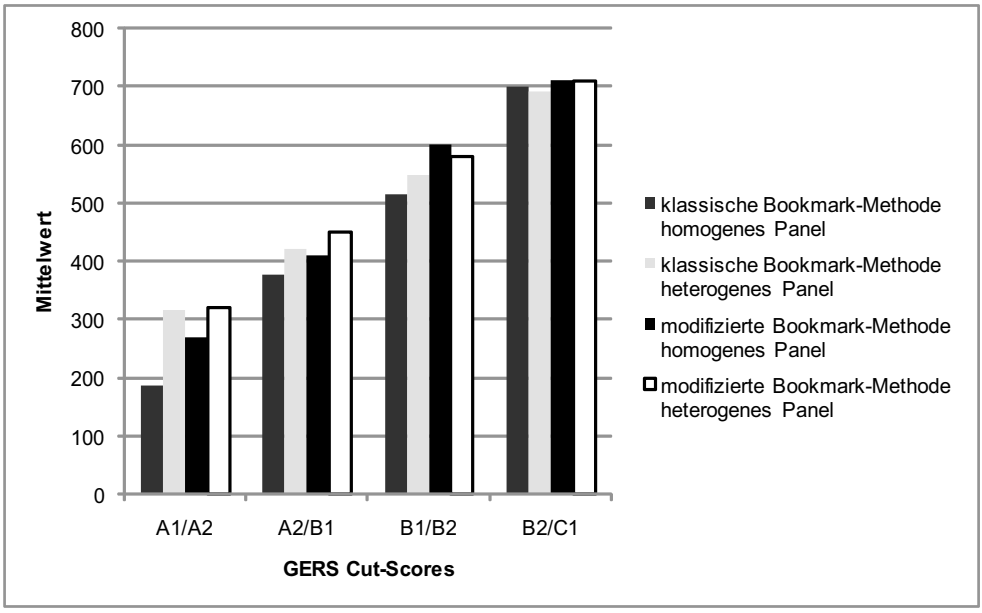


Abb. 4: Finale Cut-Scores aller vier Experimentalgruppen im Leseverstehen nach GERS-Niveau

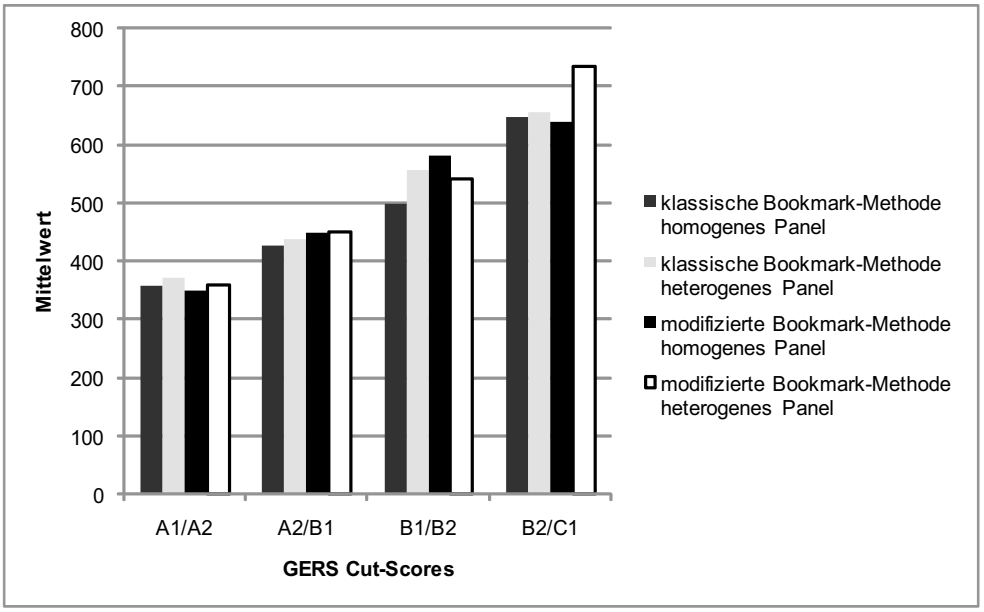


Abb. 5: Finale Cut-Scores aller vier Experimentalgruppen im Hörverstehen nach GERS-Niveau

Externe Validierung. Für jede/n Schüler/in der verwendeten Pilotierungsstichprobe wurde erhoben, wo auf dem GERS die unterrichtende Englischlehrkraft sie bzw. ihn einstuft. In einem ersten Auswertungsschritt wurde überprüft, wie gut Stufenzuordnungen von Lehrkräften und die aus dem Standard-Setting resultierende Kompetenzstufenzuordnung übereinstimmen (Kreuzklassifikation). Beim Leseverstehen kommen beide Klassifikationsansätze in 38% zu einer identischen Einstufung, in 80% der Fälle wurde innerhalb von ± 1 GERS-Stufe gleich zugeordnet (Hörverstehen: 40% bzw. 84%). Detaillierte mehrebenenanalytische Auswertungen zu den Einflussfaktoren kongruenter bzw. divergenter Klassifikationen werden andernorts beschrieben (vgl. Leucht u.a. 2009).

4. Diskussion und Ausblick

Das Setzen von Schwellenwerten auf einer metrischen Kompetenztestskala beruht beim Standard-Setting, trotz detailliert vorgegebener prozeduraler Verfahrensrichtlinien, letztlich auf Expertenurteilen. Die ersten Befunde unserer Studie belegen, dass die Höhe der finalen Cut-Scores sowohl hinsichtlich der professionalen Zusammensetzung des Beurteilerpanels als auch für prozedurale Varianten des Verfahrens sensibel ist. Die kommenden Standard-Setting Studien werden auf größerer Datenbasis zu zeigen haben, wie stabil die gefundenen Effekte sind. Dazu gehört auch die Frage, ob der gezeigte Mildeeffekt in den Lehrkräftepanels nach Bildungsgang (z.B. Hauptschul- vs. Gymnasiallehrkräfte) differenziert werden muss.

In Rückmeldeprotokolle der Expert/innen zeichneten sich darüber hinaus typische „neuralgische Punkte“ ab, die auf divergierende Repräsentationen von Schlüsselkonzepten des Standard-Settings verweisen. Hierzu zählen Unklarheiten hinsichtlich der Konzepte Itembeherrschung (*Mastery*) bzw. Lösungswahrscheinlichkeit (*Response Probability*) und die Überbetonung von „nicht passenden“ Einzelitems in Gruppen von Items, die als gleich schwierig eingeschätzt werden („*Odd-one-out-Phänomen*“). Ziel der folgenden Studienphasen ist es u.a., derartige kognitive Prozesse mit Hilfe von Think-aloud-Techniken explizit zu machen.

Seit dem Jahr 2009 stehen im Fach Englisch normierte Testaufgaben zur Verfügung, die die Tests zur Überprüfung der Bildungsstandards über Ankeritems mit denen der Vergleichsarbeiten in der Jahrgangsstufe 8 verknüpfen, sodass die Leistungen der Schüler/innen in den Vergleichsarbeiten zu den Befunden aus länderübergreifenden Stichprobentestungen in Beziehung gesetzt werden können. Durch diese Entwicklung haben sich Verwertungszusammenhang und Generalisierungsanspruch für das hier betrachtete Standard-Setting-Verfahren erheblich erweitert und damit auch die Anforderungen an die Validierung des Verfahrens. Neben Vergleichen der Kompetenzstände von Schülerschaften eines Landes, einer Region oder einer Schule innerhalb eines Erhebungsjahres sollen nun auch Cross-Grade-Vergleiche (8. und 9. Jahrgangsstufe) und damit verbundene Entwicklungsprognosen möglich werden

Abgesehen von den statistischen und interpretatorischen Schwierigkeiten, die mit Kompetenzstufungen in Cross-Grade- bzw. „Growth-to-Standard“-Anwendungen verbunden sind (vgl. Ho 2007; Lissitz/Wei 2008), ist bisher nicht untersucht worden, wie sich verschiedene Testzwecke (High-Stakes vs. Low-Stakes) auf das Verhalten von Expert/innen in Standard-Setting Verfahren auswirken. Die folgenden Projektphasen werden auch diese Fragen aufgreifen.

Literatur

Angoff, W.H. (1971): Scales, norms, and equivalent scores. In: Thorndike, R.L. (Hrsg.): Educational measurement. Washington, DC: American Council on Education, S. 508–600.

Buckendahl, C.W./Smith, R.W./Impara, J.C./Plake, B.S. (2002): A comparison of Angoff and Bookmark standard setting methods. In: Journal of Educational Measurement 39, S. 253–263.

Cizek, G.J./Bunch, M.B. (2007): Standard-setting. A guide to establishing and evaluating performance standards on tests. California: Sage Publications Inc.

Europarat (2001): Gemeinsamer europäischer Referenzrahmen für Sprachen. Berlin: Langenscheidt.

Figueras, N./North, B./Takala, S./Verhelst, N./Van Avermaet, P. (2005): Relating examinations to the Common European Framework: a manual. In: Language Testing 22, S. 261–279.

Green, D.R./Trimble, C.S./Lewis, D.M. (2003): Interpreting the results of three different standard setting procedures. In: Educational Measurement: Issues and Practice 22, S. 22–32.

Ho, A.D. (2007): Discrepancies between score trends from NAEP and state tests: A scale-invariant perspective. In: Educational Measurement: Issues and Practice 26, H. 4, S. 11–20.

Hurtz, G.M./Auerbach, M.A. (2003): A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. In: Educational and Psychological Measurement 63, S. 584–601.

Kane, M.T. (2001): So much remains the same: Conception and status of validation in setting standards. In: Cizek, G. (Hrsg.): Setting performance standards: Concepts, methods and perspectives. Mahwah, NJ: Erlbaum, S. 53–88.

Karantonis, A./Sireci, S.G. (2006): The Bookmark standard-setting method: A literature review. In: Educational Measurement: Issues and Practice 25, H. 1, S. 4–12.

Kultusministerkonferenz (KMK) (2003): Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) für den Mittleren Schulabschluss. München: Wolters-Kluwe.

Leucht, M./Tiffin-Richards, S.P./Pant, H.A./Köller, O. (2009): Diagnostische Kompetenz von Lehrkräften in der ersten Fremdsprache Englisch. Manuskript eingereicht zur Publikation.

Lissitz, R.W./Wei, H. (2008): Consistency of standard-setting in an augmented state testing system. In: Educational Measurement: Issues and Practice 27, H. 2, S. 46–55.

Messick, S. (1994): The interplay of evidence and consequences in the validation of performance assessments. In: Educational Researcher 23, H. 2, S. 13–23.

Mitzel, H.C./Lewis, D.M./Patz, R.J./Green, D.R. (2001): The Bookmark procedure: Psychological perspectives. In: Cizek, G. (Hrsg.): Setting performance standards: Concepts, methods and perspectives. Mahwah, NJ: Erlbaum, S. 249–281.

OECD (2009): PISA 2006 technical report. Paris: OECD.

Pant, H.A./Rupp, A.A./Tiffin-Richards, S./Köller, O. (2009): Validity issues in standard-setting studies. In: Studies in Educational Evaluation 35, S. 95–101.

Reckase, M.D. (2006): Psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. In: Educational Measurement: Issues and Practice 25, H. 2, S. 4–18.

Rupp, A.A./Vock, M./Harsch, C./Köller, O. (2008): Developing standards-based assessment tasks for English as a first foreign language – Context, processes, and outcomes in Germany. Münster: Waxmann.

Van der Linden, W.J. (1995): A conceptual analysis of standard-setting in large-scale assessments. In: Crocker, L./Zieky, M. (Hrsg.): Proceedings of the joint conference on standard-setting for large-scale assessments. Washington, DC: National Assessment Governing Board & National Center for Education Statistics, S. 97–118.

Van der Linden, W.J./Veldkamp, B.P./Carlson, J.E. (2004): Optimizing Balanced Incomplete Block Designs for educational assessments. In: Applied Psychological Measurement 28, S. 317–331.

Van Nijlen, D./Janssen, R. (2008): Modeling judgments in the Angoff and Contrasting-Groups method of standard setting. In: Journal of Educational Measurement 45, S. 45–63.

Wilson, M./Draney, K. (2002): A technique for setting standards and maintaining them over time. In: Nishisato, S./Baba, Y./Bozdogan, H./Kanefugi, K. (Hrsg.): Measurement and multivariate analysis. Tokyo: Springer, S. 325–332.

Wilson, M./Draney, K. (2004): Some links between large-scale and classroom assessments: The case of the BEAR assessment system. In: Yearbook of the National Society for the Study of Education 103, H. 2, S. 132–154.

Wu, M.L./Adams, R.J./Wilson, M.R. (1998): ConQuest: Multi-Aspect Test Software [computer program]. Camberwell, Victoria: Australian Council for Educational Research.

Yin, P./Schulz, E.M. (2005, April): A comparison of cut scores and cut score variability from Angoff-based and Bookmark-based procedures in standard setting. Vortrag gehalten beim Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

Anschrift der Autoren

Prof. Dr. Hans Anand Pant, Institut zur Qualitätsentwicklung im Bildungswesen (IQB),
Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin
E-Mail: hansanand.pant@iqb.hu-berlin.de

Simon P. Tiffin-Richards, M.Sc., Institut für Schulqualität der Länder Berlin und Brandenburg (ISQ), Freie Universität Berlin, Otto-von-Simson-Str. 15; D-14195 Berlin
E-Mail: simon.tiffin-richards@cms.hu-berlin.de

Prof. Dr. Olaf Köller, Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN) an der Universität Kiel, Olshansenstr. 62, D-24098 Kiel
E-Mail: koeller@ipn.uni-kiel.de

Johannes Hartig/Jana Höhler

Modellierung von Kompetenzen mit mehrdimensionalen IRT-Modellen

Projekt MIRT[1]

1. Einleitung

Kompetenzen werden im Rahmen des Schwerpunktprogramms 1293 als kontextabhängige, kognitive Leistungsdispositionen definiert (vgl. auch Weinert 2001; Hartig/Klieme 2006). Im Sinne dieser Definition kann eine Person mit einer spezifischen Kompetenz *Anforderungen in einem spezifischen Bereich von Situationen* erfolgreich bewältigen. Die zu erfassende Kompetenz ist hierbei ein *theoretisches Konstrukt* zur Beschreibung und Erklärung von Leistungsunterschieden in diesen Situationen. Wenn Kompetenzen in standardisierten Tests erfasst werden, stellt das Kompetenzkonstrukt die theoretische Verbindung zwischen dem Testverhalten und den daraus gezogenen diagnostischen Schlüssen dar. *Psychometrische Modelle* oder *Messmodelle* sind in diesem Zusammenhang das Werkzeug, um diese theoretische Verbindung in konkrete Messverfahren umzusetzen (vgl. z.B. Mislevy u.a. 2002; Wilson 2005).

In psychometrischen Modellen werden Annahmen über die Struktur des interessierenden Konstrukts sowie über die Zusammenhänge zwischen dem Konstrukt und dem beobachtbaren Testverhalten formuliert. Idealerweise sollte in einem psychometrischen Modell abgebildet werden, wie konkrete *situative Anforderungen* und *individuelle Ressourcen* bei der Bewältigung dieser Anforderungen interagieren (vgl. Embretson 1983; Hartig 2008). Die zu messenden Konstrukte werden in psychometrischen Modellen durch *latente Variablen* repräsentiert (vgl. Borsboom/Mellenbergh/van Heerden 2003). Diese Variablen dienen der Erklärung der Verteilungen und Zusammenhangsstrukturen beobachtbarer Variablen, die aus dem Testverhalten gebildet werden (vgl. Skrondal/Rabe-Hesketh 2004). Umgekehrt wird auf Basis des gewählten psychometrischen Modells vom beobachtbaren Testverhalten auf individuelle Ausprägungen oder Populationsverteilungen der latenten Variablen geschlossen.

Modelle mit latenten Variablen unterscheiden sich unter anderem hinsichtlich des Skalenniveaus der beobachteten Indikatoren und der latenten Variablen. Modelle mit *kategorialen Indikatoren* (vgl. z.B. Aufgabe gelöst/nicht gelöst) und *kontinuierlichen latenten Variablen*, auf die in diesem Forschungsvorhaben fokussiert wird, werden zum Beispiel traditionell unter dem Begriff *Item-Response-Theorie* (IRT) zusammengefasst. In den letzten Jahren wird die separate Behandlung dieser Modelle zunehmend zugunsten

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: HA 5050/2-1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

ten einer breiteren und flexibleren Sichtweise aufgegeben. Spezifische Modelle stellen in einem solchen breiteren Rahmen Spezialfälle dar, die für bestimmte Arten von beobachteten Variablen geeignet sind (vgl. Muthén 2002; Rabe-Hesketh/Skrondal/Pickles 2004; Skrondal/Rabe-Hesketh 2004).

Die Modellierung von Kompetenzen bezieht *quantitative* und *qualitative* Aspekte der Beschreibung von interindividuellen Unterschieden mit ein. Der erste *qualitative* Aspekt ist die Differenzierung von (Sub-) Dimensionen, die durch verschiedene latente Variablen repräsentiert werden. Die interindividuellen Unterschiede in den latenten Variablen stellen den *quantitativen* Aspekt der Modellierung dar. Diese beiden Aspekte werden in den folgenden Abschnitten anhand erster Projektergebnisse näher erörtert. Zusätzlich wird oft angestrebt, Personen mit unterschiedlich stark ausgeprägten (Teil-) Kompetenzen dahingehend zu beschreiben, welche situativen Anforderungen sie bewältigen können. Diese *kriterienorientierte* Beschreibung, die oft mit der Definition von „Kompetenzniveaus" oder „Kompetenzstufen" erreicht wird, ist der zweite *qualitative* Aspekt bei der psychometrischen Modellierung von Kompetenzen (vgl. z.B. Hartig 2007, 2008).

Psychometrische Modelle können ein- oder mehrdimensional sein. In der pädagogisch-psychologischen Diagnostik überwiegt die Anwendung eindimensionaler Messmodelle, die zur Modellierung von Unterschieden zwischen Personen eine einzelne latente Variable beinhalten (vgl. z.B. Gonzalez 2003; Gonzalez/Galia/Li 2004; Hartig/Jude/Wagner 2008). Zur psychometrischen Modellierung der in einem Test erfassten Kompetenz können jedoch auch mehrdimensionale psychometrische Modelle, insbesondere mehrdimensionale IRT-Modelle (MIRT-Modelle) verwendet werden, in denen das Kompetenzkonstrukt differenziert hinsichtlich mehrerer zugrunde liegender Teilkompetenzen modelliert wird (vgl. z.B. Ackerman/Gierl/Walker 2003; Adams/Wilson/Wang 1997; Hartig/Höhler 2008; McDonald 1997; Reckase 2007, 2009). Häufig werden einfache eindimensionale Modelle eher aus pragmatischen als aus theoretischen Gründen gewählt. Verglichen mit eindimensionalen Modellen kann die Erfassung einer Kompetenz mittels eines mehrdimensionalen psychometrischen Modells eine differenziertere Diagnostik und zugleich eine Prüfung von Annahmen über die Struktur der erfassten Kompetenz und Teilkompetenzen ermöglichen (vgl. z.B. Ackerman/Gierl/Walker 2003; Walker/Beretvas 2003).

Der vorliegende Artikel skizziert zunächst kurz die Zielsetzung des DFG-Projekts „Modellierung von Kompetenzen mit mehrdimensionalen IRT-Modellen". Im darauf folgenden Abschnitt werden zwei wesentliche Charakteristika zur Unterscheidung von MIRT-Modellen sowie deren inhaltliche Implikationen für die Modellierung von Kompetenzen dargestellt.

2. Zielsetzung

Die zentrale Fragestellung des DFG-Projekts „Modellierung von Kompetenzen mit mehrdimensionalen IRT-Modellen" ist, inwieweit sich sprachliche Kompetenzen, die mit demselben Aufgabenmaterial erfasst wurden, auf unterschiedliche Weise modellie-

ren lassen und welche diagnostischen Aussagen aus den verschiedenen Auswertungen resultieren. Insbesondere interessiert, inwieweit komplexere mehrdimensionale Modelle gegenüber eindimensionalen Modellen (z.B. dem eindimensionalen Raschmodell) einen Zugewinn an diagnostischer Information liefern.

In dem Projekt wurden zunächst verschiedene IRT-Modelle auf dieselben Daten angewendet. Dafür wurde auf bereits vorliegende Leistungsdaten der DESI-Studie (Deutsch Englisch Schülerleistungen International) (vgl. Beck/Klieme 2007; Klieme u.a. 2008) zurückgegriffen. Verwendet werden Daten aus den Tests für Sprachbewusstheit (Grammatik), Lese- und Hörverstehen für Englisch als Fremdsprache zum Ende der neunten Jahrgangsstufe.

3. Charakteristika mehrdimensionaler IRT-Modelle

Im Weiteren werden zwei wesentliche Charakteristika mehrdimensionaler IRT-Modelle beschrieben, die weitreichende inhaltliche Konsequenzen für die Interpretation der latenten Variablen im Modell haben. Zunächst wird auf die Komplexität des Ladungsmusters eingegangen (Between- versus Within-Item-Mehrdimensionalität), im darauf folgenden Abschnitt auf kompensatorische versus nicht-kompensatorische Verknüpfungen mehrerer latenter Variablen bei komplexer Ladungsstruktur.

3.1 Between- versus Within-Item-Mehrdimensionalität

In den meisten empirischen Anwendungen von MIRT-Modellen zur Messung von Schülerkompetenzen wird für jede zu erfassende Kompetenz jeweils eine separate Dimension (latente Variable) definiert. Im Kontext der PISA-Studien findet sich auch eine mehrdimensionale Modellierung von Subdimensionen *innerhalb* einzelner Kompetenzbereiche (vgl. z.B. Blum u.a. 2004). In beiden Fällen lädt jedes Item nur auf einer Dimension. Die Modelle sind durch eine *Einfachstruktur* oder *Between-Item*-Mehrdimensionalität (vgl. Adams/Wilson/Wang 1997) gekennzeichnet. Die latenten Variablen repräsentieren in diesen Modellen das Leistungsniveau in den durch die jeweiligen Item-Mengen definierten Bereichen.

MIRT-Modelle können jedoch auch eine komplexe Ladungsstruktur (*Within-Item*-Mehrdimensionalität) haben, bei der Items gleichzeitig durch mehrere latente Dimensionen beeinflusst werden. Bei dieser Modellierung können die latenten Variablen verschiedene Fähigkeiten repräsentieren, die bei der Leistung in bestimmten inhaltlichen Bereichen gemeinsam erforderlich sind. Für die Definition einer solchen komplexen Ladungsstruktur sind theoretisch fundierte Vorannahmen über Eigenschaften der Aufgaben und die inhaltliche Bedeutung der unterschiedlichen latenten Variablen notwendig. Walker und Beretvas (2003) analysierten zum Beispiel Daten aus einem Test zur Erfassung mathematischer Fähigkeiten mit einem zweidimensionalen IRT-Modell, in dem alle Aufgaben Indikatoren für eine allgemeine mathematische Fähigkeit darstell-

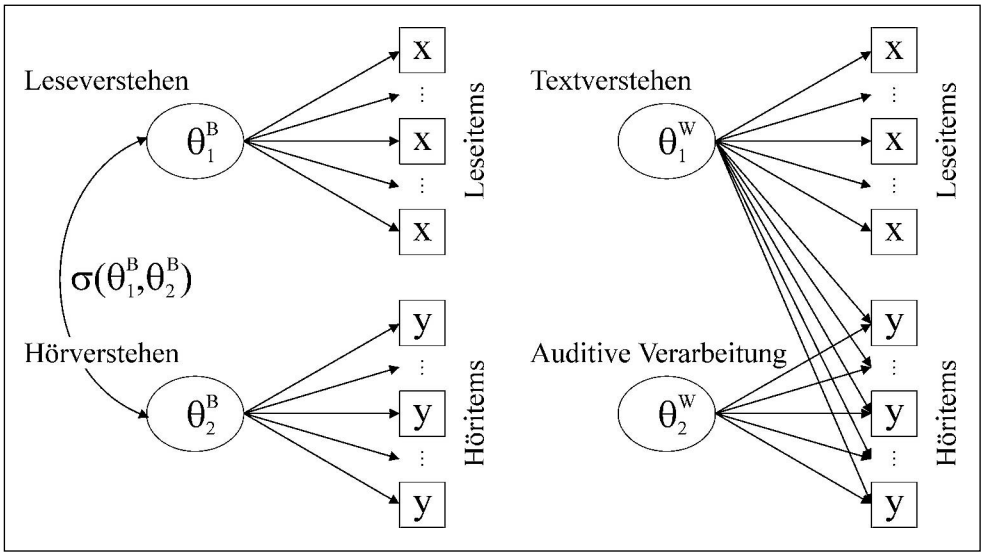


Abb. 1: Schematische Darstellung der Modellierung von Testaufgaben x, y mit (a) einem zweidimensionalen IRT-Modell mit Einfachstruktur und (b) einem zweidimensionalen IRT-Modell mit Mehrfachladungen (Quelle: Hartig/Höhler 2008).

ten. Aufgaben mit offenem Antwortformat wurden zusätzlich als Indikatoren der zweiten Dimension modelliert, die als die Fähigkeit zur Kommunikation mathematischer Inhalte interpretiert wurde. Hartig und Höhler (2008) zeigen am Beispiel von Lese- und Hörverstehensleistungen in Englisch als Fremdsprache, dass ein solches Generalfaktorenmodell in ein Modell mit Einfachstruktur überführbar ist, in dem beide Item-Gruppen auf separaten Dimensionen laden (vgl. Abbildung 1).

Eine Entscheidung zugunsten einer der beiden in Abbildung 1 dargestellten Modellvarianten auf Basis der Anpassung an die empirischen Daten ist nicht möglich, unter bestimmten Umständen können beide Modelle hinsichtlich der Anpassung sogar äquivalent sein (ebd.). Die inhaltliche Bedeutung der latenten Variablen ist jedoch grundsätzlich verschieden. Während die latenten Variablen des Modells mit Einfachstruktur direkt die Teilkompetenzen Lese- und Hörverstehen repräsentieren, ist die Interpretation der latenten Variablen im Modell mit Mehrfachladungen komplizierter. Hier repräsentiert die erste Dimension im Sinne eines Generalfaktors die Teilkompetenzen, die notwendig sind, um die *gemeinsamen Anforderungen* von Lese- und Hörverstehensaufgaben zu bewältigen. Die zweite Dimension steht hingegen für die spezifischen Kompetenzen, die *zusätzlich* für die erfolgreiche Bearbeitung von Hörverstehensaufgaben benötigt werden.

Durch die unterschiedliche Bedeutung der latenten Variablen in diesen beiden Modellierungsansätzen ergeben sich bedeutsame Implikationen für die Interpretation und Kommunikation von Fähigkeitsprofilen. So konnten Hartig und Höhler (ebd.) am Beispiel von Lese- und Hörverstehen in Englisch als Fremdsprache zeigen, dass das Modell

mit Einfachstruktur in nahezu gleichen Geschlechts- und Bildungsgangprofilen für beide Dimensionen resultierte. Dagegen sind in dem Modell mit Mehrfachladungen für die zweite, hörverstehen-spezifische Dimension die Unterschiede zwischen Bildungsgängen weniger stark ausgeprägt als für den für alle Items gemeinsamen Generalfaktor. Darüber hinaus zeigte sich für das Modell mit Mehrfachladungen, dass Mädchen im Generalfaktor zwar höhere Werte aufweisen als Jungen, dieser Effekt sich jedoch bezüglich der Leistung in der hörverstehen-spezifischen Dimension umkehrt; hier zeigen Jungen eine bessere Leistung als Mädchen.

In MIRT-Modellen mit Mehrfachladungen können Annahmen über Interaktionen zwischen den verschiedenen Teilkompetenzen und Aufgabenanforderungen überprüft werden. Solche Modelle implizieren explizite Annahmen über die Fähigkeiten, die benötigt werden, um ein Item lösen zu können und darüber, ob diese Fähigkeiten kompensatorisch verknüpft sind oder nicht (vgl. nächster Abschnitt). Besonders interessant sind Modelle mit Mehrfachladungen für die Kompetenzmodellierung bei komplexen Aufgaben, deren Lösung nicht mit einer Fähigkeitsdimension für jede Aufgabe erklärt werden kann. Auch für Fragestellungen, die sich auf spezifische Teilkompetenzen einer Kompetenzdomäne beziehen, kann dieser Modellierungsansatz eine vielversprechende Alternative darstellen.

Die geringere Komplexität und einfache Interpretation der latenten Variablen in Modellen mit Einfachstruktur kann aber auch vorteilhaft sein: In diesen Modellen stellen die geschätzten Fähigkeitswerte für die latenten Dimensionen direkte Leistungsmaße des zugrunde liegenden Items dar. In vielen Fällen werden die Leistungsmaße verschiedener Teilkompetenzen (z.B. Lese- und Hörverstehen) hoch korreliert sein, da zur erfolgreichen Bearbeitung aller Items auch gemeinsame Fähigkeiten benötigt werden. Diese Gewichtung spezifischer Fähigkeiten und deren genaue Interaktion sind jedoch nicht unbedingt von Interesse. In einem Modell mit Einfachstruktur repräsentieren die latenten Variablen die notwendige Kombination aller Fähigkeiten, die zur erfolgreichen Bearbeitung komplexer Aufgaben benötigt wird. Die Gemeinsamkeiten kommen in der Höhe der Korrelationen zwischen den latenten Variablen zum Ausdruck. Wenn das Forschungsinteresse also hauptsächlich darin besteht, deskriptive Leistungsmaße zu erhalten, unabhängig davon, wie diese miteinander interagieren, sind Modelle mit Einfachstruktur angemessener als solche mit Mehrfachladungen.

3.2 Kompensatorische versus nicht-kompensatorische Verknüpfung von Dimensionen

Bei Modellen mit Mehrfachladungen können die Lösungswahrscheinlichkeiten einzelner Items von mehreren Fähigkeitsdimensionen abhängen. Diese Fähigkeitsdimensionen können unterschiedlich integriert werden. Eine grundsätzliche Frage ist, ob die latenten Variablen kompensatorisch oder nicht-kompensatorisch miteinander verknüpft sind (vgl. Hartig/Höhler 2008). Die meisten MIRT-Modelle mit Mehrfachladungen sind kompensatorisch, eine geringe Fähigkeitsausprägung in einer Dimension kann durch

eine hohe Ausprägung in einer zweiten Dimension ausgeglichen werden und umgekehrt. In einem nicht-kompensatorischen Modell wird für die erfolgreiche Bearbeitung einer komplexen Aufgabe eine hohe Fähigkeitsausprägung in allen Dimensionen benötigt. Der Unterschied zwischen beiden Verknüpfungen kann an zwei einfachen Modellen veranschaulicht werden. Die *Item-Response-Funktion* (IRF) für ein Item im zweidimensionalen Raschmodell mit einer Itemladung auf beiden Dimensionen und einer Itemschwierigkeit von Null ist:

$$\Pr\left(x_{vi}|\theta_1,\theta_2\right)=\operatorname{logit}\left(\theta_1+\theta_2\right) \tag{1}$$

mit

$$\operatorname{logit}\left(y\right)\equiv\frac{e^{y}}{1+e^{y}} \tag{2}$$

In diesem Modell hängt die Lösungswahrscheinlichkeit von der Summe beider Fähigkeitsdimensionen θ_1 und θ_2 ab. Da eine Fähigkeit die andere kompensieren kann, resultiert die gleiche Summe $\theta_1 + \theta_2$ in der gleichen Wahrscheinlichkeit, ein Item richtig zu lösen.

Das *multicomponent latent trait model* (MLTM) von Embretson (1984) ist ein partiell nicht-kompensatorisches MIRT-Modell. Die IRF für ein Item in einer vereinfachten zweidimensionalen Version des MLTM, mit Komponentenschwierigkeiten und Rateparametern von null ist:

$$\Pr\left(x_{vi}|\theta_1,\theta_2\right)=\operatorname{logit}\left(\theta_1\right)\cdot\operatorname{logit}\left(\theta_2\right) \tag{3}$$

In dem nicht-kompensatorischen MLTM hängt die Lösungswahrscheinlichkeit von dem Produkt der Logits beider Fähigkeitsdimensionen θ_1 und θ_2 ab. Somit ist die Lösungswahrscheinlichkeit eines Items nur dann hoch, wenn alle dafür benötigten Fähigkeiten hoch ausgeprägt sind. In Abbildung 2 sind die IRFs für das kompensatorische Modell (vgl. Gleichung 1) und das nicht-kompensatorische Modell (vgl. Gleichung 2) graphisch dargestellt.

Durch die graphische Veranschaulichung wird deutlich, dass das nicht-kompensatorische Modell verglichen mit dem kompensatorischen Modell in allgemein geringer ausgeprägten Lösungswahrscheinlichkeiten resultiert. In dem nicht-kompensatorischen Modell sind die Anforderungen an die Fähigkeiten der getesteten Person höher. Die IRFs beider Modelle weichen besonders stark in Bereichen voneinander ab, in denen sich beide Dimensionen θ_1 und θ_2 unterscheiden: Wenn eine der beiden Dimensionen gering ausgeprägt ist, geht die Lösungswahrscheinlichkeit im nicht-kompensatorischen Modell gegen Null. Im kompensatorischen Modell kann die Lösungswahrscheinlichkeit dagegen substanziell von Null abweichen, wenn die Fähigkeit in einer Dimension ge-

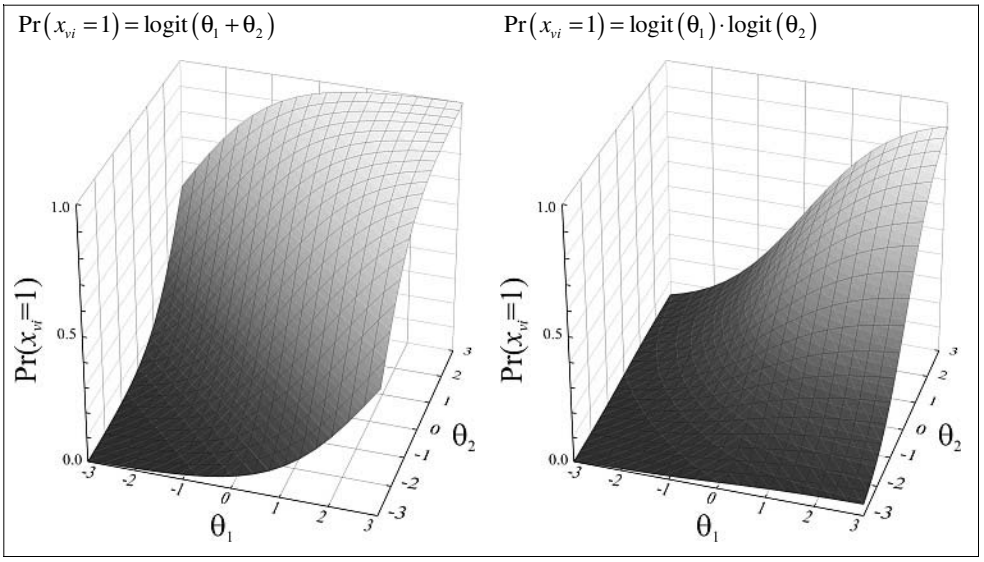


Abb. 2: IRFs für ein zweidimensionales kompensatorisches (links) und ein zweidimensionales nicht-kompensatorisches (rechts) Modell (Quelle: Hartig/Höhler eingereicht)

ring, aber in der anderen Dimension hoch ausgeprägt ist. Für die Bereiche, in denen θ_1 und θ_2 etwa gleich ausgeprägt sind, unterscheiden sich die beiden Modelle kaum.

Die Wahl einer kompensatorischen oder nicht-kompensatorischen Funktion für die Verknüpfung mehrerer latenter Variablen hat entscheidende Implikationen für die Bedeutung dieser Variablen. Eine kompensatorische Verknüpfung kann beispielsweise adäquat sein, wenn die latenten Variablen unterschiedliche Lösungsstrategien für die erfolgreiche Bearbeitung der Testaufgaben repräsentieren. Die Lösungswahrscheinlichkeit sollte hoch sein, wenn irgendeine Strategie ausreichend gut beherrscht wird und noch höher, wenn mehrere Lösungsstrategien zur Verfügung stehen.

Eine nicht-kompensatorische IRF kann dagegen angemessen sein, wenn die latenten Dimensionen Fähigkeiten repräsentieren, die unterschiedlichen (kognitiven) Operationen zugrunde liegen. Die grundlegende Annahme ist hier, dass für eine erfolgreiche Itembearbeitung alle dieser Operationen durchgeführt werden müssen (z.B. bei aufeinander aufbauenden Schritten eines Lösungsprozesses). Wenn eine dieser Fähigkeiten gering ausgeprägt ist, ist die erfolgreiche Bearbeitung eines entsprechenden Items stark erschwert, unabhängig von der Fähigkeitsausprägung in den anderen Dimensionen. Eine hohe Lösungswahrscheinlichkeit ergibt sich erst, wenn alle benötigten Fähigkeiten hoch ausgeprägt sind.

Die Anzahl der latenten Dimensionen hat unterschiedliche Implikationen für die Interpretation der Werte in θ. für die verschiedenen Modelle. In nicht-kompensatorischen Modellen mit Mehrfachladungen verändert sich die Skalierung der Fähigkeitsschätzungen, wenn weitere Dimensionen in das Modell mit aufgenommen werden. In kompen-

satorischen Modellen ist die Skalierung der einzelnen Dimensionen dagegen unabhängig von zusätzlichen Dimensionen (vgl. Reckase 2007). Trotz der mathematischen und theoretischen Unterschiedlichkeit dieser beiden Modellierungsansätze gibt es Hinweise darauf, dass kompensatorische und nicht-kompensatorische Modelle mit der gleichen Anzahl latenter Dimensionen eine sehr ähnliche Anpassung an empirische Daten aufweisen. In Abbildung 2 wird deutlich, dass die IRFs gleich sind für Bereiche, in denen die Dichte einer bivariaten Verteilung bei positiv korrelierten Dimensionen am größten ist (vgl. ebd.; Spray u.a. 1990). Für positiv korrelierte Dimensionen sollten sich daher zwischen den beiden Modellen keine bedeutsamen Unterschiede bezogen auf die empirische Anpassungsgüte ergeben. Deutliche Unterschiede sind hingegen bei negativ korrelierten latenten Variablen zu erwarten.

4. Schlussfolgerungen

Insbesondere für die Modellierung von Kompetenzen stellen MIRT-Modelle eine vielversprechende Methodologie dar. Es existiert bereits eine Bandbreite von Modellen, die für vielfältige Fragestellungen (nicht nur) innerhalb der empirischen Bildungsforschung jeweils passend sein können. Durch vertiefende Analysen mit komplexeren Modellen der IRT können zusätzliche Informationen über die Fähigkeiten der getesteten Individuen gewonnen werden. Aber auch theoretische Annahmen über spezifische (Teil-) Kompetenzen und darüber, wie diese miteinander interagieren, können mit den verschiedenen MIRT-Modellen getestet werden. Schon bei der theoretischen Entwicklung von Kompetenzmodellen sollte die psychometrische Umsetzung bedacht werden. So kann eine theoriegeleitete Itemkonstruktion bei gleichzeitiger Formulierung eines Messmodells gewinnbringend für eine möglichst differenzierte Interpretation und Kommunikation von Schülerkompetenzen eingesetzt werden. Von solchen Erkenntnissen können nicht nur die Schüler/innen, sondern auch Pädagogen/innen, Erzieher/innen und Entscheidungsträger/innen profitieren. Diese Aufgabe erfordert jedoch den möglichst frühen interdisziplinären Austausch zwischen Didaktik und Psychometrie. In der zweiten Förderphase des DFG-Projekts „Modellierung von Kompetenzen mit mehrdimensionalen IRT-Modellen“ ist die Neuentwicklung von Testaufgaben geplant, bei der die mögliche Dimensionalität der Aufgaben schon bei der Itemkonstruktion systematisch berücksichtigt werden soll.

Literatur

Ackerman, T.A./Gierl, M.J./Walker, C.M. (2003): Using multidimensional item response theory to evaluate educational and psychological tests. In: Educational Measurement: Issues and Practice 22, S. 37–53.

Adams, R./Wilson, M./Wang, W.-C. (1997): The multidimensional random coefficients multinomial logit model. In: Applied Psychological Measurement 21, S. 1–32.

Beck, B./Klieme, E. (Hrsg.) (2007): Sprachliche Kompetenzen. Konzepte und Messung. Weinheim/Basel: Beltz.

Blum, W./Neubrand, M./Ehmke, T./Senkbeil, M./Jordan, A./Ulfig, F./Carstensen, C.H. (2004): Mathematische Kompetenz. In: Prenzel, M./Baumert, J./Blum, W./Lehmann, R./Leutner, D./ Neubrand, M./Pekrun, R./Rolff, H.-G./Rost, J./Schiefele, U. (Hrsg.): PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs. Münster: Waxmann, S. 47–92.

Borsboom, D./Mellenbergh, G.J./van Heerden, J. (2003): The theoretical status of latent variables. In: Psychological Review 110, S. 203–219.

Embretson, S.E. (1983): Construct validity: Construct representation versus nomothetic span. In: Psychological Bulletin 93, S. 179–197.

Embretson, S.E. (1984): A general latent trait model for response processes. In: Psychometrika 49, S. 175–186.

Gonzalez, E.J. (2003): Scaling the PIRLS reading assessment data. In: Martin, M.O./Mullis, I.V.S./Kennedy, A.M. (Hrsg.): PIRLS 2001 technical report. Chestnut Hill, MA: Boston College.

Gonzalez, E.J./Galia, J./Li, I. (2004): Scaling methods and procedures for the TIMSS 2003 mathematics and science scales. In: Martin, M.O./Mullis, I.V.S./Chrostowski S.J. (Hrsg.): TIMSS 2003 Technical Report. International Association for the Evaluation of Educational Achievement.

Hartig, J. (2007): Skalierung und Definition von Kompetenzniveaus. In: Beck, B./Klieme, E. (Hrsg.): Sprachliche Kompetenzen. Konzepte und Messung. DESI-Ergebnisse Band 1. Weinheim/Basel: Beltz, S. 83–99.

Hartig, J. (2008): Psychometric models for the assessment of competencies. In: Hartig, J./Klieme, E./Leutner, D. (Hrsg.): Assessment of competencies in educational contexts: state of the art and future prospects. Göttingen: Hogrefe & Huber Publishers, S. 69–90.

Hartig, J./Höhler, J. (eingereicht): Multidimensional IRT models for the assessment of competencies. In: Studies in Educational Evaluation.

Hartig, J./Höhler, J. (2008): Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. In: Zeitschrift für Psychologie/Journal of Psychology 216, S. 89–101.

Hartig, J./Jude, N./Wagner, W. (2008): Methodische Grundlagen der Messung sprachlicher Kompetenzen. In: Klieme, E./Eichler, W./Lehmann, R.H./Nold, G./Schröder, K./Thomé, G./Willenberg, H. (Hrsg.): Sprachliche Kompetenzen. Leistungsverteilungen und Bedingungsfaktoren. DESI-Ergebnisse Band 2. Weinheim: Beltz Pädagogik, S. 34–54.

Hartig, J./Klieme, E. (2006): Kompetenz und Kompetenzdiagnostik. In: Schweizer, K. (Hrsg.): Leistung und Leistungsdiagnostik. Berlin: Springer, S. 127–143.

Klieme, E./Eichler, W./Helmke, A./Lehmann, R.H./Nold, G./Rolff, H.-G. u.a. (2008): Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie. Weinheim: Beltz.

McDonald, R.P. (1997): Normal-ogive multidimensional model. In: van der Linden, W.J./Hambleton, R.K. (Hrsg.): Handbook of modern item response theory. New York City, NY: Springer-Verlag, S. 257–269.

Mislevy, R.J./Wilson, M./Ercikan, K./Chudowsky, N. (2002): Psychometric principles in student assessment. CSE technical report 583. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Muthén, B. (2002): Beyond SEM: general latent variable modeling. In: Behaviormetrika 29, S. 81–117.

Rabe-Hesketh, S./Skrondal, A./Pickles, A. (2004): Generalized multilevel structural equation modelling. In: Psychometrika 69, S. 167–190.

Reckase, M.D. (2007): Multidimensional item response theory. In: Sinharay, S./Rao, C.R. (Hrsg.): Handbook of statistics, Vol. 26: Psychometrics. Amsterdam: Elsevier, S. 607–642.

Reckase, M.D. (2009): Multidimensional item response theory. New York: Springer Verlag.

Skrondal, A./Rabe-Hesketh, S. (2004): Generalized latent variable modeling. Multilevel, longitudinal and structural equation models. Boca Raton u.a.: Chapman & Hall.
Spray, J.A./Davey, T.C./Reckase, M.D./Ackerman, T.A./Carlson, J.E. (1990): Comparison of two logistic multidimensional item response theory models. Research Report ONR90-8. ACT Inc., Iowa City, IA.
Walker, C.M./Beretvas, S.N. (2003): Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. In: Journal of Educational Measurement 40, S. 255–275.
Weinert, F.E. (2001): Concept of competence: a conceptual clarification. In: Rychen, D.S./Salganik, L.H. (Hrsg.): Defining and selecting key competencies. Seattle: Hogrefe & Huber Publishers, S. 45–65.
Wilson, M. (2005): Constructing measures. An item response modelling approach. Mawah: Lawrence Erlbaum Associates.

Anschrift des Autors/der Autorin

Prof. Dr. Johannes Hartig, Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Bildungsqualität und Evaluation, Schloßstr. 29, D-60486 Frankfurt a.M.
E-Mail: hartig@dipf.de

Jana Höhler, Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Bildungsqualität und Evaluation, Schloßstraße 29, D-60486 Frankfurt a.M.
E-Mail: hoehler@dipf.de

Albert Bremerich-Vos

Modellierung von Aspekten sprachlich-kultureller Kompetenz

Anmerkungen zu den Projektberichten

Sichtet man die Studien zu den produktiven und rezeptiven Domänen sprachlicher Kompetenz, dann fällt auf, dass Struktur und Stufung der Lesekompetenz bislang empirisch am gründlichsten untersucht worden sind. Insofern überrascht es nicht, dass alle hier zu würdigenden Projekte mit Aspekten der Lesekompetenz zu tun haben.

Im Projekt *Standardsetting* geht es um eine zentrale Frage der Kompetenzmessung: Wo auf einer kontinuierlichen Leistungstestskala sollen Schwellen bzw. „Cut-Off-Scores" fixiert werden, sodass inhaltlich und psychometrisch plausible Beschreibungen von „Kompetenzstufen" bzw. „-niveaus" daraus resultieren? In den meisten US-amerikanischen Staaten werden Varianten des Bookmark-Verfahrens praktiziert, das im Bericht des Projekts detailliert beschrieben wird. Hier müssen ExpertInnen an wenigen Stellen in einem Buch, das die nach ihren Schwierigkeiten geordneten Items enthält, Markierungen anbringen. Im Projekt interessieren vor allem Aspekte der Validität des Verfahrens. Zwei Teilergebnisse liegen vor: Was die interne Validität betrifft, so urteilen homogene, nur aus Lehrkräften bestehende Panels im Großen und Ganzen „milder" als heterogene Panels, jedenfalls dann, wenn ihre Urteile im dritten Durchgang gemittelt werden. Bei der modifizierten Bookmarkmethode, die vom Berkeley Evaluation and Assessment Research Center entwickelt wurde, tritt dieser Effekt der Gruppenzusammensetzung nicht auf. Hier wird am Ende nicht gemittelt, sondern im „Konsens" entschieden.

M.E. sollte detailliert darüber informiert werden, wie dieser Konsensbildungsprozess organisiert worden ist bzw. mit welchen Mitteln versucht wurde, Attributen eines „herrschaftsfreien Diskurses" gerecht zu werden. Es könnte ja sein, dass der Panelkompositionseffekt nicht aufgrund der besseren Argumente verschwand, sondern deshalb, weil sich eine „Fraktion" mit Macht durchsetzte.

Was die externe Validität angeht, so stimmen die Expertenurteile und die Urteile der unterrichtenden Lehrpersonen in ca. 40% der Fälle überein. In ca. 80% der Fälle beträgt die Differenz eine Stufe. Für die Bewertung dieses Ergebnisses sind Informationen nötig, die im Zwischenbericht (noch) nicht enthalten sind. Das Projekt stützt sich auf Messicks sehr weites Verständnis von „Validität", wonach auch „konsequenzbezogene" Aspekte wie die angemessene Reaktion von Lehrkräften und Bildungsadministration auf die Rückmeldung von Ergebnissen kompetenzorientierter Tests unter diesen Begriff fallen. Auch wenn man diese Ansicht nicht teilt, muss man unterstreichen, dass die politisch-normative Dimension der Fixierung von Cut-off-Scores nicht getilgt werden kann. Dieser Sachverhalt wird nach meinem Eindruck im fachdidaktischen Diskurs häufig verkannt.

Das Projekt konnte sich auf den bewährten Gemeinsamen Europäischen Referenzrahmen für Sprachen beziehen. Insofern lagen die Deskriptoren für die einzelnen Niveaus bzw. Stufen bereits vor. Zu einem Bookmarkverfahren kann aber auch gehören, dass die ExpertInnen die Deskriptoren selbst definieren. Es ist zu wünschen, dass derart angereicherte Standard-Setting-Verfahren zukünftig auch in Bereichen praktiziert werden, in denen die Entwicklung von Kompetenzstufenmodellen noch in den Kinderschuhen steckt.

Herkömmliche Tests (nicht nur) der Lesekompetenz sind Statustests, bei denen es um die Ermittlung der Fähigkeitsausprägung zu einem bestimmten Zeitpunkt geht. Demgegenüber setzt ein anderes Projekt auf das *Dynamische Testen*. Es interessiert nicht nur die aktuelle Leistung, sondern es geht primär darum, wie SchülerInnen im Laufe eines Lernprozesses verschiedene Formen von Feedbacks verarbeiten, also Lernpotenziale aktivieren. So werden, was gerade didaktisch besonders bedeutsam ist, Aussagen über lernerspezifische Zonen der aktuellen und der nächsten Entwicklung möglich. Die Feedbacks zielen in erster Linie auf intendierte („textbasierte") und elaborative („vorwissensbasierte") Inferenzen, die für die Herstellung lokaler und globaler Textkohärenz nötig sind, sowie auf kognitive und metakognitive Strategien. Es wird allerdings nur ein Beispiel vorgestellt. Üblicherweise werden einfache und komplexe Formen der Rückmeldung unterschieden. Zwei einfache Versionen: Die Lernenden erfahren, ob ihre Antwort richtig oder falsch war, bei einer falschen Antwort wird die korrekte Antwort aber nicht genannt. Oder es wird nach der Antwort in jedem Fall die richtige Lösung mitgeteilt. Elaborierter sind Feedbacks, wenn über die Falsch- (bzw. auch Richtig-)Mitteilung hinausgegangen wird, wenn z.B. auf neue Beispiele, Analogien und eben auch auf Strategien hingewiesen wird. Die erste einfache Version ist offenbar nicht lernwirksam, die zweite wohl. Was elaborierte Formen des Feedbacks angeht, so verweisen die AutorInnen auf eine Metaanalyse von Bangert-Drowns u.a. (1991), wonach eine durchschnittliche Effektstärke von $d = .53$ resultierte. Es gibt aber auch eine Reihe von Studien, aus denen hervorgeht, dass ausführliches, auf tieferes Verstehen zielendes Feedback nicht zu besseren Leistungen führte als die Rückmeldung der richtigen Antwort (Dempsey/Sales 1993), insbesondere dann nicht, wenn die Fehler eher basaler Natur waren. Darüber hinaus sind die Operationalisierungen des elaborierten Feedbacks offenbar recht heterogen. Zu bedenken wären auch motivationale Aspekte vor allem von negativem Feedback.

Im Zwischenbericht sind die Konturen des Projekts erkennbar. Wie auch immer die Exempel der verschiedenen Feedbackarten im Einzelnen aussehen werden; wie dabei mehr oder weniger komplexe Inferenzen und kognitive und metakognitive Strategien berücksichtigt werden; oder auf welche Art und Weise der Leistungszuwachs bestimmt wird – das Projekt ist in didaktischer Hinsicht nicht zuletzt deshalb bedeutsam, weil es dazu beitragen kann, die insbesondere im deutschdidaktischen Diskurs oft bemühte Differenz von – überspitzt gesagt – eher „guten" Lern- und eher „bösen" Testaufgaben zu mindern.

Unter Bezug auf Überlegungen von Umberto Eco untersucht ein weiteres Projekt die Struktur einer *literarästhetischen Urteilskompetenz*. Dabei werden drei Dimensionen dieser Kompetenz unterschieden: eine semantische, eine idiolektale und eine kontextuelle. Im Rahmen von Aufgaben, bei denen es um semantische (Teil-) Kompetenz geht, müssen die ProbandInnen zeigen, dass sie Textsinn konstruieren können. Unter einem Idiolekt versteht man in der Regel den für eine/n bestimmte/n Sprecherin/Sprecher charakteristischen Sprachgebrauch. Eco spricht aber von einem „Text-Idiolekt" bzw. von „Textstrategien". Es handelt sich um Struktureigenheiten von Texten, die intersubjektiv zugänglich sind. Dabei kann es sich um konventionelle Stilmittel wie „Es war einmal", um rhetorische Mittel wie Metaphern usw. handeln. Üblicherweise spricht man – wie auch im Projekt – hier von „formalen" Aspekten.

Die kontextuelle Kompetenz schließlich macht aus, dass man literarische Texte auf der Folie von historischem Wissen (über Epochen, Gattungen, Motive, Autoren usw.) verstehen kann.

Weil zu befürchten ist, dass kontextuelle Kompetenz im angedeuteten Sinn nur bei einigen SchülerInnen vorausgesetzt werden kann, leuchtet ein, dass man aus Gründen der Testfairness Kontextinformationen in Form weiterer Texte präsentiert. Diese Texte beziehen sich, so die AutorInnen, einmal primär auf „inhaltliche", einmal mehr auf „formale" Aspekte der literarischen Texte.

Die Autorengruppe favorisiert ein zweidimensionales Modell der literarästhetischen Urteilskompetenz (LUK). Ob das bei einer Korrelation auf latenter Ebene von $r = .92$ sinnvoll ist, sei dahingestellt. Was die kriteriale Validität angeht, so ergibt sich u.a. eine unkorrigierte Korrelation mit Lesekompetenz, erhoben in Form von einigen Items zu kontinuierlichen Sachtexten, von .59. Diese Korrelation ist fast identisch mit der, die Artelt und Schlagmüller (2004, S. 178) im PISA-Kontext für die Skalen für literarische und kontinuierliche Texte berichten; sie beträgt .55. Insofern drängt sich die Frage auf, inwiefern im Projekt tatsächlich eine spezifisch literarästhetische Kompetenz erhoben wird. Das Vorhaben der Gruppe zu untersuchen, „inwieweit sich LUK auch dann von allgemeiner Lesekompetenz abgrenzen lässt, wenn ihre Operationalisierung auch literarische Texte beinhaltet", erscheint insofern als besonders dringlich. Intuitiv wenig einsichtig ist m.E. darüber hinaus z.B. der Sachverhalt, dass die Mathematiknote mit der Lesekompetenz im Allgemeinen mit −.09 korreliert, mit LUK aber mit −.20 (negativ, insofern die kleinere Note die bessere ist). Im Mathematikunterricht spielen doch vor allem ästhetikferne Sachtexte eine Rolle.

Für eine detaillierte Beurteilung des Projekts dürfte vor allem eine Inspektion der Aufgaben zum „Text-Idiolekt" bzw. zu „formalen" Merkmalen literarischer Texte nötig sein. Dabei wäre vor allem zu prüfen, inwiefern sich diese Aufgaben z.B. von den auf literarische Stimuli bezogenen PISA-Aufgaben unterscheiden.

Im Projekt BITE untersuchen die AutorInnen die Fähigkeit von SchülerInnen und Lehrpersonen, Texte und bildhafte Darstellungen integrativ zu verarbeiten, wobei vor allem auch interessiert, inwiefern die Lehrerkompetenz über unterrichtliches Handeln die Schülerkompetenz und -motivation beeinflusst. Sie stützen sich auf ein Modell von

Schnotz und Bannert und unterscheiden – illustriert anhand eines Exempels aus dem Biologieunterricht – im Hinblick auf die Text-Bild-Integration drei Anforderungsniveaus: das „Ablesen" von Detailinformationen, die Angabe einfacher Relationen und die Nennung komplexer Relationen. Die modellkonforme Konstruktion der Aufgaben beruht auf einer umfangreichen Lehrwerksanalyse. Insofern dürften die Aufgaben unterrichtsvalide sein. Auffällig ist allerdings, dass im Modell nicht auch auf Reflektieren und Bewerten abgehoben wird. Dabei könnte es doch auch darum gehen, dass die SchülerInnen ansatzweise lernen (sollten), Aspekte der „Rhetorik" bildhafter Darstellungen zu durchschauen.

Das eigentliche Projekt ist längsschnittlich angelegt, ausgewertet wurde bislang eine querschnittliche Pilotierungsstudie. Zentral sind zwei Befunde: HauptschülerInnen in achten Klassen erreichen im Mittel nicht das Kompetenzniveau von GymnasiastInnen in fünften Klassen, und Bild-Text-Kombinationen sind im Unterricht offensichtlich nicht expliziter Lehr- bzw. Lerngegenstand.

Für die nähere Zukunft hat sich die Projektgruppe u.a. eine rationale Aufgabenanalyse vorgenommen, d.h. es sollen für die Lösung der einzelnen Aufgaben spezifische kognitive Prozesse bestimmt und auf dieser Basis Itemschwierigkeiten vorausgesagt werden. In diesem Kontext könnten m.E. von der Gruppe nicht genannte Arbeiten Mosenthals und Kirschs hilfreich sein. Sie haben nicht nur eine sowohl auf kontinuierliche Texte als auch auf diskontinuierliche „Dokumente" bezogene Theorie vorgelegt, wonach die Aufgabenschwierigkeit wesentlich vom Typ der gefragten Information (z.B. eher „konkret" oder eher „abstrakt"), vom Zuordnungstyp (z.B. Lokalisieren) und von der Plausibilität von Distraktoren abhängt. Sie haben auch ein Modell für die Messung der Komplexität von „Dokumenten" wie Karten, Diagrammen, Tabellen usw. vorgelegt (vgl. Mosenthal/Kirsch 1998). Die Ergebnisse der Studie des Projekts werden nicht zuletzt für die erste Phase der Lehrerausbildung relevant sein, spielt doch die Fähigkeit, Bild-Text-Kombinationen zu verstehen, in allen Fächern eine große Rolle.

Im Projekt *MIRT* geht es am Beispiel von DESI-Aufgaben zur Sprachbewusstheit und zum Lese- und Hörverstehen vor allem um die Frage, ob mehrdimensionale Modelle diagnostisch informativer sind als eindimensionale. Sind einzelne Items mit mehr als einer Kompetenz assoziiert (Modell mit Mehrfachladungen), z.B. mit zwei (Teil-) Kompetenzen, kann der Fall eintreten, dass man die erste latente Variable im Sinne eines Generalfaktors und die zweite als zusätzliche Teilkompetenz zu verstehen hat. So verhält es sich bei den DESI-Lese- und Hörverstehensaufgaben: Die erste Variable bezeichnet die für die Bewältigung von Anforderungen nötige (Teil-) Kompetenz, die Lese- und Hörverstehensaufgaben gemeinsam auszeichnen. Die zweite Variable repräsentiert dann die zusätzliche Teilkompetenz, die für Hörverstehen spezifisch ist. Ob man mehrdimensionale Modelle mit Einfach- oder Mehrfachladung wählt, kann wie bei DESI diagnostisch relevante Implikationen haben. Bei einem Modell mit Mehrfachladung sind die Unterschiede zwischen Bildungsgängen in der nur auf das Hörverstehen bezogenen Dimension weniger markant als bei einem Modell mit Einfachladung, und in dieser Dimension schneiden Jungen sogar besser ab als Mädchen. Kann man daraus also u.a.

schließen, dass Jungen vor allem mit Hilfe von Aufgaben gefördert werden sollten, die nicht speziell auf das Hörverstehen zielen? Sollten sie z.B. üben, im Umgang mit Gehörtem *und* Gelesenem mehr oder weniger komplexe Schlüsse zu ziehen?

Die AutorInnen unterscheiden darüber hinaus kompensatorische und nicht-kompensatorische Varianten von Modellen mit Mehrfachladungen. Im ersten Fall kann eine gering ausgeprägte Teilfähigkeit durch eine andere kompensiert werden, im zweiten müssen erhebliche Fähigkeiten in allen Dimensionen vorliegen. M.E. dürfte es nicht einfach sein, bei Kompetenzen im sprachlichen Bereich Beispiele für die kompensatorische Verknüpfung von latenten Variablen zu finden.

Hartig und Höhler betonen selbst, dass für die üblichen mit Testungen verfolgten Ziele Modelle mit Einfachstruktur in der Regel angemessen seien. Insofern drängt sich der Eindruck auf, dass der – auf abstrakter Ebene ja sehr einleuchtende – diagnostische Zusatznutzen einer komplexen mehrdimensionalen Modellierung zurzeit noch nicht auf Anhieb erkennbar ist.

Ein Zwischenbericht kann die Neugier der Lesenden nur teilweise stillen. Dennoch lässt sich resümieren, dass alle Projekte auf zentrale Aspekte sprachlicher Kompetenz zielen und dass allen Vorhaben methodische Standards zugrunde liegen, die z.B. der Deutschdidaktik noch weitgehend fremd sind. Insofern ist zu hoffen, dass sie einige Strahlkraft entfalten.

Literatur

Artelt, C./Schlagmüller, M. (2004): Der Umgang mit literarischen Texten als Teilkompetenz im Lesen? Dimensionsanalysen und Ländervergleiche. In: Schiefele, U./Artelt, C./Schneider, W./Stanat, P. (Hrsg.): Struktur, Entwicklung und Förderung von Lesekompetenz – Vertiefende Analysen im Rahmen von PISA 2000. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 169–196.

Bangert-Drowns, R.S./Kulik, C.C./Kulik, J.A./Morgan, A. (1991): The instructional effect of feedback in test-like events. In: Review of Educational Research 61, S. 213–238.

Dempsey, J./Sales, G. (1993) (Hrsg.): Interactive instruction and feedback. Englewood, New Jersey.

Mosenthal, P.B./Kirsch, I.S. (1998): A new measure for assessing document complexity: The PMOSE/IKIRSCH document readability formula. In: Journal of Adolescent and Adult Literacy 41, S. 638–657.

Anschrift des Autors

Prof. Dr. Albert Bremerich-Vos, Germanistik/Linguistik/Sprachdidaktik, Fakultät für Geisteswissenschaften, Universität Duisburg-Essen, Universitätsstr. 12, D-45117 Essen
E-Mail: Albert.Bremerich-Vos@uni-due.de

Ellen Gausmann/Sabina Eggert/Marcus Hasselhorn/Rainer Watermann/ Susanne Bögeholz

Wie verarbeiten Schüler/innen Sachinformationen in Problem- und Entscheidungssituationen Nachhaltiger Entwicklung?

Ein Beitrag zur Bewertungskompetenz

Projekt Bewertungskompetenz[1]

Problem- und Entscheidungssituationen für eine nachhaltige bzw. zukunftsfähige Gestaltung unseres Planeten sind nicht nur faktisch, sondern auch gesellschaftlich und damit ethisch komplex. Darüber hinaus weisen sie in der Regel mehrere gleichwertige Handlungsoptionen auf (vgl. Eggert/Bögeholz 2006). Für Schüler/innen besteht die Herausforderung, für beispielsweise anthropogen bedingte Umweltprobleme – und damit für Entscheidungssituationen Nachhaltiger Entwicklung – Lösungsvorschläge bzw. Handlungsoptionen zu entwickeln und diese reflektieren zu können. Dazu benötigen sie Bewertungskompetenz. Bewertungskompetenz erlaubt es, in komplexen Problem- und Entscheidungssituationen geeignete Handlungsoptionen zu entwickeln, diese miteinander vergleichen zu können und auf Basis von Werten und Normen tragfähige und zukunftsfähige Entscheidungen treffen und reflektieren zu können (vgl. u.a. ebd.; Bögeholz 2007). Ziel ist es, die Umwelt nachhaltig mitzugestalten. Nachhaltiges Mitgestalten der Umwelt ist ein zentrales Anliegen unserer Gesellschaft, die sich dem Leitbild der Nachhaltigen Entwicklung verpflichtet hat (vgl. De Haan u.a. 2008). Das Leitbild fordert, dass wir bei unseren Entscheidungen und Handlungen mit bedenken, dass wir andere, jetzt oder in Zukunft lebende Menschen durch unser Handeln nicht darin beeinträchtigen, ihre (Grund-)Bedürfnisse zu befriedigen (vgl. WCED 1987). Weiterhin ist für nachhaltiges Handeln der Einbezug und die Verarbeitung von Sachinformationen zu allen drei Sphären Nachhaltiger Entwicklung, d.h. zu Ökologie, Ökonomie und

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: BO 1730/3-2) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

Sozialem und deren Zusammenwirken zentral (vgl. Scott/Gough 2003). Informationssuch- und -verarbeitungsprozesse sind damit wichtige Bestandteile von Bewertungskompetenz, einer zentralen Säule der Bildungsstandards für den Unterricht in den naturwissenschaftlichen Fächern.

Ziel des Artikels ist die Herleitung eines Rahmenkonzeptes der Bewertungskompetenz als Basis für die Modellierung der Teilkompetenz „Generieren und Reflektieren von Sachinformationen". In diesem Artikel geht es primär um eine theoretische Fundierung und damit verbunden um einen ersten empirischen Zugang zur Erfassung dieser Teilkompetenz. Konkret zeigt der vorliegende Artikel auf, wie offene Aufgaben angelegt sein sollten, um herauszufinden, ob Schüler/innen a) die Verbindungen von Ökologie, Ökonomie und Sozialem erkennen, b) Handlungsoptionen auf dieser Basis entwickeln und c) Handlungsoptionen im Hinblick auf ihre Tragfähigkeit beurteilen. Er stellt damit ein Rahmenkonzept für die Aufgabenentwicklung und Aufgabenanalyse für eine zukünftige Modellierung und Testentwicklung für die Teilkompetenz „Generieren und Reflektieren von Sachinformationen" bereit.

1. Theoretischer Ansatz und Fragestellung

Um Bewertungskompetenz empirisch messbar zu machen, bedarf es zunächst eines theoretischen Modells, welches zentrale Teilkompetenzen sowie mögliche Graduierungen im Sinne von Kompetenzniveaus beschreibt. Für Bewertungskompetenz im Kontext Nachhaltiger Entwicklung haben Eggert und Bögeholz (2006; Bögeholz 2007) in ihrem Göttinger Modell vier Teilkompetenzen identifiziert. Dies sind zum einen konzeptuelles Wissen über Nachhaltige Entwicklung („Kennen und Verstehen Nachhaltiger Entwicklung") sowie ethisches Basiswissen über Werte und Normen („Kennen und Verstehen von Werten und Normen"). Zum anderen sind für die Lösung von Problem- und Entscheidungssituationen Nachhaltiger Entwicklung prozedurale Kompetenzen wie Informationen suchen und verarbeiten („Generieren und Reflektieren von Sachinformationen") sowie Handlungsoptionen bewerten und Entscheidungen treffen („Bewerten, Entscheiden und Reflektieren") zentral. Letztere sind dabei wichtige Teilaspekte von Entscheidungsfindungsprozessen (vgl. Betsch/Haberstroh 2005).

Für den naturwissenschaftlichen Unterricht bedeutet dies, dass Schüler/innen Problem- und Entscheidungssituationen zunächst als relevant erkennen und beschreiben müssen. Zur Lösung müssen sie tragfähige Handlungsoptionen im Sinne des Leitbilds Nachhaltiger Entwicklung entwickeln. Für eine Beurteilung tragfähiger Handlungsoptionen ist darüber hinaus eine kritische Reflexion über Sachinformationen notwendig. Diese Prozesse werden in der Teilkompetenz „Generieren und Reflektieren von Sachinformationen" zusammengefasst.

Die Handlungsoptionen müssen im Folgeschritt bewertet werden. Dazu gehört das Vergleichen der Vor- und Nachteile unter Anwendung von Entscheidungsstrategien sowie eine kritische Reflexion von Entscheidungsfindungen. Diese Aspekte zeichnen die Teilkompetenz „Bewerten, Entscheiden und Reflektieren" aus. Für alle beschriebenen

Teilkompetenzen wurden theoretisch vier Kompetenzniveaus a priori formuliert. Für „Bewerten, Entscheiden und Reflektieren“ konnten diese Niveaus mit Modifizierungen empirisch bestätigt werden (vgl. Eggert/Bögeholz 2010).

Für die Teilkompetenz „Generieren und Reflektieren von Sachinformationen“ ist auf theoretischer Ebene davon auszugehen, dass qualitative Unterschiede in Form von Kompetenzniveaus durch eine steigende Verbindung von Sachinformationen der drei Sphären Nachhaltiger Entwicklung gekennzeichnet sind. Darüber hinaus sollte eine steigende Kompetenz mit einer qualitativ unterschiedlichen Reflexionsfähigkeit einhergehen. Die Reflexionsfähigkeit drückt sich in der Qualität der Beurteilung der Tragfähigkeit von Handlungsoptionen aus. Ein vergleichbares Graduierungsprinzip wurde für die Verwendung von (natur-)wissenschaftlichen Informationen bei der Bearbeitung von Entscheidungssituationen zu gesellschaftlich relevanten Themen (socioscientific issues) bereits empirisch nachgewiesen (vgl. Roberts/Wilson/Draney 1997; Wilson/Sloane 2000). Die vorliegende Arbeit baut auf den Erkenntnissen des Science Education for Public Understanding Project (SEPUP) auf – kontextualisiert sie jedoch für Gestaltungsaufgaben Nachhaltiger Entwicklung und trägt damit zentralen Herausforderungen des Leitbildes Rechnung (vgl. bspw. WCED 1987; SRU 1994; Scott/Gough 2003). Zentrale Anforderungen für Schüler/innen sind für diesen Kontext aus fachdidaktischer Perspektive

a) die Zusammenhänge zwischen ökologischen, ökonomischen und sozialen Aspekten zu erkennen,
b) Handlungsoptionen zu entwickeln, die geeignet zu nachhaltigen Entwicklungen beitragen, sowie
c) Handlungsoptionen im Hinblick auf ihre Tragfähigkeit als Beitrag zu nachhaltigen Entwicklungen zu beurteilen.

Die Forschungsarbeit in diesem Projekt des Schwerpunktprogramms „Kompetenzmodelle“ zielt derzeit darauf, die Teilkompetenz empirisch zugänglich zu machen und zu überprüfen. Dabei ist als Vorarbeit zunächst zu klären, wie Schüler/innen mit den drei aufgezeigten Anforderungen umgehen. Konkret soll dabei – aufbauend auf den Erfahrungen eines ersten empirischen Zugangs – in dem vorliegenden Artikel ein tragfähiges Rahmenkonzept entwickelt werden, das grundsätzlich eine Modellierung der Teilkompetenz „Generieren und Reflektieren von Sachinformationen“ in künftigen Studien ermöglicht.

2. Methodisches Vorgehen

Ausgehend vom Göttinger Modell werden alle Teilkompetenzen von Bewertungskompetenz zunächst getrennt voneinander operationalisiert. Es wird angestrebt, jede Teilkompetenz eindimensional zu modellieren. Dafür werden jeweils Aufgaben entwickelt, welche in mehreren Schritten überarbeitet und optimiert werden. Eine Erprobung der ersten Aufgaben erfolgte mit je zwei Schüler/innen der Jahrgangsstufen 6, 8, 10 und 12.

2.1 Aufgabenentwicklung

Für die Teilkompetenz „Generieren und Reflektieren von Sachinformationen“ wurden Aufgaben entworfen, die von fachdidaktischen Expert/innen begutachtet wurden. Nach einer Optimierung wurden die Aufgaben in einer qualitativen Studie (N = 8; davon 2 Schülerinnen) erprobt. Zum Einsatz kam die Methode des Lauten Denkens (vgl. Ericsson/Simon 1993).

Als Aufgabenkontexte wurden realweltliche Problem- und Entscheidungssituationen Nachhaltiger Entwicklung ausgewählt, die im Spannungsfeld zwischen Ökologie, Ökonomie und Sozialem stehen. Die Aufgaben lassen sich zwei Aufgabentypen zuordnen: zum einen Aufgaben, bei denen Sachinformationen für die Entwicklung von Handlungsoptionen verarbeitet werden müssen (Aufgabentyp I; Abb. 1), und zum anderen Aufgaben, bei denen Sachinformationen zu einzelnen Handlungsoptionen beurteilt und damit reflektiert werden müssen (Aufgabentyp II; Abb. 1).

Abb. 1: Beispiele zu den Aufgabentypen „Generieren“ und „Reflektieren“ von Sachinformationen (gekürzte Versionen)

Beide Aufgabentypen sind durch mehrere Aspekte und Verbindungen gekennzeichnet, die bei der Bearbeitung berücksichtigt werden sollten (Abb. 2). Die Aufgaben des Typs I lassen sich dadurch beschreiben, dass eine Naturressource durch eine hohe Nachfrage

bedroht wird (ökologisch-ökonomische Verbindung). Die (Über-)Nutzung dieser Naturressource hat positive und negative Auswirkungen auf die beteiligten Personengruppen. In der Regel profitieren einige Personengruppen aufgrund der hohen Nachfrage (ökonomisch-soziale Verbindung), andere werden durch den Rückgang der Naturressource benachteiligt (ökologisch-soziale Verbindung). In der Beispielaufgabe „Königspythons aus Ghana“ werden Königspythons aufgrund der weltweiten hohen Nachfrage wild gefangen und gehen damit dem Ökosystem verloren. Damit wird das dynamische Gleichgewicht zwischen Nagetieren und Königspythons gestört. Dies hat negative Auswirkungen auf die Ernte von Kleinbauern/bäuerinnen. Im Gegensatz zu den verschlechterten Lebensbedingungen der Kleinbauern/bäuerinnen profitieren die Arbeiter/innen auf den Schlangenfarmen durch eine erhöhte Nachfrage an Königspythons.

Bei den Aufgaben des Typs I werden die Schüler/innen zunächst aufgefordert, die dargestellte Problemsituation zu beschreiben. Dabei sind die oben beschriebenen Verbindungen der einzelnen Aspekte zu berücksichtigen. Anschließend sollen die Schüler/innen tragfähige Handlungsoptionen (= Lösungsmöglichkeiten) entwickeln (vgl. Abb. 1).

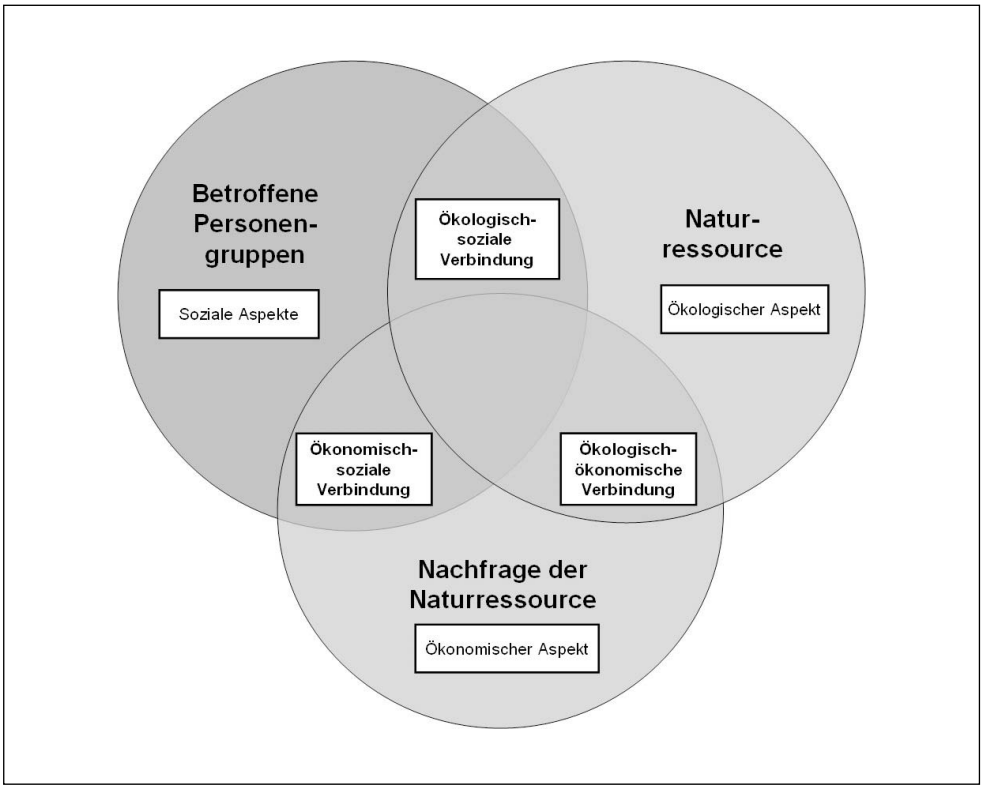


Abb. 2: Relevante Aspekte und Verbindungen der drei Sphären Nachhaltiger Entwicklung in den Aufgabentypen I und II

Bei den Aufgaben des Typs II müssen präsentierte Sachinformationen zu einzelnen Handlungsoptionen kritisch geprüft und in Bezug auf ihre Tragfähigkeit beurteilt werden (vgl. Abb. 1). Die Vorgehensweise folgt damit strukturell dem erfolgreichen Ansatz von Reflexionsaufgaben zur Messung der Teilkompetenz „Bewerten, Entscheiden und Reflektieren“ (vgl. Eggert/Bögeholz 2010). In der Beispielaufgabe „Bananenanbau in Costa Rica“ werden verschiedene Projekte zum Bananenanbau vorgestellt. Die Projekte sind dabei so konstruiert, dass sich Vorteile in einer oder zwei Sphären negativ auf die andere bzw. die anderen Sphären Nachhaltiger Entwicklung auswirken (vgl. Abb. 2). Diese Interdependenz soll von den Schüler/innen erkannt, beschrieben und bei der Beurteilung der Tragfähigkeit der Projekte im Hinblick auf nachhaltige Entwicklungen berücksichtigt werden.

2.2 Auswertung von Schülerantworten

Die Auswertung der Schülerantworten innerhalb der Protokolle Lauten Denkens erfolgte computergestützt in Anlehnung an die Qualitative Inhaltsanalyse nach Mayring (2008). Dabei wurden in einem iterativen Verfahren sowohl deduktiv als auch induktiv aus dem Datenmaterial Kategorien zur Auswertung entwickelt. Die Analyse erfolgte durch zwei Wissenschaftlerinnen unabhängig voneinander. Ergebnis dieser qualitativen Analyse ist ein Kodierleitfaden. Dieser bildet die Grundlage für die Konstruktion quantitativer offener Aufgaben sowie deren Auswertung für die Teilkompetenz „Generieren und Reflektieren von Sachinformationen“.

3. Ergebnisse

Die Analysen in Bezug auf die Bearbeitung des Aufgabentyps I zeigen, dass viele der Schüler/innen in der Lage waren, eine zentrale Verbindung zwischen den Sphären Ökologie, Ökonomie und Sozialem für die Aufgabe „Königspythons aus Ghana“ zu beschreiben (vgl. Abb. 3).

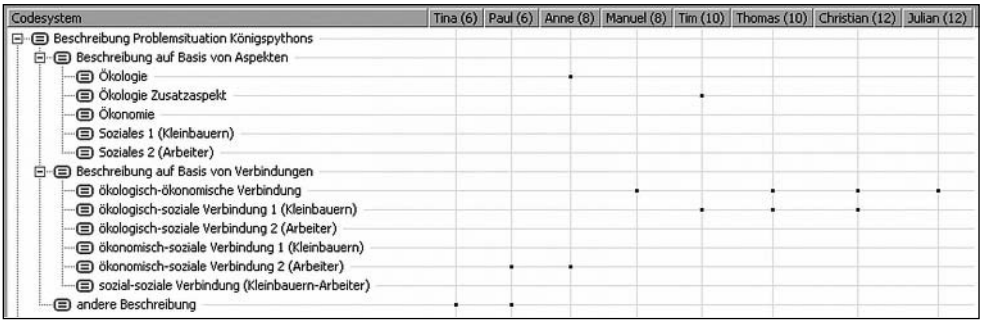


Abb. 3: Auszug aus dem Kodiersystem in Bezug auf die Beschreibung der Problemsituation

Am häufigsten wurden die ökologisch-ökonomische und die ökologisch-soziale Verbindung erläutert. Die ökologisch-ökonomische Verbindung bezieht sich auf den Zusammenhang zwischen Nachfrage und Rückgang der Naturressource, hier am Beispiel von Thomas expliziert:

> *„Ja, das Problem ist wahrscheinlich, dass die Nachfrage stark gestiegen ist und man dadurch mehr Tiere züchten muss, wobei die Weibchen von weiter weg geholt werden und in diesen Gebieten, wo die Weibchen halt herkommen, es halt weniger gibt […]"* (Thomas, 10. Jg.).

Die ökologisch-soziale Verbindung stellt den Zusammenhang zwischen Rückgang der Naturressource und negativen Auswirkungen für eine der beteiligten Personengruppen dar:

> *„Ja. (...)[2]. Ach, zum Problem (.) wäre noch zu sagen, dass die Pythons diese kleinen Nagetiere fressen und diese ja auch (..) dadurch dafür sorgen, dass diese Nagetiere, die Ernte der Menschen nicht vernichten. Das wär halt (..) auch noch ein Problem, wenn die Pythons dann wegfallen. Das auch (.) die Ernte der Leute darunter leidet"* (Christian, 12. Jg.).

Eine sozial-soziale Verbindung, d.h. die Abhängigkeit der beteiligten Personengruppen untereinander (Kleinbauern/bäuerinnen und Arbeiter/innen auf Schlangenfarmen), wurde jedoch von keiner der befragten Personen angesprochen.

Im Anschluss an die Beschreibung der Problemsituation sollten die Schüler/innen tragfähige Handlungsoptionen generieren. Die Ergebnisse zeigten, dass Handlungsoptionen auf Basis einzelner Sphären entwickelt wurden, wobei der ökologische und der ökonomische Aspekt dominierten (vgl. Abb. 4). In Bezug auf Handlungsoptionen, die auf Basis von Verbindungen entwickelt wurden, war die ökologisch-ökonomische Verbindung zentral. Die Beispiele von Julian und Tim zeigen eine derartige Argumentationsweise:

> *„Zweite Lösungsmöglichkeit wäre, dass die USA und Deutschland die Annahme von Königspythons verweigern und eben in eigenen Aufzuchtterrarien, wo die richtigen klimatischen Bedingungen gemacht werden, aufzüchtet und halt nur noch aus diesem Bedarf an die Leute verkauft. Dann ist eben die Anzahl der Königspythons in Deutschland und den USA reduziert. Aber damit muss man halt leben, um die Natur in Afrika zu schützen"* (Julian, 12. Jg.).

> *„Ja, also, man müsste die Pythons, die man gefangen hat, wieder an dem Ort aussetzen, wo man sie (.) gefangen genommen hat. Und (..) ja, man müsste versuchen Weibchen in den Farmen zu halten, die (.) aus Eiern geschlüpft sind. Also nicht alle Jungen ins Ausland zu verkaufen, sondern auch daran zu denken, dass man nicht immer nur wieder Wilde fangen muss, sondern eigene hat"* (Tim, 10. Jg.).

2 Die Punkte in Klammern repräsentieren die unterschiedliche Länge an Sprechpausen.

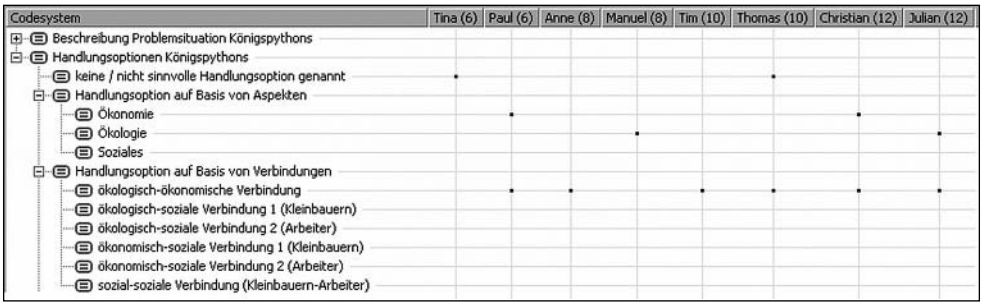


Abb. 4: Auszug aus dem Kodiersystem in Bezug auf die Generierung von Handlungsoptionen

Verbindungen, die die beteiligten Personengruppen einschlossen, wurden in keinem Fall berücksichtigt.

In der Aufgabe des Aufgabentyps II wurden die Schüler/innen dazu aufgefordert, vorgestellte Bananenanbauprojekte (= Handlungsoptionen) in Costa Rica anhand der präsentierten Sachinformationen auf ihre Tragfähigkeit hin zu beurteilen, sowie Ratschläge für eine Verbesserung der Projekte zu geben (vgl. Abb. 5). In Hinblick auf die Tragfähigkeit des Projekts B waren Aussagen zu dem reduzierten Einsatz von Pestiziden/Giften prominent, was sich positiv auf die Arbeitsbedingungen der Arbeiter/innen auswirkt (vgl. Abb. 5: ökologisch-soziale Verbindung). Das Beispiel von Thomas zeigt eine für diese Verbindung typische Argumentationslinie:

> *„Es werden aber halt immer noch Schadstoffe entsorgt, die werden halt gerecht entsorgt und die Plastiksäcke, die blauen werden auch entsorgt. Die Gifte kommen nicht mehr ins Grundwasser, das Meer wird damit geschützt, die Bewohner auf Costa Rica werden damit geschützt, da sie nicht mehr so viele Gifte abbekommen und für die Arbeiter ist es auch besser, dass sie Schutzkleidung haben und so“* (Thomas, 10. Jg.).

Lediglich zwei Schüler/innen erkannten, dass es sich bei Projekt B immer noch um eine Monokultur handelt und somit immer noch Pestizide eingesetzt werden müssen. Am Beispiel von Tim wird deutlich, was das Projekt B nicht lösen kann bzw. inwiefern es modifiziert werden müsste:

> *„Ja also, es ist auf jeden Fall schon mal ein Lösungsansatz, da die meisten der schädlichen Substanzen aufgefangen werden, in kontrollierten Kanälen und die Arbeiter Schutzanzüge bekommen. Jedoch wird das Problem von den (.) Monokulturen noch nicht gelöst. Es werden ja trotzdem nur Bananen angepflanzt. (...) Ja, dass halt noch andere Bäume oder Baumarten (.) zwischen die Bananenstauden gepflanzt werden und somit auch die Monokultur verhindert wird“* (Tim, 10. Jg.)

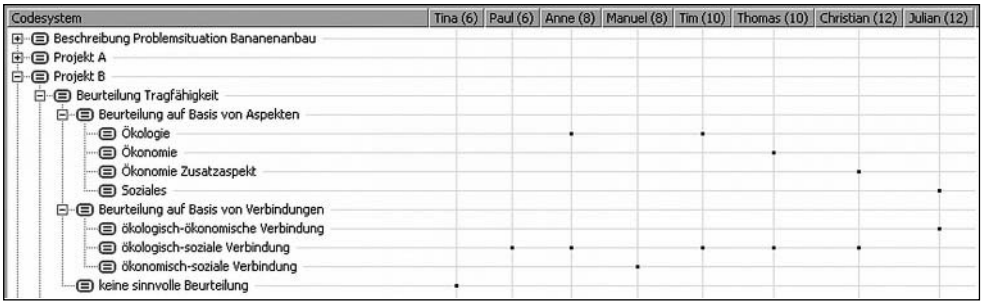


Abb. 5: Auszug aus dem Kodiersystem in Bezug auf das Reflektieren von Sachinformationen

4. Diskussion, Schlussfolgerungen und Ausblick

Am Beispiel der Aufgaben „Königspythons aus Ghana“ und „Bananenanbau in Costa Rica“ wurde illustriert, inwiefern ökologische, ökonomische und soziale Aspekte untereinander in Verbindung gebracht werden. Es konnte gezeigt werden, dass Schüler/innen der Sekundarstufen in Abhängigkeit der präsentierten Aufgaben bestimmte Aspekte Nachhaltiger Entwicklung sowie bestimmte Verbindungen zwischen diesen Aspekten erkannten bzw. bei den von ihnen generierten Handlungsoptionen berücksichtigten. Gleiches konnte auch für die Beurteilung der Tragfähigkeit von präsentierten Handlungsoptionen gezeigt werden. Jedoch dominierte bei den untersuchten Schüler/innen die Berücksichtigung lediglich einer Verbindung zwischen zwei Sphären Nachhaltiger Entwicklung. Inwiefern auch mehrere Verbindungen beim Einsatz optimierter Aufgaben berücksichtigt werden können, ist eine offene Forschungsfrage.

Aufgrund des hier verwendeten „sparsamen“ – aber klassischen und viel zitierten – Rahmenkonzeptes Nachhaltiger Entwicklung (vgl. Scott/Gough 2003; s. Abb. 2), welches die beteiligten Personengruppen zunächst nicht explizit differenziert, war eine Berücksichtigung beispielsweise der sozial-sozialen Verbindungen eher unwahrscheinlich. Damit war es bislang auch schwer, Gerechtigkeitsüberlegungen innerhalb als auch zwischen Generationen angemessen zu berücksichtigen. Dennoch gab die Studie wertvolle Hinweise für die Entwicklung eines tragfähigen Rahmenkonzepts für die Teilkompetenz „Generieren und Reflektieren von Sachinformationen“, das der realweltlichen faktischen Komplexität von Problem- und Entscheidungssituationen Nachhaltiger Entwicklung näher kommt. Die hier vorgestellten Aufgaben und das bei der Aufgabenanalyse resultierende Kodiersystem dienten somit als Ausgangspunkt für weiterführende systematische Überlegungen zur Umgrenzung der Klasse von Anforderungssituationen, die von Schüler/innen bewältigt werden sollen. Abbildung 6 zeigt am Beispiel „Königspythons aus Ghana“ ein Rahmenkonzept auf, das für die künftige Aufgabenkonstruktion und -analyse in quantitativen Studien entwickelt werden konnte. Das Rahmenkonzept hat den Anspruch, dass alle Problem- und Entscheidungssituationen Nachhaltiger Entwicklung vergleichbar aufgearbeitet werden können.

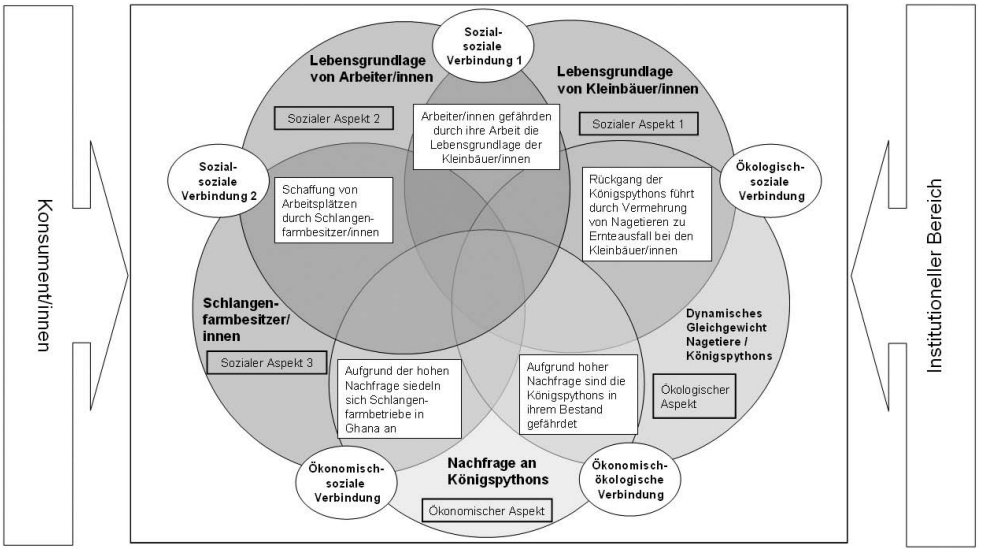


Abb. 6: Rahmenkonzept zur Fundierung der Teilkompetenz „Generieren und Reflektieren von Sachinformationen“

Das hier vorgeschlagene Rahmenkonzept erlaubt damit sowohl die Beschreibung von komplexen Problem- und Entscheidungssituationen, als auch die Verortung von entwickelten Handlungsoptionen sowie die Beurteilung der Tragfähigkeit von präsentierten (nachhaltigen) Handlungsoptionen. Es trägt den zentralen Anforderungen Rechnung, die mit dem Leitbild der Nachhaltigen Entwicklung verbunden sind: Gesamtvernetzung von Ökologie, Ökonomie und Sozialem, Grundbedürfnisorientierung sowie intra- und intergenerationelle Gerechtigkeit (vgl. WCED 1987; SRU 1994; De Haan u.a. 2008). Darüber hinaus wird der Bedeutung des institutionellen Bereichs (u.a. Gesetzgebung; vgl. Scott/Gough 2003) sowie der Konsument/innen („Eigenständiges Handeln“; vgl. De Haan u.a. 2008) Rechnung getragen, die letztlich Einfluss auf das System in seiner Gesamtvernetzung nehmen und (mit-)steuernd zu nachhaltigen Entwicklungen beitragen können.

Das hier präsentierte Konzept beschreibt einen Rahmen, innerhalb dessen sich die Aufgabentypen I und II verorten lassen, doch nicht alle fünf Aspekte und alle fünf Verbindungen müssen zwangsläufig in jeder Aufgabe repräsentiert sein. Damit lassen sich über die Komplexität von Aufgaben Aufgabenschwierigkeiten theoretisch steuern. Aufgaben, die künftig für die Entwicklung eines Messinstrumentes für diese Teilkompetenz zum Einsatz kommen, sollten jedoch neben einem ökologischen, einem ökonomischen und einem sozialen Aspekt sowie der ökologisch-sozialen, der ökologisch-ökonomischen und der ökonomisch-sozialen Verbindung mindestens zwei betroffene Personengruppen als soziale Aspekte einbeziehen. Erst dann können beispielsweise die sozialen Auswirkungen der Problematik der Ablösung traditioneller durch moderne Wirtschaftsformen abgebildet werden (sozial-soziale Verbindung). Die Beschreibung der Problem-

situation erfordert damit das Erkennen von vier zentralen – qualitativ unterschiedlichen – Verbindungen.

Auf der Basis des Rahmenkonzeptes lassen sich vorläufige Überlegungen für eine mögliche Graduierung ausführen (vgl. Tab. 1). Zunächst ist zu vermuten, dass die Beschreibung einer Problem- und Entscheidungssituation bzw. die Generierung und Beurteilung von Handlungsoptionen auf der Basis von Einzelaspekten ein basales Niveau der untersuchten Teilkompetenz darstellt. Normatives Ziel einer Bildung für Nachhaltige Entwicklung muss es sein, die Gesamtvernetzung zu erfassen. Somit stellen Verbindungen zwischen ökologischen, ökonomischen und (qualitativ unterschiedlichen) sozialen Aspekten sicherlich elaboriertere Schülerantworten dar.

Niveau	**Beschreibung**
	Problemsituation wird [1] … bzw. Handlungsoptionen werden [2] …
1	● [1] mit Alltagswissen abgebildet und/oder ● [1] bzw. [2] auf Basis eines bzw. mehrerer Aspekte beschrieben bzw. entwickelt
2	● [1] bzw. [2] auf Basis einer Verbindung beschrieben bzw. entwickelt ● [2] auf Basis einer Verbindung im Hinblick auf ihre Tragfähigkeit beurteilt
3	● [1] bzw. [2] auf Basis von zwei oder drei qualitativ unterschiedlichen Verbindungen beschrieben bzw. entwickelt ● [2] auf Basis von zwei oder drei qualitativ unterschiedlichen Verbindungen im Hinblick auf ihre Tragfähigkeit beurteilt
4	● [1] bzw. [2] auf Basis von vier qualitativ unterschiedlichen Verbindungen beschrieben bzw. entwickelt ● [2] auf Basis von vier qualitativ unterschiedlichen Verbindungen im Hinblick auf ihre Tragfähigkeit beurteilt

Tab. 1: Graduierungsüberlegungen für die Teilkompetenz „Generieren und Reflektieren von Sachinformationen“

Das in diesem Beitrag entwickelte Rahmenkonzept bildet ein Referenzsystem für eine systematische Entwicklung von Diagnose- *und* Lernaufgaben für die Teilkompetenz „Generieren und Reflektieren von Sachinformationen“. Für die Kompetenzmodellierung bildet es die Basis für die Entwicklung eines quantitativen Messinstruments. Diesbezügliche Erhebungen werden derzeit durchgeführt. Ziel ist es, nach der Entwicklung des Messinstruments zu explorieren, inwieweit die in Tabelle 1 dargestellten Niveaustufen sich in der Realität zeigen lassen. Des Weiteren sollen die Zusammenhänge zu den anderen Teilkompetenzen von Bewertungskompetenz untersucht sowie experimentelle Validierungen der fokussierten Kompetenz beispielsweise in Abgrenzung zur Problemlösekompetenz vorgenommen werden. Das Projekt als Ganzes liefert Grundlagenwissen für eine kompetenzorientierte Gestaltung von Bewertungsunterricht zum Kontext Nachhaltiger Entwicklung zur Umsetzung der nationalen Bildungsstandards im Fach Biologie.

Literatur

Betsch, T./Haberstroh, S. (2005): Current Research on Routine Decision Making: Advances and Prospects. In: Betsch T./Haberstroh S.: The Routines of Decision Making. Mahwah, NJ: Erlbaum Associates, S. 359–376.

Bögeholz, S. (2007): Bewertungskompetenz für systematisches Entscheiden in komplexen Gestaltungssituationen Nachhaltiger Entwicklung. In: Krüger, D./Vogt, H. (Hrsg): Theorien in der biologiedidaktischen Forschung. Berlin: Springer, S. 209–220.

De Haan, G./Kamp, G./Lerch, A./Martignon, L./Müller-Christ, G./Nutzinger, H.G. (2008): Nachhaltigkeit und Gerechtigkeit: Grundlagen und schulpraktische Konsequenzen. Berlin: Springer.

Eggert, S./Bögeholz, S. (2006): Göttinger Modell der Bewertungskompetenz – Teilkompetenz „Bewerten, Entscheiden und Reflektieren" für Gestaltungsaufgaben Nachhaltiger Entwicklung. In: Zeitschrift für Didaktik der Naturwissenschaften 12, S. 199–217.

Eggert, S./Bögeholz, S. (2010): Students' Use of Decision Making Strategies With Regard to Socioscientific issues – An Application of the Rasch Partial Credit Model. Science Education 94, H. 2, S. 230–258.

Ericsson, A.K./Simon, H.A. (1993): Protocol analysis – verbal reports as data. Cambridge, MA: MIT Press.

Mayring, P. (2008): Qualitative Inhaltsanalyse. Weinheim: Beltz.

Roberts, L./Wilson, M./Draney, K. (1997): The SEPUP Assessment System: An Overview. In: BEAR Report Series, SA-97-1. Berkeley: University of California.

Scott, W./Gough, S. (2003): Sustainable Development and Learning. London and New York: Routledge Falmer.

Sachverständigenrat für Umweltfragen (SRU) (1994): Umweltgutachten 1994. Stuttgart: Metzler-Poeschel.

World Commission on Environment and Development (WCED) (1987): Our common future. Oxford: Oxford University Press.

Wilson, M./Sloane, K. (2000): From Principles to Practices: An Embedded Assessment System. In: Applied Measurement in Education 13, H. 2, S. 181–208.

Anschrift der Autor/innen

Ellen Gausmann, Georg-August-Universität Göttingen, Albrecht-von-Haller-Institut für Pflanzenwissenschaften, Didaktik der Biologie, Waldweg 26, D-37073 Göttingen
E-Mail: egausma@gwdg.de

Dr. rer. nat. Sabina Eggert, Georg-August-Universität Göttingen, Albrecht-von-Haller-Institut für Pflanzenwissenschaften, Didaktik der Biologie, Waldweg 26, D-37073 Göttingen
E-Mail: seggert1@gwdg.de (Korrespondenz an diese Adresse)

Prof. Dr. phil. Marcus Hasselhorn, Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Bildung und Entwicklung, Schlossstraße 29, D-60486 Frankfurt a.M.
E-Mail: hasselhorn@dipf.de

Prof. Dr. phil. Rainer Watermann, Georg-August-Universität Göttingen, Pädagogisches Seminar, Schulpädagogik und Empirische Schulforschung, Waldweg 26, D-37073 Göttingen
E-Mail: rwaterm@gwdg.de

Prof. Dr. rer. nat. Susanne Bögeholz, Georg-August-Universität Göttingen, Albrecht-von-Haller-Institut für Pflanzenwissenschaften, Didaktik der Biologie, Waldweg 26, D-37073 Göttingen
E-Mail: sboegeh@gwdg.de

Samuel Greiff/Joachim Funke

Systematische Erforschung komplexer Problemlösefähigkeit anhand minimal komplexer Systeme

Projekt Dynamisches Problemlösen[1]

1. Einführung

Bis vor wenigen Jahren basierten psychologische Testverfahren zur Kompetenz- und Fähigkeitsmessung fast ausschließlich auf konventionellen Papier-und-Bleistift-Methoden. Mit dem Aufkommen von Computern ergaben sich neue und effiziente Möglichkeiten zur Erfassung von Fähigkeiten, die heute in modernen diagnostischen Verfahren wie *Computer Adaptive Testing* (CAT) münden. Neben einer höheren Effizienz haben sich dank der technischen Entwicklung aber auch neuartige Konstrukte entwickelt, die über klassische Formate nicht erfassbar waren. Eines dieser Konstrukte ist *komplexes Problemlösen*, das per se dynamisch und interaktiv ist (vgl. Funke 2003), sodass eine Testung nur computerbasiert möglich ist.

Komplexes Problemlösen (KPL) hat in den vergangenen Jahrzehnten im Hauptinteresse experimentalpsychologischer Forschung gestanden. Demgegenüber vernachlässigt wurde die Individualdiagnostik von KPL, die lediglich vereinzelt Berücksichtigung fand (vgl. z.B. Beckmann/Guthke 1995; Wagener 2001). Zugleich ist ein aufkeimendes Interesse an cross-curricularen Kompetenzen und damit auch an KPL in internationalen Bildungsstudien wie PISA zu beobachten (vgl. Klieme/Leutner/Wirth 2005). Als erste Konsequenz wurden im Rahmen einer Felderprobung sowie nachfolgend in einer nationalen Ergänzungsstudie Deutschlands zu PISA 2000 Hinweise auf die Messbarkeit und das Potential dynamischer Problemlösefähigkeit erbracht (vgl. Klieme u.a. 2001): Über einen semantisch in den Kontext der Raumfahrt eingebetteten finiten Automaten[2] zur Erfassung der Problemlösefähigkeit konnten die beiden Facetten Wissenserwerb und Wissensanwendung erfasst werden. Explorative Faktorenanalysen, lineare Strukturgleichungsmodelle und multidimensionale Skalierungen zeigten empirisch, dass KPL, analytisches Problemlösen, fachspezifische Kompetenzen und Testintelligenz

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: Fu 173/11-2) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

2 Finite Automaten sind Systeme, die eine begrenzte (= finite) Zahl qualitativ unterschiedlicher Zustände annehmen können. Sie beschreiben auf abstrakter Ebene eine Vielzahl alltäglicher Systeme unterschiedlicher Komplexitätsstufen wie Fahrkartenautomaten oder Computerbetriebssysteme und werden in der Forschung häufig zur Messung der Problemlöseleistung herangezogen.

zwar korrelierte, jedoch voneinander abgrenzbare Konstrukte waren; KPL ließ sich dabei am besten separieren (vgl. Leutner u.a. 2005). Begrifflich betrachten wir KPL und dynamisches Problemlösen als identisch, da KPL im Gegensatz zum analytischen Problemlösen in sich stets dynamisch ist.

Auch aus einer praktischen Perspektive finden sich zahlreiche anwendungsbezogene Implikationen von KPL. Eine Vielzahl an Aktivitäten lässt sich formal als komplexe Problemlöseprozesse beschreiben, bspw. medizinische Notfälle, aber auch die Verwaltung der eigenen Finanzen oder die Bedienung von Fahrkartenautomaten am Bahnhof. Gemein sind diesen Aktivitäten die folgenden Merkmale: (a) eine hohe Anzahl an Variablen ist involviert und die Gesamtinformation über das System muss von Problemlösenden adäquat reduziert werden (Komplexität); (b) verschiedene Variablen beeinflussen eines oder mehrere Resultate (Vernetztheit); (c) das zugrunde liegende System ist nicht statisch (Dynamik); (d) die über das System vorliegende Information ist nicht erschöpfend (Intransparenz) und (e) Ziele können einander widersprechen und möglicherweise nicht simultan erreicht werden (Polylely).

Diese von Dörner (1986) in der Theorie der operativen Intelligenz benannten fünf Eigenschaften komplexer Probleme korrespondieren mit fünf Anforderungen an eine problemlösende Person: (a) die Reduktion überbordender Information auf einen handhabbaren Umfang (Informationsreduktion); (b) die Bildung adäquater Situationsmodelle zum Verständnis der gegebenen Situation (Modellbildung); (c) die Prognose weiterer Entwicklungen aufgrund der gegebenen Situation und im Lichte getroffener Maßnahmen (Prognose); (d) die Beschaffung fehlender, aber für die Problemlösung notwendige Information (Informationssuche und -generierung) und (e) das Treffen von Wertentscheidungen und Prioritätensetzungen, mit denen Ziele gesetzt und Zielkonflikte gelöst werden können (Bewertung).

Aus theoretischer Sicht sollte ein Messverfahren Indikatoren zu jeder dieser Anforderungen enthalten; ein derartiges Facettendiagnosticum existiert bislang aber nicht. Ergänzend bestätigen zwar aus empirischer Sicht die oben berichteten PISA-Ergebnisse vorläufig die konvergente und divergente Validität von KPL, der verwendete Raumfahrt-Automat war allerdings ein ad hoc konstruiertes Testverfahren mit unklaren psychometrischen Eigenschaften, unklarem Messbereich und fehlender theoretischer Anbindung.

Wir stellen hier ein neuartiges Diagnosticum, MicroDYN, mit überprüfbaren psychometrischen Eigenschaften, theoretischem Bezug und der Möglichkeit, einzelne Facetten der Problemlösefähigkeit zu evaluieren, vor. Im empirischen Teil dieses Artikels konzentrieren wir uns auf Modellbildung als einen der fünf Indikatoren Dörners und leiten für diesen ein vorläufiges Kompetenzmodell ab.

2. Der MicroDYN-Ansatz

Ungeachtet des gestiegenen Interesses an der individualdiagnostischen Erfassung von KPL besteht nach wie vor ein grundlegender Mangel an gut eingeführten Testverfahren. Zusätzlich existiert kaum Einvernehmen darüber, wie KPL zu operationalisieren und zu

messen ist. Selbst für bestehende Tests gibt es keine hinreichend gesicherten theoretischen Grundlagen, an denen die Messung ansetzen könnte.

Neben diesen insgesamt unbefriedigenden Aspekten ist gegen die in PISA durchgeführte Form der Messung ein Einwand anzuführen, der im Übrigen alle simulierten Mikrowelten, wie sie erstmals von Dörner in den 1970er Jahren entwickelt wurden (vgl. Funke/Frensch 2007), betrifft: Die gesamte Testung besteht aus einem einzelnen Item, das über eine geraume Zeitspanne bearbeitet wird (*one item testing*). Sämtliche Bearbeitungsschritte hängen dabei von vorherigen Entscheidungen und der durchgängig unveränderten Systemstruktur ab. Im Ergebnis basieren die Aussagen über individuelle Problemlösefähigkeit auf der Leistung in diesem einem Item, was grundlegenden psychometrischen Anforderungen widerspricht. Einige Autor/innen versuchen dieses Problem über (a) die Vorgabe eines Systems, das aus mehreren unabhängigen Teilsystemen besteht und die fälschlicherweise separat ausgewertet werden (vgl. z.B. Müller 1993; Wagener 2001), oder über (b) mehrere Fragen zu einem System (wie im finiten Automaten aus PISA) zu lösen. Dies macht die Items aber nur scheinbar unabhängiger und das grundsätzliche Problem besteht weiterhin: (a) Ein umfangreiches System aus unabhängigen Subsystemen bleibt letztlich auch nur ein einzelnes Item und (b) eine Vielzahl an Items (Fragen) zu einem einzelnen Szenario könnte bestenfalls als lokal abhängiges Item-Bündel verstanden werden.

Vor dem dargestellten Hintergrund stellt sich die Frage, wie dynamisches Problemlösen über psychologische Messverfahren überhaupt getestet werden kann. Wir nehmen an, dass interindividuelle Unterschiede im Kontext linearer Strukturgleichungsmodelle erfassbar werden. Dieser Formalismus (s. Abschnitt 3) wurde verschiedentlich als ökologisch valide bezeichnet und ist bereits häufig in experimentellen Arbeiten als Indikator der Problemlöseleistung verwendet worden (vgl. Funke 2001) – dort allerdings als *one item testing*. Wir wählen nun einen modifizierten Ansatz: Anstatt nur ein einzelnes Item darzubieten, bearbeiten Proband/innen unter strikter Zeitbegrenzung eine ganze Serie *minimal komplexer Systeme*. Wir nennen diesen Zugang MicroDYN, um damit die Orientierung an der kleinsten Einheit der Komplexität zu verdeutlichen und zugleich eine Referenz an den zugrunde liegenden DYNAMIS-Ansatz zu machen (vgl. ebd.).

Der MicroDYN-Ansatz vermag einige Versäumnisse bestehender Messverfahren zu lösen oder zumindest deutlich abzumildern. Er bietet dabei die folgenden Vorteile: (a) durch die Anbindung an Dörners Anforderungen an Problemlösende ist ein gewisser theoretischer Bezug gewährleistet (Theoriebezug); (b) durch die Vorgabe von etwa 15 Items mit kurzer Bearbeitungszeit, aber hinreichender Komplexität wird ein entscheidender Nachteil bisheriger Diagnostik überwunden (Itemindependenz); (c) empirisch nachweisbare Facetten der Problemlöseleistung im Sinne der Dörnerschen Anforderungen können zuverlässig gemessen werden (Facettendiagnostik); (d) Items sind einfach zu entwickeln und hinsichtlich ihrer Schwierigkeit frei variierbar (infiniter Itempool); (e) psychometrische Eigenschaften wie Reliabilität können standardmäßig überprüft werden (psychometrisch orientierte Testentwicklung); und (e) Alltagsaktivitäten können über MicroDYN-Items modelliert werden (ökologische Validität).

3. Die Items

Ein typisches MicroDYN-Item (illustriert in Abbildung 1) besteht aus exogenen und endogenen Variablen (im Beispiel ein 3x3-System). Die exogenen Variablen können im Gegensatz zu den endogenen Variablen aktiv manipuliert werden. Denkbare Verknüpfungen zwischen den Variablen sind Haupteffekte (HE), multiple Effekte (ME), multiple Abhängigkeiten (MA), Eigendynamiken (ED) und Nebeneffekte (NE). *Haupteffekte* beschreiben kausale Relationen einer exogenen auf eine endogene Variable. Wirkt eine exogene Variable auf mehrere endogene, so ist dies ein *multipler Effekt*. Wird umgekehrt eine endogene Variable von mehreren exogenen beeinflusst, wird dies *multiple Abhängigkeit* genannt. Diese drei Effekte können aktiv manipuliert werden und sind nur dann zu beobachten, wenn die exogenen Variablen (durch direkte Eingabe) Werte ungleich 0 aufweisen. Wirkt eine endogene Variable auf andere endogene, ist dies ein *Nebeneffekt*. Wirkt sie hingegen auf sich selbst (mit einem Gewicht ungleich 1), wird dieser Spezialfall eines Nebeneffektes *Eigendynamik* (als Wachstums- oder Schrumpfungsprozess) genannt. NE und ED können nicht aktiv manipuliert werden und hängen grundsätzlich von den Zuständen der endogenen Variablen ab, sind also weitestgehend unabhängig von den aktiven Manipulationen der Proband/innen. Über die Verwendung adäquater Strategien können sie aber eindeutig identifiziert werden.

Ein Item wird von Proband/innen stets in drei Schritten durchlaufen: (a) Explorationsphase, (b) Modellbildungsphase und (c) Steuerphase.

In der *Explorationsphase* (a) können Proband/innen das System eigenständig und vollständig frei explorieren. Ihre Aufgabe besteht darin, sich mit dem System und seiner

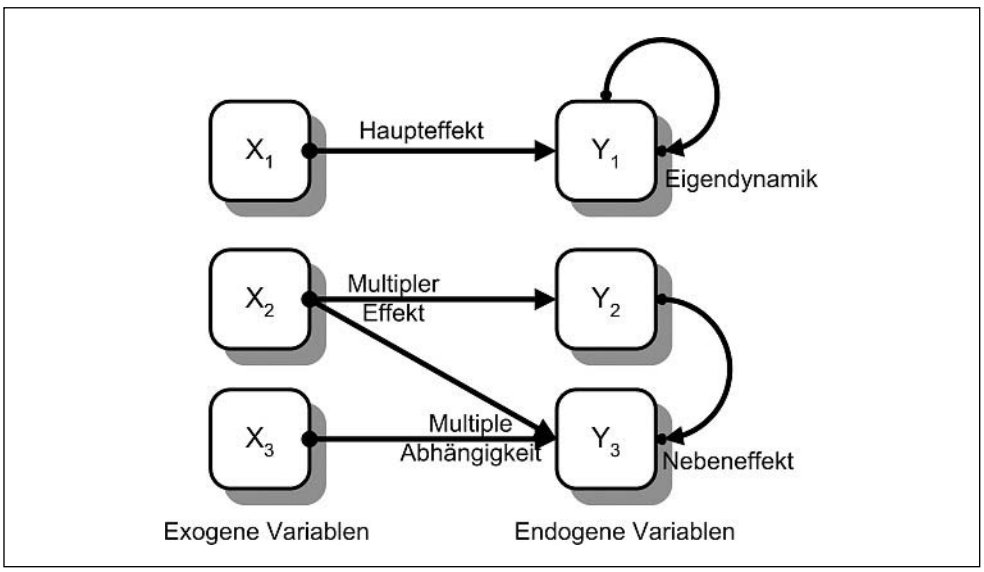


Abb. 1: Struktur eines MicroDYN-Items mit 3 exogenen und 3 endogenen Variablen und den 5 Effektqualitäten

Struktur vertraut zu machen. Die hinterlegten Verknüpfungen sind dabei für die Proband/innen nicht ersichtlich.

Die Systemstruktur soll im Anschluss oder parallel zur ersten Phase in der *Modellbildungsphase* (b) expliziert und aufgezeichnet werden. Phasen (a) und (b) dauern gemeinsam etwa 3 bis 4 Minuten.

In der *Steuerphase* (c) werden Proband/innen mit vorgegebenen Zielwerten in den endogenen Variablen konfrontiert, die sie durch adäquate Manipulation der exogenen Variablen in mehreren Schritten erreichen sollen. Phase (c) dauert 1½ Minuten.

Proband/innen bearbeiten insgesamt 12 bis 15 unabhängige Systeme (Dauer pro Item 4 bis 5 Minuten), was einer angemessenen Itemzahl entspricht und die Testzeit auf ökonomisch vertretbare 60 Minuten begrenzt. Anhand dieser drei Phasen können Aussagen über die Anforderungen der Informationssuche und -generierung (Phase a), der Modellbildung (Phase b) sowie der Prognose (Phase c) getroffen werden. Indikatoren für die beiden Anforderungen der Informationsreduktion und der Bewertung sind über speziell konstruierte Items möglich und derzeit in Vorbereitung.

4. Aktuelle Forschung

Obwohl der beschriebene Itemtyp bereits vielfach als Operationalisierung komplexer Problemlösefähigkeit verwendet wurde, blieb dabei grundsätzlich unklar, ob diese Operationalisierungen reliabel und valide waren. Insbesondere die Vergleichbarkeit zwischen Studien war aufgrund unterschiedlicher Systemstrukturen nicht gewährleistet. Wir glauben, dass ein Blick auf Systemseite und die Aufschlüsselung der Itemschwierigkeit in einzelne systemimmanente Merkmale wichtige Hinweise auf die Anforderungen solcher Systeme, aber auch auf eine mögliche Kompetenzstruktur auf Personenseite geben können. In einer Aufgabenanalyse finden wir sieben Systemdimensionen mit potentiellem Einfluss auf die Itemschwierigkeit (Tabelle 1). Diesen Einfluss haben wir in der vorliegenden Untersuchung zu quantifizieren versucht mit dem Ziel, (a) einen Beitrag zur Messbarkeit komplexer Problemlösefähigkeit zu liefern und (b) ein vorläufiges Kompetenzmodell für den Aspekt der Modellbildung abzuleiten. Wir beschränken uns dabei auf diesen Aspekt der Problemlösefähigkeit.

4.1 Design

In einem Messwiederholungsdesign (n = 48; 39 weiblich, 9 männlich; Alter M = 23.42, S = 3.02) bearbeiteten Proband/innen 15 MicroDYN-Items mit einer Gesamtdauer von ca. 60 Minuten. Der Fokus lag dabei auf den ersten drei Dimensionen aus Tabelle 1. Es wurden vornehmlich Haupteffekte untersucht, lediglich eine Interaktion für die als a priori besonders relevant angenommenen Faktoren Effektzahl und Effektqualität integrierten wir. Der dreistufige Faktor *Effektqualität* (HE, ME, NE; s. Abbildung 1; MA und ED wurden nicht berücksichtigt; Ergebnisse hierzu werden kurz in der Diskussion erwähnt)

Systemmerkmal	mögliche Ausprägungen; Merkmalserklärung
Effektqualität	Haupteffekt, multipler Effekt, multiple Abhängigkeit, Eigendynamik, Nebeneffekt; qualitativ unterschiedliche Verknüpfungen
Effektzahl	frei variierbar; Zahl der Effekte in einem gegebenen System
Variablenzahl	frei variierbar; Zahl der exogenen und endogenen Variablen
relative Effektstärke	frei variierbar; relativer Grad eines Effektes; beinhaltet auch Vorzeichen
Start- und Zielwerte	frei variierbar; Zielwerte nur in der Steuerphase für endogene Variablen
Dispersion	niedrig bis stark; Clusterung bzw. Verteilung der Effekte auf exogene und endogene Variablen
Konfiguration	frei variierbar; Anordnung der Elemente und der Verknüpfungen

Tab. 1: Anhand der Aufgabenanalyse identifizierte Systemmerkmale und ihre Erläuterung

und der zweistufige Faktor *Effektzahl* (2 & 4; s. Abbildung 2) wurden in einem 2 × 3-Design vollständig gekreuzt. Weiter wurde der dreistufige Faktor *Zahl der Variablen* (2 × 2, 3 × 3, 4 × 4; s. Abbildung 3) isoliert variiert. Über Kovarianzanalysen wurde der Einfluss der Itemposition auspartialisiert.

4.2 Hypothesen

In der vorliegenden Untersuchung wurden nur drei der in Tabelle 1 aufgeführten Faktoren getestet: Effektqualität, Effektzahl und Variablenzahl. Für *Effektqualität* (EQ) nehmen wir die beste Erkennbarkeit bei HE an, gefolgt von ME, während NE schwerlich erkennbar sein sollten. Dies folgt aus der Manipulierbarkeit sowie der Zahl notwendiger Explorationsschritte, um die jeweiligen Effekte zu erkennen. Eine höhere *Effektzahl* (EZ) sollte die Schwierigkeit eines Systems merklich erhöhen. Hinsichtlich der Interaktion EQ × EZ vermuten wir keine statistisch bedeutsamen Effekte. Bei steigender *Variablenzahl* sollte ceterus paribus die Leistung der Proband/innen sinken.

4.3 Abhängige Variable

Die Wissensabfrage erfolgte nach der Methode der Kausaldiagramm-Analyse (vgl. Funke 1995), in der Proband/innen die vermutete Systemstruktur auf Papier aufzeichnen und anhand vorgegebener Kategorien die angenommene Stärke eines Pfadgewichtes angeben. Als Indikator wurde die *Güte der Kausaldiagramme* (GdK) in einer

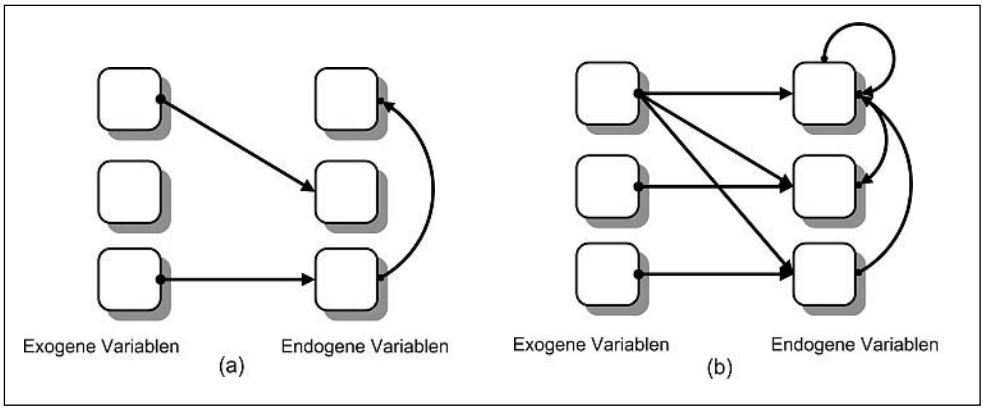


Abb. 2: Zwei Items mit (a) niedriger bzw. (b) hoher Anzahl an Effekten bei konstant gehaltener Zahl der Variablen

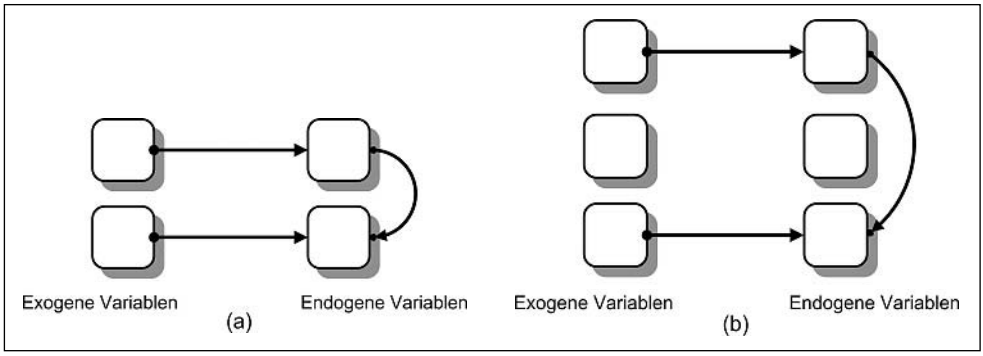


Abb. 3: Zwei Items mit (a) 2 bzw. (b) 3 exogenen und endogenen Variablen bei konstant gehaltener Effektzahl

Gewichtung von 0,75 Relations- und 0,25 Stärkewissen verwendet. Richtungswissen (Vorzeichen) wurde nicht ausgewertet. GdK erreicht ein Maximum von 1 (Systemwissen absolut richtig) und ein Minimum von −1 (Systemwissen absolut falsch). GdK hatte sich in einer Simulationsstudie mit insgesamt über 40 möglichen Indikatoren gemessen an einem Kriterium bestehend aus Expertenbeurteilungen fiktiver Kausaldiagramme als überlegener Indikator gezeigt.

4.4 Ergebnisse

Die Voraussetzungen für Kovarianzanalysen waren in hinreichender Weise erfüllt.

Die Effekte der Systemmerkmale auf GdK finden sich in Tabelle 2 und bestätigten generaliter unsere Hypothesen. Die Qualität eines Effektes wirkte sich signifikant ($p < .001$) auf das Kausalwissen aus. HE ($M = .71$; SE = .03) und ME ($M = .70$; SE = .04)

waren leichter als NE (M = .60; SE = .04; p < .001; Kontraste insgesamt nicht dargestellt). HE und ME unterschieden sich dabei nicht (p > .10). Dies mag daran liegen, dass Nebeneffekte nur beobachtet, nicht aber aktiv manipuliert werden können. Bei höherer Effektzahl stieg die Schwierigkeit eines Systems deutlich an (M_{EZ2} = .75; SE_{EZ2} = .04; M_{EZ4} = .60; SE_{EZ4} = .04; p < .001). Dieser Effekt blieb unverändert, wenn die Effektzahl als zufälliger Faktor verstanden wurde (nicht dargestellt). Zwischen der Effektzahl und ihrer Qualität bestand keine Interaktion (Abbildung 4). Eine unterschiedliche Variablenzahl (M_{2x2} = .71; $SE_{2\times2}$ = .04; $M_{3\times3}$ = .65; $SE_{3\times3}$ = .05; $M_{4\times4}$ = .61; $SE_{4\times4}$ = .06) beeinflusste GdK nicht in der Omnibustestung. Ein geplanter linearer Kontrast allerdings zeigte ein stetiges Ansteigen der Schwierigkeit mit zunehmender Variablenzahl (p < .05).

Um erste Aussagen über die psychometrischen Eigenschaften MicroDYNs treffen zu können, verstanden wir die experimentellen Stimuli als eine Testversion mit 15 Items. Cronbachs α als Reliabilitätsschätzung fiel mit .94 sehr zufriedenstellend aus.

Faktor	Stufen	*F*	*df*	*p*	η^2	KIη^2(.95)
Effektqualität	HE vs ME vs NE	9.95	2/90	<.001**	.18	[.05;.31]
Effektzahl	2 vs 4	37.74	1/45	<.001**	.46	[.23;.60]
IA EQ x EZ		1.61	2/90	>.10	.04	[.00;.12]
Variablenzahl	2 vs 3 vs 4	2.10	2/92	>.10	.04	[.00;.13]

Tab. 2: ANCOVA-Ergebnisse für die getesteten Effekte
Anmerkungen: Kovariate ist Itemposition. KI = Konfidenzintervall; IA = Interaktion; EQ = Effektqualität; EZ = Effektzahl; HE = Haupteffekt; ME = multipler Effekt; NE = Nebeneffekt

5. Implementation

Die Software wurde in enger Kooperation mit dem Deutschen Institut für Internationale Pädagogische Forschung (DIPF, Frankfurt a.M.) und der Firma SOFTCON (München) entwickelt. Die endgültige Version (verfügbar ab Herbst 2009) ist als Autorensystem in die frei zugängliche Plattform TAO (vgl. Latour/Martin 2007) integriert und wird erheblichen Freiraum in Bezug auf graphische Gestaltung, Semantik und Itementwicklung lassen. In Abbildung 5 ist ein Screenshot der derzeitigen Software illustriert.

Neben dem eigentlichen System mit den exogenen Variablen links und den endogenen rechts werden eine Historie sowie ein Zeitbalken dargeboten. Ein Undo- und ein Resetbutton ermöglichten es den Proband/innen in der Explorationsphase, vorherige Schritte zu korrigieren. In der Steuerphase werden zusätzlich extern vorgegebene Zielwerte angezeigt. Der Verlauf wird am unteren Seitenrand dokumentiert.

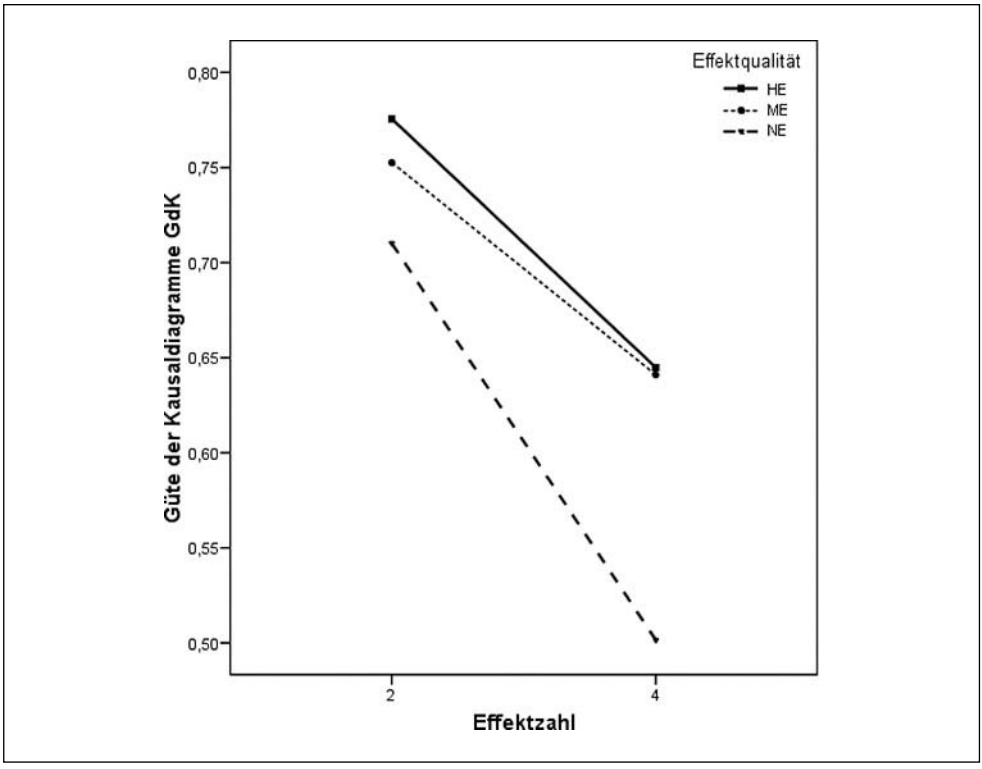


Abb. 4: Mittlere Werte „Güte der Kausaldiagramme" (GdK) in Abhängigkeit von der Effektzahl (2 & 4) und der Effektqualität (HE = Haupteffekt, ME = multipler Effekt, NE = Nebeneffekt)

6. Diskussion

6.1 Kompetenzmodell

Auf Grundlage der Ergebnisse entwickelten wir ein vorläufiges Kompetenzniveaumodell, das in Abbildung 6 dargestellt ist und für den Aspekt der Modellbildung eine mögliche Kompetenzstruktur menschlichen Verhaltens in dynamischen Systemen beschreibt. Hierin integrierten wir die relevanten Merkmale Effektzahl und -qualität. Ergebnisse zu den verbliebenen Itemmerkmalen aus Tabelle 1 wurden hier nicht berichtet, weisen aber darauf hin, dass diese in ihrem Einfluss auf die Schwierigkeit vernachlässigt werden können. Bezogen auf die hier ausschließlich verwendeten 4×4-Systeme bilden gering und stark vernetzte Systeme zwei Kompetenzniveaus, innerhalb derer unterschiedliche Effektqualitäten als verschieden gut erkennbar angenommen werden. HE, ME und MA sind gleich gut erkennbar, gefolgt von ED. Am schwersten zu erschließen sind NE (Ergebnisse zu MA und ED aus nicht dargestelltem Experiment).

Abb. 5: Screenshot der MicroDYN-Software.
Anmerkung: Im oberen Teil befinden sich links vier exogene und rechts vier endogene Variablen. Im unteren Teil ist die Interventionshistorie angezeigt. Rechts oben ist die Restzeit visualisiert.

6.2 Einschränkungen in der vorliegenden Studie

Aus inhaltlicher Sicht stellt sich die grundlegende Frage nach dem Wesen komplexer Problemlösefähigkeit und ob diese überhaupt im Rahmen standardisierter Testverfahren erfassbar ist. Diese Diskussion ist nicht neu, wird aber durch die Kürze und die Vielzahl an Systemen im MicroDYN-Ansatz aktuell. Abstriche in der ökologischen Validität sind u.E. unvermeidbar und im Übrigen bei allen psychologischen Konstrukten (bspw. Intelligenz) gegeben, wenn auch komplexe Probleme hiervon besonders betroffen sein mögen. Systemschwierigkeit auf einzelne Merkmale zurückzuführen ist nicht unumstritten, entsteht die Komplexität doch gerade aus dem Zusammenspiel vieler Einzelvariablen. Die Befunde sprechen allerdings für das gewählte Vorgehen: Systemeigenschaften (als fixe Effekte verstanden) wirken sich deutlich und den Erwartungen entsprechend auf den Umgang mit diesen Systemen aus und unter psychometrischen Aspekten sind unabhängige Items mit kurzer Darbietungsdauer nachgerade unumgänglich. Ein weiterer wesentlicher Einwand tangiert den Realitätsbezug der verwendeten Probleme.

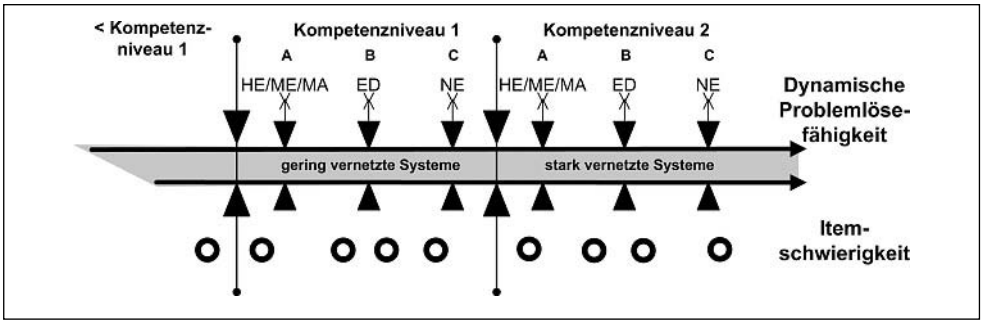


Abb. 6: Angenommenes Kompetenzmodell mit drei qualitativen Kompetenzniveaus und Items variierender Schwierigkeit

Tatsächlich weisen MicroDYN-Systeme an der Oberfläche nur wenige Gemeinsamkeiten mit alltäglichen Situationen auf, können diese aber formal gut modellieren (vgl. Beckmann/Guthke 1995). Im Übrigen erleben Proband/innen die Aufgaben als durchaus komplex und dynamisch. Auch ist es keinesfalls realitätsfern, sich mit alltäglichen Problemen (man denke z.B. an einen Fahrkartenautomat oder die Einstellung eines Thermostats) nur wenige Minuten zu beschäftigen und dann zum nächsten Problem zu wechseln („Wo ist Gleis 21a?“ oder Programmierung des DVD-Rekorders). Modellbildung beschreibt außerdem nur eine der fünf identifizierten Anforderungen an einen Problemlöser. Eine Integration der verbleibenden vier ist in MicroDYN aber möglich und wurde im Abschnitt 3 bereits angedeutet.

6.3 Ausblick

Im naturwissenschaftlichen Unterricht sind Schüler/innen häufig mit unbekannten Systemen konfrontiert, in denen sie selbstständig explorieren und experimentieren müssen. Aber nicht nur dort, sondern auch in vielfältigen anderen Bereichen spielt die Fähigkeit, mit dynamischen Systemen umzugehen, eine wichtige Rolle. Existiert – unabhängig von der semantischen und situativen Einbettung – eine solche breite Kompetenz? In der hier berichteten Erhebung wurde semantisch ein (vorwissenarmes) Chemielabor gewählt. Die Frage nach dem Einfluss des situativen und semantischen Kontexts, die im Rahmen fachübergreifender Kompetenzdiagnostik zentral ist, wird über unterschiedliche Oberflächenstrukturen bei gleich bleibender Systemstruktur derzeit in MicroDYN untersucht. Wie aber sähe ein entsprechend domänenunspezifischer Test aus? Dieser Frage konnte bisher nur unzureichend nachgegangen werden, da mit Ausnahme von experimentellen *ad hoc* Konstruktionen keine adäquaten Erhebungsinstrumente existierten. An diesem Punkt möchten wir anknüpfen und Impulse für die Entwicklung einer fundierten Messung geben, da dynamische Problemlösefähigkeit als cross-curriculare Kompetenz in seiner Relevanz für Bildung und Unterricht keinesfalls zu unterschätzen ist. Sie stellt ein Konstrukt mit inkrementellem Potential dar, das in vielfältigen Situationen von Lernen und Unterricht relevant ist. Wir wünschen uns –

dies sollte deutlich geworden sein – eine weniger an inhaltlichen Konzepten, sondern mehr an testtheoretischer Güte orientierte Messtradition, wie es in anderen Bereichen pädagogischer Bildungsforschung schon seit Jahren Standard ist. Wenn diese Arbeit einen Anstoß in diese Richtung liefern kann, ist viel erreicht.

Danksagung

Wir danken Ursula Pöll, Britta Veith und Sascha Wüstenberg für ihre Hilfe bei der Datenerhebung.

Literatur

Beckmann, J./Guthke, J. (1995): Complex problem solving, intelligence, and learning ability. In: Frensch, P.A./Funke, J. (Hrsg.): Complex problem solving: The European perspective. Hillsdale, NJ: Lawrence Erlbaum, S. 177–200.

Dörner, D. (1986): Diagnostik der operativen Intelligenz. In: Diagnostica 32, S. 290–308.

Funke, J. (1995): Experimental research on complex problem solving. In: Frensch, P.A./Funke, J. (Hrsg.): Complex problem solving: The European perspective. Hillsdale, NJ: Lawrence Erlbaum, S. 243–268.

Funke, J. (2001): Dynamic systems as tools for analysing human judgement. In: Thinking and Reasoning 7, S. 69–89.

Funke, J. (2003): Problemlösendes Denken. Stuttgart: Kohlhammer.

Funke, J./Frensch, P.A. (2007): Complex problem solving: The European perspective – 10 years after. In: Jonassen, D.H. (Hrsg.): Learning to solve complex scientific problems. New York: Lawrence Erlbaum, S. 25–47.

Klieme, E./Funke, J./Leutner, D./Reimann, P./Wirth, J. (2001): Problemlösen als fächerübergreifende Kompetenz. Konzeption und erste Resultate aus einer Schulleistungsstudie. In: Zeitschrift für Pädagogik 47, S. 179–200.

Klieme, E./Leutner, D./Wirth, J. (Hrsg.) (2005): Problemlösekompetenz von Schülerinnen und Schülern. Wiesbaden: VS Verlag für Sozialwissenschaften.

Latour, T./Martin, R. (2007): TAO, an open and versatic computer-based assessment platform based on semantic web technology. In: ERCIM News 71, S. 32–33.

Leutner, D./Wirth, J./Klieme, E./Funke, J. (2005): Ansätze zur Operationalisierung und deren Erprobung im Feldtest zu PISA 2000. In: Klieme, E./Leutner, D./Wirth, J. (Hrsg.): Problemlösekompetenz von Schülerinnen und Schülern. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 21–36.

Müller, H. (1993): Komplexes Problemlösen: Reliabilität und Wissen. Bonn: Holos.

Wagener, D. (2001): Psychologische Diagnostik mit komplexen Szenarios. Taxonomie, Entwicklung, Evaluation. Lengerich: Pabst Science Publishers.

Anschrift der Autoren

Dipl.-Psych. Samuel Greiff, Psychologisches Institut der Universität Heidelberg,
Hauptstr. 47–51, D-69117 Heidelberg
E-Mail: samuel.greiff@psychologie.uni-heidelberg.de

Prof. Dr. Joachim Funke, Psychologisches Institut der Universität Heidelberg, Hauptstr. 47–51,
D-69117 Heidelberg
E-Mail: joachim.funke@psychologie.uni-heidelberg.de

Klaus Lingel/Nora Neuenhaus/Cordula Artelt/Wolfgang Schneider

Metakognitives Wissen in der Sekundarstufe: Konstruktion und Evaluation domänenspezifischer Messverfahren

Projekt EWIKO[3]

1. Einführung

1.1 Projektbeschreibung und theoretischer Hintergrund

Das Projekt EWIKO („Entwicklung von Wissenskomponenten") untersucht die Entwicklung und Interaktion von metakognitivem Wissen und inhaltlichem Wissen in den Fächern Deutsch, Mathematik und Englisch im Verlauf der Sekundarstufe I (Jahrgangsstufen 5 bis 8). Die längsschnittliche Anlage der Studie erlaubt die Beschreibung von inter- und intraindividuellen Entwicklungsverläufen im metakognitiven und inhaltlichen Wissen. In der Vorbereitung auf die Längsschnittuntersuchung wurden Messverfahren zur Erfassung des domänenspezifischen, deklarativen metakognitiven Wissens erarbeitet bzw. bestehende Instrumente modifiziert. Der vorliegende Beitrag gibt einen Überblick über Konstruktion und Evaluation dieser Instrumente.

1.2 Theoretischer Hintergrund

Metakognitives Wissen und inhaltliches Wissen gelten neben allgemeinen kognitiven, motivationalen und volitionalen Schülermerkmalen als individuelle Determinanten der Schulleistung (vgl. Helmke/Rindermann/Schrader 2008). Das Modell des „*good information processing*" (vgl. Pressley/Borkowski/Schneider 1989) spannt einen theoretischen Rahmen, in dem die Interaktion aus Inhaltswissen, dem Wissen um Strategien und ihre Umsetzung (metakognitives Wissen), Motivation und Verarbeitungskapazität im Prozess schulischen Lernens und Leistens hervorgehoben wird. Von besonderer Bedeutung ist das in der Informationsverarbeitung besonders saliente metakognitive Wissen: es koordiniert und strukturiert komplexere Verarbeitungsprozesse (vgl. Pressley 1994).

Metakognitives Wissen enthält eine deklarative und eine prozedurale Komponente. Gemäß dem Ansatz von Flavell (1979) versteht man unter deklarativem metakognitiven

3 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: SCHN 315/36-1 und AR 301/8-1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen" (SPP 1293).

Wissen das verbalisierbare Wissen über Faktoren, die Einfluss auf kognitive Prozesse nehmen. Dieses bezieht sich auf Aspekte der eigenen Person, der vorliegenden Aufgabe und der zur Verfügung stehenden Strategien bzw. der Interaktion dieser Aspekte. Insbesondere das Wissen, wo und wann bestimmte Strategien einzusetzen sind (konditionales Wissen sensu Jacobs/Paris 1987), entfaltet spezifische Verhaltenswirksamkeit, da dieses Wissen eine notwendige (aber nicht hinreichende) Bedingung des kompetenten Einsatzes metakognitiver Strategien zur Steuerung kognitiver Prozesse (prozedurales metakognitives Wissen) darstellt (vgl. Carr/Biddlecomb 1998; Borkowski/Chan/Muthukrishna 2000).

Schlagmüller, Visé und Schneider (2001) berichten über psychometrische Schwierigkeiten, das deklarative metakognitive Wissen über Gedächtnisprozesse auf Grundlage von Selbstberichtsverfahren reliabel und valide zu erfassen. Die üblicherweise eingesetzten Verfahren, die auf die allgemeine Nutzungshäufigkeit von Strategien rekurrieren, erfordern den Abruf von strategischen Lernerfahrungen und die Abstraktion dieser episodischen Gedächtnisinhalte über eine Vielzahl von Anwendungssituationen. Freiere – und wesentlich aufwendigere – Erhebungsverfahren (z.B. Interviews, Lerntagebücher) setzen u.a. die Explikation mehr oder weniger prozeduralisierten Handlungswissens voraus (vgl. Spörer/Brunstein 2006). Die komplexen Anforderungen, die eine sachgemäße Bearbeitung dieser Instrumente stellt, sind in Schülerstichproben in unterschiedlichem Ausmaß erfüllt. Dies kann zu in ihrer Validität eingeschränkt interpretierbaren Messungen strategischer Kompetenzen bzw. metakognitiven Wissens führen (vgl. Artelt 2000; Samuelstuen/Bråten 2007).

Werden dagegen Verfahren eingesetzt, in denen die Abstraktionsanforderungen durch die Vorgabe von konkreten Anwendungssituationen reduziert sind und kaum Gedächtnis- bzw. Verbalisierungsanforderungen bestehen, da lediglich eine Effektivitätsbeurteilung bereits vorgegebener Strategien vorzunehmen ist, lassen sich auch schon bei sehr jungen Schüler/innen reliable und valide Einschätzungen des metakognitiven Wissens erzielen. Durch den Einbezug von Merkmalen der Anwendungssituation werden zusätzlich konditionale metakognitive Wissensaspekte berücksichtigt.

Für ein solches, durch Effektivitätsbeurteilungen von Strategiealternativen operationalisiertes, deklaratives metakognitives Wissen liegen im Bereich des Lern- und Gedächtnisstrategiewissens bereits querschnittliche Befunde aus dem Übergang von der Elementar- zur Primarstufe vor (vgl. Justice 1986). Van Kraayenoord und Schneider (1999) erfassten auf diese Weise am Ende der Grundschulzeit (3. und 4. Jahrgangsstufe) deklaratives Gedächtnis- und Lesestrategiewissen (Würzburger Metamemory-Batterie, Index of Reading Awareness aus Jacobs/Paris 1987) und fanden relativ enge Zusammenhänge beider Strategiemaße mit der Leseleistung. Eine Folgeuntersuchung an derselben Stichprobe vier Jahre später (vgl. Roeschl-Heils/Schneider/Van Kraayenoord 2003) replizierte diesen Zusammenhang und zeigte darüber hinaus eine hohe Stabilität des Konstrukts: zwei unterschiedliche Verfahren zur Erfassung des Lesestrategiewissens (vgl. Index of Reading Awareness bzw. Würzburger Lesestrategietest, WLST, Schlagmüller/Schneider 2007) zeigten über das Vierjahresintervall Korrelationen von $r = .50$.

In einer der wenigen Längsschnittstudien zur Entwicklung des Strategiewissens untersuchten Annevirta u.a. (2007) Kinder zu drei Messzeitpunkten vom Kindergarten bis zum zweiten Schuljahr in ihren Effektivitätsbeurteilungen über Lern- und Verstehensstrategien und konnten v.a. mit Eintritt in die Schule eine beschleunigte Zunahme des metakognitiven Wissens sowie enge Zusammenhänge mit der Lese- und Hörverstehensleistung der Kinder nachweisen.

Der von Schlagmüller und Schneider (2007) für den Einsatz in der Sekundarstufe entwickelte Würzburger Lesestrategietest (WLST), der ebenfalls auf der Effektivitätsbeurteilung von Strategiealternativen basiert, weist neben psychometrischer Zuverlässigkeit ein hohes Maß an Gültigkeit als Indikator für das Wissen über Lese- und Textverständnisstrategien auf, was sich in substanziellen Zusammenhängen zwischen Lesestrategiewissen und Lesekompetenz ausdrückt. Darüber hinaus lassen sich auch die aufgrund von Lernerfahrung und Leistungszuwachs erwarteten Unterschiede zwischen Jahrgangsstufen und Schularten im Strategiewissen nachweisen (vgl. auch Artelt u.a. 2009).

Diese Beobachtungen stehen in Übereinstimmung mit einer konstruktivistischen Perspektive der Entwicklung von metakognitivem Wissen (vgl. Borkowski/Chan/Muthukrishna 2000; Carr/Biddlecomb 1998). Die Entwicklung hängt demnach im Wesentlichen von den schulischen Lern- und Förderbedingungen ab und sollte je nach schulischer Domäne mit den Gelegenheiten zum intentionalen Lernen variieren. Daher ist v.a. im Verlauf der Sekundarstufe eine substanzielle Zunahme des deklarativen metakognitiven Wissens zu erwarten (vgl. Pressley/McCormick 1995; Schneider 2010). Der Wissenserwerb ist dabei als induktiver Prozess zu verstehen: zunächst aufgabenspezifisch erworbene Erfahrungen werden zunehmend innerhalb und über die Domänen hinweg generalisiert (vgl. Borkowski/Chan/Muthukrishna 2000). In Bezug auf die Annahme einer domänenspezifischen Entwicklung metakognitiven Wissens und die Frage, wann und in welchem Maße allgemeines metakognitives Wissen generiert wird, gibt es bislang allerdings kaum empirische Evidenz. Entsprechend der konstruktivistischen Perspektive gehen wir für den in unserer Studie zunächst untersuchten Altersbereich (10–11 Jahre) von einer domänenspezifischen Struktur des deklarativen metakognitiven Wissens aus.

1.3 Fragestellung

Zur Messung des deklarativen metakognitiven Wissens am Beginn der Sekundarstufe standen in den schulischen Kerndomänen, namentlich Lesen, Mathematik, Fremdsprachenerwerb und (selbstreguliertes) Lernen bislang keine validen Instrumente zur Verfügung. Daher mussten die für Schüler/innen am Ende der Sekundarstufe I vorliegenden Verfahren (WLST und ein von Artelt im Rahmen von PISA 2003 entwickeltes Verfahren zur Erfassung mathematischen Strategiewissens, vgl. Ramm u.a. 2006) auf die Gegebenheiten am Beginn der Sekundarstufe I angepasst werden. Zur Erfassung des Strategiewissens in Englisch und im selbstregulierten Lernen mussten von Grund auf neue Verfahren entwickelt werden.

Die hier berichtete Untersuchung wurde mit dem Ziel durchgeführt, die konstruierten Verfahren hinsichtlich ihrer Einsetzbarkeit in der ersten Phase der Längsschnittstudie (Jahrgangsstufe 5 und 6) zu evaluieren. Die Instrumente können dann als tauglich angesehen werden, wenn sie das im untersuchten Altersbereich vorfindbare Leistungsspektrum (in Abhängigkeit von Jahrgangsstufe und Schulart) reliabel und valide abbilden.

2. Methodisches Vorgehen

2.1. Stichprobe

Die psychometrischen Eigenschaften der konstruierten Verfahren wurden an einer Stichprobe von insgesamt 798 bayerischen Schüler/innen in 33 Klassen der fünften und sechsten Jahrgangsstufe von Hauptschule und Gymnasium untersucht (Tabelle 1). Die Gesamtstichprobe teilt sich in vier Gruppen auf. Jede dieser Substichproben bearbeitete aus testökonomischen Gründen jeweils zwei der vier konstruierten metakognitiven Wissenstests (zur Zellbesetzung s. Tabelle 2).

	5. Jahrgang		6. Jahrgang		Gesamt
	Jungen	Mädchen	Jungen	Mädchen	
Hauptschule	127	135	83	73	418 (52%)
Gymnasium	93	87	143	57	380 (48%)
Gesamt	442 (56%)		356 (44%)		798

Tab. 1: Übersicht über die Gesamtstichprobe (in Klammern Prozentanteile an der Gesamtstichprobe)

2.2 Instrumente

Testkonstruktion

Die metakognitiven Wissenstests bestehen aus Lernszenarien, die Lernanforderungen beschreiben, denen die Altersgruppe in den zu untersuchenden Domänen begegnet. Jedem dieser Szenarien sind mehr oder weniger effektive Strategievorschläge zugeordnet. Die Vorschläge für effektive Strategien (metakognitive Strategien der Planung, Überwachung und Regulation kognitiver Prozesse bzw. kognitive Strategien der Wiederholung, Elaboration und Organisation) wurden aus einschlägigen Arbeiten (Metaanalysen, Interventions- und Trainingsstudien) zu Effekten spezifischer Lern- und Arbeitsstrategien abgeleitet. Diese effektiven Strategievorschläge wurden um wenig effektive, alterstypische Mängelstrategien ergänzt.

Die Schüler/innen erhalten die Instruktion, die Strategievorschläge im jeweils vorgegebenen Szenario mit Noten von 1 bis 6 hinsichtlich ihrer Angemessenheit und Nützlichkeit zu bewerten.

Die Auswertung der Schülerantworten erfolgt über einen Vergleich der Schülerurteile mit dem Urteil von Expert/innen. Dieser Vergleich greift nicht auf absolute Unterschiede im Urteil, sondern auf relative Aussagen zur Über- bzw. Unterlegenheit der Strategien zurück. Urteilen die Schüler/innen in Übereinstimmung mit dem Expertenurteil und bewerten eine angemessene und effektive Strategiealternative mit einer besseren Note als einen unangemessenen und ineffektiven Strategievorschlag, so erhalten sie einen Punkt, der in den Testsummenwert eingeht. Dieser Summenwert wird als Indikator für den metakognitiven Wissensstand interpretiert.

In der Entwicklungsphase der Strategiewissenstests wurde je nach Inhaltsgebiet ein Itempool aus sechs bis acht Szenarien mit jeweils bis zu sieben Strategiealternativen generiert.

Auf Grundlage der Validierung durch die Expertinnen wurden die Instrumente um nicht valide Strategievergleiche reduziert. Als nicht valide wurden Vergleiche zwischen Strategien angesehen, in denen das Expertenurteil keine eindeutige Überlegenheit der einen gegenüber der anderen Strategie zeigte. Im Anschluss wurden aus testökonomischen Gründen für jeden der vier untersuchten Inhaltsbereiche fünf Szenarien mit jeweils fünf bis sechs Strategiealternativen selegiert.

Metakognitiver Wissenstest im Bereich Lesen

Zur Erfassung des metakognitiven Wissens im Bereich *Lesen* wurde basierend auf Aufgaben zum Lesestrategiewissen aus PISA 2000 (vgl. Artelt/Schiefele/Schneider 2001; Schlagmüller/Schneider 2007) und weiteren Aufgaben, die in PISA 2009 eingesetzt wurden (vgl. Artelt u.a. 2009), eine der Altersgruppe angemessene Adaption und Ergänzung vorgenommen. Um die in den Szenarien umschriebenen Behaltens- und Verstehensanforderungen beim Lernen aus Texten zu erfüllen, wurden 28 Strategien vorgeschlagen. Die deutschlandweite Befragung von 19 Expert/innen aus der psychologischen Lernstrategieforschung ergab 38 inhaltlich valide Vergleiche zwischen diesen Strategien.

Metakognitiver Wissenstest im Bereich Mathematik

Zur Erfassung des metakognitiven Wissens im Bereich *Mathematik* wurde ein Instrument von Artelt (vgl. Ramm u.a. 2006) modifiziert und um altersangemessene Elemente erweitert. Es wurden 27 Strategien zu den vier Schritten des mathematischen Problemlöseprozesses vorgegeben (vgl. Garofalo/Lester 1985): Orientierung (Aufgabenverständnis und -repräsentation), Organisation (Planung der Lösungsstruktur), Ausführung (Überwachung und Regulation des Lösungsprozesses) und Evaluation der Lösung. Die 19 befragten Expert/innen aus mathematikdidaktischen Universitätsinstituten konnten in 30 Strategievergleichen klare Effektivitätsunterschiede ausmachen.

Metakognitiver Wissenstest im Bereich Englisch

Für den Bereich *Englisch* als Fremdsprache war es nicht möglich auf bereits bestehende Verfahren zurückzugreifen, da der Bedeutung von Strategien für den Fremdsprachener-

werb in der psychologischen Forschung bisher nur wenig Beachtung geschenkt wurde. Die 26 Strategievorschläge basieren auf theoretischen Erkenntnissen zur Wirksamkeit spezifischer Strategien für den Erwerb verschiedener schriftlicher und mündlicher Kompetenzen beim Erlernen einer Fremdsprache (Erwerb von Vokabular, Aussprache, Kommunikation, Textverstehen) (vgl. Cohen 1998). Die 17 befragten Expert/innen kamen in 32 Strategievergleichen zu klaren Präferenzurteilen.

Bereichsübergreifender metakognitiver Wissenstest zu Strategien des selbstregulierten Lernens

Zur Erfassung des zum *selbstregulierten Lernen* zur Verfügung stehenden Strategiewissens, das sich bereichsübergreifend auf Lernaktivitäten und Strategien in schultypischen Lernsituationen (Hausaufgaben, Vorbereitung auf eine Klassenarbeit) bezieht, wurden Szenarien mit 25 Strategien zu Kontrolle und Regulation in den Bereichen Motivation, Aufmerksamkeit, Arbeitsorganisation und Memorieren beschrieben. Daraus resultieren nach Befragung von 19 Expert/innen aus der psychologischen Lernstrategieforschung 27 valide Strategievergleiche.

3. Ergebnisse

Zur Überprüfung der psychometrischen Eigenschaften der metakognitiven Wissenstests werden Lage- und Dispersionsmaße sowie die interne Konsistenz der Skalen als Reliabilitätsindikator inspiziert. Hinweise auf Konstruktvalidität liefern Mittelwertsvergleiche zwischen den Schularten und Jahrgangsstufen.

3.1 Deskriptive Statistiken

Die in Tabelle 2 abgetragenen Lage- und Dispersionsmaße lassen nicht auf Decken- oder Bodeneffekte schließen. Die vier Skalen bilden also das über die Bildungsgänge und Jahrgangsstufen vorfindbare Leistungsspektrum im metakognitiven Wissen adäquat ab. Die auf den Paarvergleichen basierenden Scores bilden mit 27 (Lernen) bis 38 Items (Lesen) reliable Wissensindikatoren: Die internen Konsistenzen (Cronbachs Alpha) liegen mit .85 für Lesen, .87 für Mathematik, .75 für Englisch und .81 für selbstreguliertes Lernen alle im akzeptablen Bereich.

Rasch-Analysen belegen, dass sich die vier Skalen als näherungsweise rasch-homogene latente Konstrukte abbilden lassen (für nähere Ausführungen zu den metakognitiven Wissensskalen Lesen, Mathematik und selbstreguliertes Lernen s. Neuenhaus u.a. im Druck).

3.2 Unterschiede im deklarativen metakognitiven Wissen in Abhängigkeit von Schulart und Jahrgangsstufe

Um artifizielle Effekte von Schulform und Jahrgangsstufe aufgrund unterschiedlicher Zellbesetzungen auszuschließen, wurden univariate Varianzanalysen mit den metakognitiven Wissenstests als abhängige Variablen über Schulform und Jahrgangsstufe durchgeführt. Diese erbrachten lediglich signifikante Haupteffekte. Da sich keine signifikanten Interaktionseffekte zeigten, sind Mittelwertsvergleiche durch *t*-Tests interpretierbar (s. Tabelle 2): demnach verfügen Gymnasiast/innen über ein höheres metakognitives Wissen als die Hauptschüler/innen. Die Effektstärken (*d*) liegen durchgängig mit Werten über 1,0 im Bereich starker Effekte. Ebenso erweisen sich die Unterschiede zwischen den beiden untersuchten Jahrgangsstufen als bedeutsam. Demnach verfügen die Schüler/innen der 6. Jahrgangsstufe über ein höheres Strategiewissen als die Schüler/innen der 5. Jahrgangsstufe. Diese Effekte sind geringer als die der Schulform und liegen im Bereich kleiner bis moderater Effektstärken.

			N	***M***	***SD***	***T***	***df***	***p***	***d***
Lesen (0–38)	Schulform	Hauptschule	209	16,4	6,2	9,8	397	<.001	1,0
		Gymnasium	190	22,6	6,5				
	Jahrgangsstufe	5	223	18,0	6,9	4,6	397	<.001	0,5
		6	176	21,2	6,9				
Mathematik (0–30)	Schulform	Hauptschule	203	16,4	5,4	12,7	404	<.001	1,3
		Gymnasium	203	22,9	4,9				
	Jahrgangsstufe	5	230	18,6	6,3	4,4	398	<.001	0,4
		6	176	21,1	5,5				
Englisch (0–32)	Schulform	Hauptschule	215	12,7	4,2	11,8	391	<.001	1,2
		Gymnasium	178	17,8	4,3				
	Jahrgangsstufe	5	212	14,6	4,9	2,1	391	< .05	0,2
		6	181	15,6	5,0				
Lernen (0–27)	Schulform	Hauptschule	209	13,1	5,0	13,2	398	<.001	1,3
		Gymnasium	191	19,2	4,1				
	Jahrgangsstufe	5	219	15,0	5,4	4,1	398	<.001	0,4
		6	181	17,2	5,4				

Tab. 2: Deskriptive Statistiken der vier Strategiewissensskalen in Abhängigkeit von Schulform und Jahrgangsstufe (in Klammern der Wertebereich der eingesetzten Verfahren) sowie Kennwerte der Mittelwertsvergleiche

4. Diskussion

Zur Vorbereitung auf eine Längsschnittstudie am Beginn der Sekundarstufe I wurden Messinstrumente zur längsschnittlichen Erfassung des metakognitiven Wissens in Lesen, Mathematik, Englisch sowie selbstreguliertem Lernen für den Einsatz in der 5. und 6. Jahrgangsstufe entwickelt und auf ihre psychometrische Güte überprüft.

Die metakognitiven Wissenstests wurden teils völlig neu konstruiert (Englisch und selbstreguliertes Lernen), teils adaptiert (Mathematik und Lesen). Die Befunde einer Expertenbefragung lassen auf die inhaltliche Validität der Verfahren schließen. Die Pilotierungsergebnisse zeigen, dass die neu konstruierten bzw. adaptierten Verfahren eine reliable und das Leistungsspektrum im Wesentlichen abdeckende Messung metakognitiven Wissens gewährleisten. Zusätzlich durchgeführte Rasch-Analysen lassen die Annahme näherungsweise rasch-homogener Skalen zu. Mithin ist eine Anpassung des Schwierigkeitsniveaus zu späteren Messzeitpunkten durch die Einbeziehung neuer, für die Schüler/innen schwieriger zu beurteilender Aufgaben im Anker-Item-Design möglich.

In allen vier metakognitiven Wissenstests zeigen die Pilotierungsergebnisse deutliche Unterschiede zwischen den Schularten und Jahrgangsstufen: Gymnasiast/innen erzielen höhere Leistungen als Hauptschüler/innen und die Schüler/innen der 6. Jahrgangsstufe höhere Leistungen als die Schüler/innen der 5. Jahrgangsstufe. Die Unterschiede zwischen den Schularten fallen dabei größer aus als die zwischen den Jahrgangsstufen. Da Stichprobeneffekte im vorliegenden querschnittlichen Design nicht ausgeschlossen werden können, sind für weitergehende Interpretationen jedoch die in der Hauptuntersuchung geplanten längsschnittlichen Analysen abzuwarten.

Bereits verfügbare Resultate aus der ersten Erhebungswelle der Hauptuntersuchung lassen sich als Belege für die angenommene Domänenspezifität werten: trotz relativ hoher Zusammenhänge zwischen den fachspezifischen Skalen beschreibt eine domänenspezifische, mehrdimensionale Struktur das latente Konstrukt metakognitives Wissen besser (vgl. Neuenhaus u.a. im Druck). Die längsschnittlichen Analysen werden zeigen, ob der zu erwartende Zuwachs des metakognitiven Wissens in der Sekundarstufe I in Richtung einer Wissensgeneralisierung oder -differenzierung verläuft.

Insgesamt erlauben die konstruierten Verfahren die Erfassung des metakognitiven Wissenstandes über das gesamte Leistungsspektrum, das die Schüler/innen im Übergang zur Sekundarstufe I zeigen, und bilden somit eine solide Grundlage für die längsschnittliche Erhebung der Wissensentwicklung in diesem Altersbereich.

5. Praktische Implikationen

Die Befunde von EWIKO werden zeigen, inwieweit sich die Erkenntnisse aus der entwicklungspsychologischen bzw. grundlagenorientierten Forschungstradition zur Rolle des metakognitiven Wissens auf den Erwerb der schulischen Kernkompetenzen Lesen, Mathematik, Fremdspracherwerb und selbstreguliertes Lernen übertragen lassen. Die

fachbezogene Erfassung des Strategiewissens erlaubt darüber hinaus eine Auflösung nach kontextuellen Unterschieden zwischen diesen untersuchten Domänen.

Um abzuschätzen, wie groß der relative Beitrag metakognitiven Wissens zum schulischen Lernen ist, werden eine Vielzahl von theoretisch als bedeutsam für den schulischen Erfolg angesehenen Wirkfaktoren auf Schülerseite in die Untersuchung einbezogen und längsschnittlich abgebildet. Somit ergibt sich eine breit angelegte verlaufsdiagnostische Beobachtung der Schuljahre, die zwischen der in der Öffentlichkeit gut rezipierten Schulleistungsstudien liegen, die den Bildungserfolg am Ende von Bildungsphasen bilanzieren (z.B. IGLU am Ende der Primarstufe, PISA bzw. TIMSS am Ende der Sekundarstufe I bzw. Sekundarstufe II).

Die theoretisch bedeutsamen Erkenntnisse über die längsschnittlichen Wirkungen reichhaltigen metakognitiven Wissens in der Sekundarstufe lassen auch belastbare Aussagen über Relevanz und Nutzen von Programmen zum Aufbau dieser Kompetenz zu, deren grundsätzliche Förderbarkeit in einer Vielzahl von gut evaluierten Interventionsstudien belegt ist.

Die im Zuge der Untersuchung entwickelten Skalen zur verhaltensnahen Messung metakognitiven Wissens in der Sekundarstufe I füllen mit der ökonomischen, reliablen und validen Erfassung eines allgemein in seiner Bedeutung für die schulische Leistung anerkannten Konstruktes eine Lücke im Instrumentarium der (empirischen) Bildungsforschung. Daneben können die Instrumente Beiträge zur individualdiagnostischen Identifikation von Förderbedarf ebenso leisten wie zur ökonomischen und ökologisch validen Evaluation von Förderprogrammen.

Literatur

Annevirta, T./Laakkonen, E./Kinnunen, R./Vauras, M. (2007): Developmental dynamics of metacognitive knowledge and text comprehension skill in the first primary school years. In: Metacognition and Learning 2, S. 21–39.

Artelt, C. (2000): Strategisches Lernen. Münster: Waxmann.

Artelt, C./Beinicke, A./Schlagmüller, M./Schneider, W. (2009): Diagnose von Strategiewissen beim Textverstehen. In: Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie 41, S. 96–103.

Artelt, C./Schiefele, U./Schneider, W. (2001): Predictors of reading literacy. In: European Journal of Psychology of Education 16, S. 363–383.

Borkowski, J./Chan, L./Muthukrishna, N. (2000): A process-oriented model of metacognition: Links between motivation and executive functioning. In: Schraw, G./Impara, J. (Hrsg.): Issues in the Measurement of Metacognition. Lincoln, NE: Buros Institute of Mental Measurements, S. 1–41.

Carr, M./Biddlecomb, B. (1998): Metacognition in mathematics from a constructivist perspective. In: Hacker, D./Dunlosky, J./Graesser, A. (Hrsg.): Metacognition in educational theory and practice. Mahwah, NJ: Erlbaum, S. 69–91.

Cohen, A. (1998): Strategies in learning and using a second language. New York: Addison Wesley Longman.

Flavell, J. (1979): Metacognition and cognitive monitoring – A new area of cognitive-developmental inquiry. In: American Psychologist 34, S. 906–911.

Garofalo, J./Lester, F. (1985): Metacognition, Cognitive Monitoring and Mathematical Performance. In: Journal for Research in Mathematics Education 16, S. 163–176.
Helmke, A./Rindermann, H./Schrader, F.-W. (2008): Wirkfaktoren akademischer Leistungen in Schule und Hochschule. In: Schneider, W./Hasselhorn, M. (Hrsg.): Handbuch der Pädagogischen Psychologie. Göttingen: Hogrefe, S. 145–155.
Jacobs, J./Paris, S. (1987): Children's metacognition about reading: Issues in definition, measurement, and instruction. In: Educational Psychologist 22, S. 255–278.
Justice, E. (1986): Developmental changes in judgements of relative strategy effectiveness. In: British Journal of Developmental Psychology 4, S. 75–82.
Neuenhaus, N./Artelt, C./Lingel, K./Schneider, W. (im Druck): Fifth graders metacognitive knowledge: general or domain specific? In: European Journal of the Psychology of Education.
Pressley, M. (1994): Embracing the complexity of individual differences in cognition: Studying good information processing and how it might develop. In: Learning and Individual Differences 6, S. 259–284.
Pressley, M./Borkowski, J./Schneider, W. (1989): Good information processing: What it is and how education can promote it. In: International Journal of Educational Research 13, S. 857–867.
Pressley, M./McCormick, C. (1995): Advanced educational psychology for educators, researchers, and policymakers. New York: Harper Collins.
Ramm, G./Prenzel, M./Baumert, J./Blum, W./Lehmann, R./Leutner, D./Neubrand, M./Pekrun, R./Rolff, H.-G./Rost, J./Schielefe, U. (Hrsg.) (2006): Dokumentation der Erhebungsinstrumente. Münster: Waxmann.
Roeschl-Heils, A./Schneider, W./Van Kraayenoord, C. (2003): Reading literacy, metacognition and motivation: A follow-up study of German students in Grades 7 and 8. In: European Journal of the Psychology of Education 18, S. 75–86.
Samuelstuen, M./Bråten, I. (2007): Examining the validity of self-reports on scales measuring students' strategic processing. In: British Journal of Educational Psychology 77, S. 351–378.
Schneider, W. (2010): Metacognition and memory development in childhood and adolescence. In: Waters, H./Schneider, W. (Hrsg.): Metacognition, strategy use, and instruction. New York: Guilford Press, S. 54–83.
Schlagmüller, M./Schneider, W. (2007): Der Würzburger Lesestrategie-Wissenstest (WLST 6–12). Göttingen: Hogrefe.
Schlagmüller, M./Visé, M./Schneider, W. (2001): Zur Erfassung des Gedächtniswissens bei Grundschulkindern: Konstruktionsprinzipien und empirische Bewährung der Würzburger Testbatterie zum deklarativen Metagedächtnis. In: Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie 33, S. 91–102.
Spörer, N./Brunstein, J. (2006): Erfassung selbstregulierten Lernens mit Selbstberichtsverfahren – Ein Überblick zum Stand der Forschung. In: Zeitschrift für Pädagogische Psychologie 20, S. 147–160.
Van Kraayenoord, C./Schneider, W. (1999): Reading achievement, metacognition, reading self-concept and interest: A study of German students in grades 3 and 4. In: European Journal of the Psychology of Education 14, S. 305–324.

Anschrift der Autor/innen

Dipl.-Psych. Klaus Lingel, Lehrstuhl für Psychologie IV, Röntgenring 10, D-97070 Würzburg
E-Mail: lingel@uni-wuerzburg.de

Dipl.-Psych. Nora Neuenhaus, Otto-Friedrich-Universität Bamberg, Lehrstuhl für Empirische Bildungsforschung, Markusplatz 3, D-96045 Bamberg
E-Mail: nora.neuenhaus@uni-bamberg.de

Prof. Dr. Cordula Artelt, Otto-Friedrich-Universität Bamberg, Lehrstuhl für Empirische Bildungsforschung, Markusplatz 3, D-96045 Bamberg
E-Mail: cordula.artelt@uni-bamberg.de

Prof. Dr. Wolfgang Schneider, Lehrstuhl für Psychologie IV, Röntgenring 10, D-97070 Würzburg
E-Mail: schneider@psychologie.uni-wuerzburg.de

Jens Fleischer/Joachim Wirth/Stefan Rumann/Detlev Leutner

Strukturen fächerübergreifender und fachlicher Problemlösekompetenz

Analyse von Aufgabenprofilen

Projekt Problemlösen[1]

1. Einleitung

Die Fähigkeit zum Problemlösen stellt eine zentrale Qualifikation in verschiedensten schulischen und außerschulischen Lern- und Leistungsbereichen dar. Entsprechend wird Problemlösen in vielen Schulfächern, insbesondere in der Mathematik und den naturwissenschaftlichen Fächern, als zu erwerbende fachliche Kompetenz definiert, welche die Verfügbarkeit und Anwendung von Wissensbeständen einer spezifischen Fachdomäne erfordert (vgl. Blum u.a. 2006; KMK 2005). Darüber hinaus wird Problemlösen auch als fächerübergreifende Kompetenz betrachtet, bei der sich das zur Lösung notwendige Wissen nicht einer einzelnen spezifischen Fachdomäne zuordnen lässt (vgl. OECD 2004). Dieser fächerübergreifenden Problemlösekompetenz wird für beruflichen Erfolg eine zentrale Rolle beigemessen (vgl. ebd.).

Im Hinblick auf die Art einer Problemstellung lässt sich analytisches von dynamischem Problemlösen unterscheiden. Analytisches Problemlösen zeichnet sich dadurch aus, dass alle für die Problemlösung relevanten Informationen in der Problemstellung gegeben sind oder erschlossen werden können. Problemlösen kann damit zu einem großen Teil als schlussfolgernde Anwendung vorhandenen Wissens bezeichnet werden. Beim dynamischen Problemlösen hingegen muss ein Großteil der lösungsrelevanten Informationen in einer explorativen Interaktion mit der Problemsituation erst generiert werden (vgl. Leutner/Funke u.a. 2005).

Aus lehr-lernpsychologischer sowie fachdidaktischer Perspektive stellt sich vorrangig die Frage nach der Förderbarkeit der Problemlösekompetenz. Hierfür ist zunächst die allgemeine und fächerspezifische Struktur der Problemlösekompetenz aufzuklären, um darauf basierend geeignete Trainingsprogramme zur gezielten Förderung entwickeln zu können. Die Strukturfrage gewinnt spätestens seit den Ergebnissen der PISA-Studie 2003 Relevanz, die darauf hindeuten, dass fächerübergreifende analytische Problemlösekompetenz eine kognitive Ressource zum Aufbau fachlicher Problemlösekompetenz darstellt (vgl. Leutner u.a. 2004).

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: LE 645/12-1 und AR 301/8-1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

2. Zielsetzung und Arbeitsprogramm des Projekts

Die übergeordnete Zielsetzung des vorliegenden Projekts ist die Entwicklung und Validierung von Kompetenzstrukturmodellen des analytischen Problemlösens. Im Mittelpunkt steht die Frage, ob sich fächerübergreifende sowie fachspezifische Komponenten der Problemlösekompetenz in Abgrenzung zur Intelligenz identifizieren und inhaltlich beschreiben lassen.

Das Projekt gliedert sich in der ersten Zweijahresphase des DFG-Schwerpunktprogramms in zwei Schritte. Zur Identifizierung potenziell relevanter Komponenten der Problemlösekompetenz wurden in einem ersten Schritt die Anforderungen von Testaufgaben zur Erfassung fächerübergreifender sowie fachbezogener Problemlösekompetenz analysiert. Daran anschließend werden in einem zweiten Schritt neue Testverfahren zur Erfassung zentraler Komponenten der fächerübergreifenden und fachlichen Problemlösekompetenz entwickelt und administriert. Anhand dieser Daten soll die Binnenstruktur der Kompetenzen untersucht werden. Die Ergebnisse beider Projektschritte sollen in einem Kompetenzstrukturmodell fächerübergreifenden und fachlichen Problemlösens zusammengeführt werden.

Im Folgenden wird der erste Schritt des Projektes dargestellt. Wir konzentrieren uns dabei auf fächerübergreifendes und mathematisches Problemlösen. Aspekte des naturwissenschaftlichen Problemlösens werden im Rahmen eines assoziierten Projekts untersucht (s. Stawitz u.a. 2009).

3. Theoretischer Hintergrund

In der Psychologie herrscht weitgehende Einigkeit darüber, dass ein Problem aus einer Problemsituation (Ausgangszustand), einem mehr oder weniger genau definierten Zielzustand und einem nicht unmittelbar einsichtigen Lösungsweg besteht (vgl. Mayer 1990). Das Lösen von Problemen lässt sich damit definieren als „zielorientiertes Denken und Handeln in Situationen, für deren Bewältigung keine routinierten Vorgehensweisen verfügbar sind“ (Klieme u.a. 2001, S. 185). Vergleichbare Konzeptualisierungen der Begriffe „Problem“ und „Problemlösen“ finden sich auch in der Mathematik (vgl. Reiss/Törner 2007; Schoenfeld 1985).

3.1 Problemlösen als Prozess

Basierend auf Überlegungen der Gestaltpsychologie beschreibt Polya (1945) die Bearbeitung eines Problems als eine Abfolge von vier Prozessschritten: (1) understanding the problem, (2) devising a plan, (3) carrying out the plan und (4) looking back (i.S. von Bewertung der Problemlösung). Diese Prozessschritte liefern die Grundlage für neuere Ansätze sowohl zum fächerübergreifenden als auch zum fachlichen Problemlösen in der Mathematik. So beschreibt das PISA 2003 „Assessment Framework“ für den Test zur fächerübergreifenden Problemlösekompetenz eine vergleichbare Abfolge von Prozess-

schritten (vgl. OECD 2003). In der Mathematik unterscheiden Carlson und Bloom (2005) die Prozesse orientation, planning, executing und checking. Eine ähnliche Darstellung von Prozessschritten findet sich auch in der Beschreibung des mathematischen Modellierungskreislaufs, der die theoretische Basis für die Mathematiktests in den PISA-Studien darstellt (vgl. Blum u.a. 2004).

3.2 Problemlösen als Kompetenz

Problemlösekompetenz und Intelligenz

Im Rahmen der Strukturforschung zu kognitiven Fähigkeitskonstrukten wird Problemlösen mitunter als Komponente der Intelligenz betrachtet (vgl. Sternberg/Kaufmann 1998). Problemlösen im Sinne einer Kompetenz lässt sich jedoch konzeptuell von Intelligenz trennen: Problemlösekompetenz zeichnet sich im Gegensatz zu Intelligenz durch eine höhere Bereichsspezifität, prinzipielle Erlernbarkeit sowie eine stärkere Orientierung an den zu bewältigenden Anforderungen aus (vgl. Leutner/Funke u.a. 2005). Dennoch ist die Befundlage zum korrelativen Zusammenhang zwischen Problemlösen und Intelligenz nach wie vor uneinheitlich (vgl. Leutner 2002). Hinweise auf eine strukturelle Trennung zwischen Problemlösekompetenz und Intelligenz liefern jedoch Ergebnisse der PISA-Studien (vgl. Leutner u.a. 2004). Wie Leutner, Wirth, Klieme und Funke (2005) zeigen konnten, lassen sich die Leistungen auf den Skalen dynamisches Problemlösens, analytisches Problemlösens und Intelligenz nur sehr schlecht durch einen gemeinsamen Faktor erklären. Modelle mit drei latenten Dimensionen bilden die Kompetenzstruktur wesentlich besser ab.

Komponenten der Problemlösekompetenz

Im Verlauf des zuvor skizzierten Problemlöseprozesses lassen sich potenziell relevante Komponenten der Problemlösekompetenz identifizieren, die für eine erfolgreiche Lösung fächerübergreifender und fachlicher Probleme relevant sind. Hierzu gehört die Verfügbarkeit und Anwendung von inhaltsspezifischem Sach- und Handlungswissen (vgl. Schoenfeld 1985). Sachwissen ist definiert als Wissen über Objekte und Zustände, mit dessen Hilfe die Problemsituation und der angestrebte Zielzustand intern repräsentiert und damit erfasst werden können. Handlungswissen ist definiert als Wissen über Operationen zur Veränderung der Problemsituation und die Überführung dieser in den angestrebten Zielzustand sowie die Fähigkeit, eine kognitive Operation oder Handlung auch tatsächlich ausführen zu können (vgl. Süß 1996). Eine weitere Komponente stellt konditionales Wissen dar, welches die Umstände der Anwendung von Operationen beschreibt (vgl. Paris/Lipson/Wixson 1983). Darüber hinaus ist die Verfügbarkeit und die Anwendung von allgemeinen Problemlösestrategien und Heuristiken, welche die Suche nach relevanten Informationen, alternativen Problemrepräsentationen oder Teilzielen strukturieren, eine weitere Komponente der Problemlösekompetenz (vgl. Gick 1986). Hinzu kommt die Fähigkeit zur Selbstregulation, die notwendig ist, um Problemlöseprozesse zu planen, zu überwachen, zu bewerten und gegebenenfalls zu modifizieren (vgl. Davidson/Deuser/Sternberg 1994).

3.3 Fächerübergreifende und fachliche Problemlösekompetenz in PISA

Die beschriebenen konzeptuellen Ähnlichkeiten zwischen Problemlösen als fächerübergreifender Kompetenz und ihrer fachlichen Ausdifferenzierung in der Mathematik zeigt sich auch empirisch an vergleichsweise hohen Korrelationen dieser Kompetenzen in den PISA-Studien. Die messfehlerbereinigte Korrelation von mathematischer Kompetenz und (fächerübergreifender) Problemlösekompetenz liegt bei $r = .89$ (vgl. OECD 2005). Trotz dieser hohen Korrelation zeigen sich – relativiert am OECD-Durchschnitt der Kompetenzskalen – deutliche Niveauunterschiede zwischen der (im internationalen Vergleich hohen) Problemlösekompetenz und der (im internationalen Vergleich durchschnittlichen) mathematischen Kompetenz deutscher Schülerinnen und Schüler (vgl. Leutner u.a. 2004). Diese Diskrepanz lässt sich im Sinne der Potenzialausschöpfungshypothese interpretieren, wonach fächerübergreifende Problemlösekompetenz eine kognitive Ressource zum Aufbau fachlicher Kompetenzen darstellt, die im Unterricht jedoch nicht hinreichend ausgeschöpft zu werden scheint (vgl. ebd.).

Ergebnisse der PISA 2003-Messwiederholungstudie zur Untersuchung der Kompetenzentwicklung u.a. in Mathematik im Laufe der 10. Klasse unterstützen diese Vermutung. Die Ergebnisse zeigen, dass fächerübergreifende Problemlösekompetenz und fachlicher Kompetenzerwerb in engem Zusammenhang stehen, und deuten darauf hin, dass fächerübergreifende und fachliche Problemlösekompetenz aus verschiedenen, sich teilweise überlappenden Komponenten bestehen (vgl. Leutner/Fleischer/Wirth 2006).

4. Methodisches Vorgehen

Ziel des ersten Projektschritts ist es, möglichst umfassend strukturelle Gemeinsamkeiten und Unterschiede bezüglich der Konstruktion und der kognitiven Anforderungsprofile von Testaufgaben zur Erfassung fächerübergreifender und fachlicher Problemlösekompetenz zu identifizieren, um dadurch Hinweise auf relevante Komponenten der Problemlösekompetenz ableiten zu können. Hierfür wurden die Testaufgaben der PISA-Studie 2003 zum analytischen Problemlösen ($N = 28$ Analyseeinheiten) sowie eine nach Itemschwierigkeit, übergreifender Idee und Kompetenzcluster (s. OECD 2003) stratifizierte Auswahl von Aufgaben des Mathematiktests ($N = 58$ Analyseeinheiten) einer detaillierten Aufgaben- und Anforderungsanalyse unterzogen.

Nach wie vor liegt keine theoretisch hinreichend fundierte und empirisch überprüfte Taxonomie kognitiver Prozesse vor, die es ermöglichen würde, Testaufgaben nach kognitiven Anforderungen zu klassifizieren. Für den Bereich der fächerübergreifenden und der mathematischen Problemlösekompetenz existieren jedoch Arbeiten zu potenziell relevanten Aufgabenmerkmalen, welche für das Projekt genutzt werden konnten (vgl. z.B. Dossey/McCrone/O'Sullivan 2006; Jordan u.a. 2006). Zur Analyse der PISA-Aufgaben wurde ein Kategoriensystem zur Klassifikation von Aufgabenmerkmalen entwi-

ckelt und pilotiert. Die endgültige Version des Kategoriensystems enthält 36 detailliert beschriebene Beurteilungsfacetten (Beurteilungsdimensionen), die u.a. folgende Bereiche abdecken: Aufgabenstellung, Sprache, Aufgabenkontext, Themengebiet, curriculare Wissensstufe, erforderliche Wissensinhalte, kognitive Komplexität und Komplexität der Prozessschritte. Anhand dieses Kategoriensystems wurden die PISA-Aufgaben durch drei geschulte Beurteiler/innen unabhängig voneinander und ohne Kenntnis der jeweiligen Domäne (Mathematik vs. Problemlösen) nach den Beurteilungsfacetten klassifiziert.

Wir gehen davon aus, dass fächerübergreifende Problemlösekompetenz eine kognitive Ressource zum Aufbau fachlicher Problemlösekompetenz darstellt. Daher sollten die (fächerübergreifenden) Problemlöseaufgaben in den Facetten Operatoren-Anforderungsbereich, sprachliche und kognitive Komplexität sowie Komplexität der Prozessschritte einem höheren Anforderungsniveau zugeordnet werden. Wir erwarten außerdem, dass die Problemlöseaufgaben entsprechend ihres fächerverbindenden Charakters mehreren Themengebieten zugeordnet werden. Die Mathematikaufgaben sollten sich insbesondere durch einen stärkeren Formalisierungsgrad der Sprache, eine eindeutige Zuordnung zum Themengebiet der Mathematik sowie höhere Anforderungen an das inhaltsspezifische Sachwissen auszeichnen. Ergebnisse dieses ersten Projektschritts werden im folgenden Abschnitt zusammenfassend dargestellt.

5. Ergebnisse zur vergleichenden Analyse der PISA-Testaufgaben

Als Indikator für die Objektivität der Beurteilungsfacetten wurden die mittlere prozentuale Übereinstimmung, generalisierte Kappa-Koeffzienten sowie – für ordinal- und intervallskalierte Facetten – Krippendorfs Alpha über alle drei Beurteiler/innen berechnet (vgl. Fleiss 1971; Krippendorf 2004). Tabelle 1 stellt exemplarisch die Ergebnisse der Beurteilerübereinstimmung für ausgewählte Facetten über die gesamte Bandbreite des Kategoriensystems hinweg dar. Für weitere Analysen wurden Kappa-Koeffizienten von über .40 als Indikator für eine hinreichend hohe Beurteilerübereinstimmung betrachtet (vgl. Banerjee u.a. 1999).

Im Folgenden werden zentrale Ergebnisse der Aufgabenklassifikation für eine Auswahl von Beurteilungsfacetten zusammenfassend berichtet. Zur Prüfung von Unterschieden zwischen den Domänen wurden, je nach Skalenniveau der Facetten, Mann-Whitney *U*-Tests, Likelihood ratio chi^2-Tests sowie *t*-Tests für unabhängige Stichproben berechnet. Das Signifikanzniveau wurde auf 5% festgesetzt und eine sequentielle α-Fehler-Adjustierung für multiple Tests nach Holm (1979) vorgenommen.

Aufgabenstellung (Typ/Anforderungsniveau der Operatoren): Die Art der Aufgabenstellung unterscheidet sich signifikant zwischen Mathematik- und Problemlöseaufgaben ($\chi^2_{(5)} = 22.30$; $p = .001$). Während die Aufgabenstellung der Problemlöseaufgaben häufig aus einer Aufforderung besteht (67,9% der Fälle), besteht sie bei den Mathematikaufgaben häufig aus einer Frage (41,3%) oder aus einer Kombination aus Frage und Aufforderung (24,1%). Die in den Problemlöseaufgaben verwendeten Operatoren

Beurteilungsfacette	**Mittlere prozentuale Übereinstimmung**	**Kappa-Koeffizient**	**Krippendorfs *α***
Aufgabenstellung: Typ/Operatoren			
Typ (Frage, Aufforderung, Ergänzung etc.)	89,9%	0.87	
Operatoren-Anforderungsbereich (niedrig, mittel, hoch)	77,5%	0.60	0.71
Sprache: Komplexität/Formalisierungsgrad			
Komplexität (niedrig, mittel, hoch)	77,5%	0.67	0.86
Formalisierungsgrad (schwach, stark)	84,5%	0.65	0.65
Aufgabenkontext: Bezug/Funktion			
Bezug (unabhängig, persönlich, gesellschaftsbezogen etc.)	85,3%	0.79	
Funktion: Motivation (nicht erfüllt, erfüllt)	87,6%	0.75	0.75
Funktion: Modellbildung (nicht erfüllt, erfüllt)	85,3%	0.69	0.69
Funktion: Handlungsanleitung (nicht erfüllt, erfüllt)	94,6%	0.71	0.71
Themengebiet (Mathe, Physik, Chemie etc.)	96,1%	0.90	
Curriculare Wissensstufe (niedrig, mittel, hoch)	88,9%	0.78	0.83
Kognitive Komplexität (gering, mittel, hoch)	76,0%	0.63	0.78
Wissen: Art/Allgemeinheit			
Art (Sachwissen, Handlungswissen)	87,6%	0.71	
Sachwissen-Allgemeinheit/Niveau (Ratingskala 4-stufig)	65,5%	0.50	0.79
Handlungswissen-Allgemeinheit/Niveau (Ratingskala 4-stufig)	71,3%	0.59	0.81
Prozessschritte: Anforderungsniveau/ Komplexität			
Verstehen der Problemsituation (Ratingskala 3-stufig)	61,2%	0.31	0.46
Erkennen relevanter Bedingungen (Ratingskala 3-stufig)	76,7%	0.62	0.75
Problem-/Aufgabenbearbeitung (Ratingskala 3-stufig)	73,6%	0.60	0.71
Interpretieren/Validieren (nicht erforderlich, erforderlich)	96,9%	0.91	0.91

Tab. 1: Koeffizienten der Beurteilerübereinstimmung (mittlere prozentuale Übereinstimmung, generalisiertes Kappa und Krippendorfs α) für ausgewählte Beurteilungsfacetten des Kategoriensystems

(Denk- und Handlungsaufforderungen) lassen sich außerdem einem signifikant höheren kognitiven Anforderungsniveau zuordnen als bei den Mathematikaufgaben ($U = 473.5$; $p < .001$).

Sprache (Komplexität/Formalisierungsgrad): Problemlöseaufgaben weisen eine signifikant höhere sprachlogische Komplexität im Vergleich zu Mathematikaufgaben auf ($U = 310.5$; $p = <.001$). Allerdings haben Mathematikaufgaben einen signifikant höheren Formalisierungsgrad der Sprache (Fachsprache und formale Darstellung) als Problemlöseaufgaben ($U = 563.0$; $p = .002$).

Aufgabenkontext (Bezug/Funktion)/Themengebiet: Problemlöse- und Mathematikaufgaben unterscheiden sich signifikant hinsichtlich des persönlichen Bezugs des Aufgabenkontextes ($\chi^2_{(4)} = 41.00$; $p < .001$). Während Problemlöseaufgaben in den meisten Fällen einen Kontext mit persönlichem Bezug (Freizeit, Familie, Beruf, Schule) zu den Aufgabenbearbeitenden aufweisen (89,3%), ist dies bei den Mathematikaufgaben seltener der Fall (37,9%). Bezüglich der Funktion des Aufgabenkontextes (Motivation, Modellbildung, Handlungsanleitung) zeigen sich keine signifikanten Unterschiede. Die überwiegende Mehrheit der Mathematikaufgaben lässt sich eindeutig dem Themengebiet der Mathematik zuordnen (93,1%), während dies nur für 32,1% der Problemlöseaufgaben zutrifft, die in den meisten Fällen (57,1%) keinem Themengebiet eindeutig zugeordnet werden können ($\chi^2_{(2)} = 35.79$; $p < .001$).

Curriculare Wissensstufe/kognitive Komplexität: Die Problemlöseaufgaben unterscheiden sich signifikant von den Mathematikaufgaben hinsichtlich des curricularen Niveaus des zur Lösung der Aufgaben notwendigen Wissens ($U = 251.5$; $p < .001$). Mathematikaufgaben erfordern in der Regel mindestens einfaches Wissen der Sekundarstufe I (84,2%). Demgegenüber erfordern die Problemlöseaufgaben in den meisten Fällen lediglich Grundkenntnisse (81,5%), die teilweise bereits in der Grundschule vermittelt wurden. Dennoch ist die kognitive Komplexität (u.a. Notwendigkeit zum strategisch planvollen Vorgehen) der Problemlöseaufgaben signifikant höher als die der Mathematikaufgaben ($U = 418.0$; $p < .001$).

Wissen (Art/Allgemeinheit): Problemlöse- und Mathematikaufgaben unterscheiden sich bezüglich der Art des zur Lösung notwendigen Wissens signifikant voneinander ($\chi^2_{(2)} = 15.78$; $p = .001$). Bei Problemlöseaufgaben steht häufiger Handlungswissen im Vordergrund (96,4%) als bei Mathematikaufgaben (60,3%), bei denen außerdem häufiger eine Kombination von Sach- und Handlungswissen erforderlich ist (31,0%) als bei Problemlöseaufgaben (3,6%). Das zur Lösung der Problemlöseaufgaben notwendige Sachwissen ist im Vergleich zu den Mathematikaufgaben allgemeiner ($t_{adj(83)} = -5.84$; $p < .001$), während bezüglich des Allgemeinheitsgrads des notwendigen Handlungswissens kein statistisch signifikanter Unterschied festgestellt werden kann.

Prozessschritte (Anforderungsniveau/Komplexität): Problemlöseaufgaben stellen höhere Anforderungen an das Erkennen relevanter ggf. einschränkender Bedingungen der Aufgabenstellung ($t_{(84)} = -3.72$; $p < .001$). Bezüglich der Anforderungen an die Prozesse Aufgabenbearbeitung und Validierung der Aufgabenlösung vor dem Hintergrund der Ausgangssituation lassen sich keine signifikanten Unterschiede feststellen. Problemlöseaufgaben scheinen zwar ebenfalls höhere Anforderungen an das Verstehen der

Problemsituation zu stellen, jedoch ist es angebracht, die betreffende Facette aufgrund geringer Beurteilerübereinstimmung zu überarbeiten.

6. Diskussion und Ausblick

Die Aufgabenklassifikation zeigt, dass der persönliche Bezug der beschriebenen Problemstellungen ein charakteristisches Merkmal der fächerübergreifenden Problemlöseaufgaben ist. In der PISA-Studie wurde fächerübergreifende Problemlösekompetenz über fächerverbindende Aufgaben operationalisiert, was in der Aufgabenklassifikation dazu hätte führen sollen, dass jede Problemlöseaufgabe mehreren Fach- bzw. Themengebieten zugeordnet wird. Dieser fächerverbindende Charakter der Aufgaben zeigt sich jedoch nicht. Die überwiegende Mehrheit der Problemlöseaufgaben wurde entweder ausschließlich der Mathematik oder gar keinem Themengebiet zugeordnet (89,2%).

Mathematikaufgaben sind stärker formalisiert und stellen damit höhere Anforderungen an die Fähigkeit zu Dekodierung der lösungsrelevanten Informationen. Hierfür ist inhaltsspezifisches Sachwissen notwendig, was sich auch am höheren curricularen Niveau des zur Lösung der Aufgaben notwendigen Wissens zeigt. Im Gegensatz dazu scheint bei den Problemlöseaufgaben eher Handlungswissen eine relevante Komponente zu sein.

Das relativ hohe Anforderungsniveau der in den Problemlöseaufgaben verwendeten Operatoren sowie die höhere sprachlogische Komplexität der Aufgaben deuten auf das kognitive Potenzial hin, welches sich bei der erfolgreichen Bearbeitung fächerübergreifender Problemlöseaufgaben zu manifestieren scheint. Insbesondere die höheren Anforderungen an das Erkennen lösungsrelevanter Bedingungen und an das planvolle und strategische Vorgehen deuten auf die Relevanz konditionalen Wissens und der Fähigkeit zur Selbstregulation hin. Dies scheinen eher relevante Komponenten der fächerübergreifenden und weniger der fachspezifischen Problemlösekompetenz zu sein.

Die bisherigen Ergebnisse des ersten Projektschritts zeigen strukturelle Gemeinsamkeiten und Unterschiede bezüglich der Konstruktion und der kognitiven Anforderungen der PISA-Testaufgaben zur fächerübergreifenden und zur mathematischen Problemlösekompetenz. Auf Basis dieser Ergebnisse werden im zweiten Schritt des laufenden Projekts neue Messinstrumente zur Erfassung relevanter Komponenten der fächerübergreifenden und fachlichen Problemlösekompetenz entwickelt und eingesetzt, um so die Binnenstruktur der Kompetenzen aufklären zu können. Die Ergebnisse beider Projektschritte sollen in einem Kompetenzstrukturmodell fächerübergreifenden und fachlichen Problemlösens zusammengeführt werden.

In der kommenden zweiten Förderphase des Schwerpunktprogramms soll die Ausprägung auf einzelnen Komponenten der Problemlösekompetenz experimentell variiert werden, um Ursache-Wirkungsrelationen prüfen zu können. Dadurch soll das Kompetenzstrukturmodell experimentell validiert und zu einem Entwicklungsmodell ausgebaut werden.

Literatur

Banerjee, M./Capozzoli, M./McSweeney, L./Sinha, D. (1999): Beyond kappa: A review of interrater agreement measures. In: Canadian Journal of Statistics 27, S. 3–23.

Blum, W./Drüke-Noe, C./Hartung, R./Köller, O. (Hrsg.) (2006): Bildungsstandards Mathematik: konkret. Berlin: Cornelson.

Blum, W./Neubrand, M./Ehmke, T./Senkbeil, M./Jordan, A./Ulfig, F./Carstensen, C. (2004): Mathematische Kompetenz. In: PISA-Konsortium Deutschland (Hrsg.): PISA 2003: Der zweite Vergleich der Länder in Deutschland – Was wissen und können Jugendliche? Münster: Waxmann, S. 47–92.

Carlson, M.P./Bloom, I. (2005): The cyclic nature of problem solving: An emergent multidimensional problem-solving framework. In: Educational Studies in Mathematics 58, S. 45–75.

Davidson, J.E./Deuser, R./Sternberg, R.J. (1994): The role of metacognition in problem solving. In: Metcalfe, J./Shimamura, A.P. (Hrsg.): Metacognition: knowing about knowing. Cambridge, MA: MIT Press, S. 207–226.

Dossey, J.A./McCrone, S.A./O'Sullivan, C. (2006): Problem solving in the PISA and TIMSS 2003 assessments (NCES 2007-049). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Fleiss, J.L. (1971): Measuring nominal scale agreement among many raters. In: Psychological Bulletin 76, S. 378–382.

Gick, M.L. (1986): Problem-solving strategies. In: Educational Psychologist 21, S. 99–120.

Holm, S. (1979): A simple sequentially rejective multiple test procedure. In: Scandinavian Journal of Statistics 6, S. 65–70.

Jordan, A./Ross, N./Krauss, S./Baumert, J./Blum, W./Neubrand, M./Löwen, K./Brunner, M./Kunter, M. (2006): Klassifikationsschema für Mathematikaufgaben: Dokumentation der Aufgabenkategorisierung im COACTIV-Projekt. Berlin: Max-Planck-Institut für Bildungsforschung.

Klieme, E./Funke, J./Leutner, D./Reimann, P./Wirth, J. (2001): Problemlösen als fächerübergreifende Kompetenz? Konzeption und erste Resultate aus einer Schulleistungsstudie. In: Zeitschrift für Pädagogik 47, S. 179–200.

KMK (2005): Beschlüsse der Kultusministerkonferenz – Bildungsstandards im Fach Chemie für den mittleren Schulabschluss (Beschluss von 16. Dezember 2004). München: Wolters Kluwer.

Krippendorf, K. (2004): Content analysis. An introduction to its methodology. Thousand Oaks, CA: SAGE.

Leutner, D. (2002): The fuzzy relationship of intelligence and problem solving in computer simulations. In: Computers in Human Behavior 18, S. 685–697.

Leutner, D./Fleischer, J./Wirth, J. (2006): Problemlösekompetenz als Prädiktor für zukünftige Kompetenz in Mathematik und in den Naturwissenschaften. In: Prenzel, M./Baumert, J./Blum, W./Lehmann, R./Leutner, D./Neubrand, M./Pekrun, R./Rost, J./Schiefele, U. (Hrsg.): PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres. Münster: Waxmann, S. 119–137.

Leutner, D./Funke, J./Klieme, E./Wirth, J. (2005): Problemlösefähigkeit als fächerübergreifende Kompetenz. In: Klieme, E./Leutner, D./Wirth, J. (Hrsg.): Problemlösekompetenz von Schülerinnen und Schülern. Diagnostische Ansätze, theoretische Grundlagen und empirische Befunde der deutschen PISA-2000-Studie. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 11–19.

Leutner, D./Klieme, E./Meyer, K./Wirth, J. (2004): Problemlösen. In: PISA-Konsortium Deutschland (Hrsg.): PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs. Münster: Waxmann, S. 147–175.

Leutner, D./Wirth, J./Klieme, E./Funke, J. (2005): Ansätze zur Operationalisierung und deren Erprobung im Feldtest zu PISA 2000. In: Klieme, E./Leutner, E./Wirth, J. (Hrsg.): Problemlösekompetenz von Schülerinnen und Schülern. Diagnostische Ansätze, theoretische Grundlagen und empirische Befunde der deutschen PISA-2000-Studie. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 21–36.
Mayer, R.E. (1990): Problem solving. In: Eysenck, M.W. (Hrsg.): The Blackwell dictionary of cognitive psychology. Oxford, England: Basil Blackwell, S. 284–288.
OECD (2003): The PISA 2003 assessment framework – Mathematics, reading, science and problem solving knowledge and skills. Paris: OECD.
OECD (2004): Problem solving for tomorrow's world. First measurements of cross-curricular competencies from PISA 2003. Paris: OECD.
OECD (2005): PISA 2003 – Technical report. Paris: OECD.
Paris, S.G./Lipson, M.Y./Wixson, K.K. (1983): Becoming a strategic reader. In: Contemporary Educational Psychology 8, S. 293–316.
Polya, G. (1945): How to solve it. Princeton, NJ: Princeton University Press.
Reiss, K./Törner, G. (2007): Problem solving in the mathematics classroom: The German perspective. In: ZDM – The International Journal on Mathematics Education 39, S. 431–441.
Schoenfeld, A.H. (1985): Mathematical problem solving. Orlando: Academic Press.
Stawitz, H./Rumann, S./Fleischer, J./Wirth, J. (2009): Vergleich von Aufgabenmerkmalen in Large-Scale Assessments. In: Höttecke, D. (Hrsg.): Chemie- und Physikdidaktik für die Lehramtsausbildung. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Schwäbisch-Gmünd 2008. Berlin: LIT, S. 386–388.
Sternberg, R.J./Kaufman J.C. (1998): Human abilities. In: Annual Review of Psychology 49, S. 479–502.
Süß, H.-M. (1996): Intelligenz, Wissen und Problemlösen. Göttingen: Hogrefe.

Anschrift der Autoren

Dipl. Psych. Jens Fleischer, Universität Duisburg-Essen, FB Bildungswissenschaften, Lehrstuhl für Lehr-Lernpsychologie, Universitätsstr. 2, D-45117 Essen
E-Mail: jens.fleischer@uni-due.de

Prof. Dr. Joachim Wirth, Ruhr-Universität Bochum, Fakultät für Philosophie und Erziehungswissenschaft, Lehrstuhl für Lehr-Lernforschung, Universitätsstr. 150, D-44721 Bochum
E-Mail: joachim.wirth@rub.de

Prof. Dr. Stefan Rumann, Universität Duisburg-Essen, FB Chemie, Institut für Didaktik der Chemie, Schützenbahn 70, D-45127 Essen
E-Mail: stefan.rumann@uni-due.de

Prof. Dr. Detlev Leutner, Universität Duisburg-Essen, FB Bildungswissenschaften, Lehrstuhl für Lehr-Lernpsychologie, Universitätsstr. 2, D-45117 Essen
E-Mail: leutner@uni-due.de

Melanie Schütte/Joachim Wirth/Detlev Leutner

Selbstregulationskompetenz beim Lernen aus Sachtexten

Entwicklung und Evaluation eines Kompetenzstrukturmodells

Projekt Selbstregulationskompetenz[1]

1. Ausgangsfrage und Zielsetzung des Projekts

Die Fähigkeit, den eigenen Lernprozess selbst zu planen, zu überwachen und zu steuern, gehört zu den zentralen Kompetenzen, die Schülerinnen und Schüler im Laufe ihrer Schullaufbahn erwerben und nutzen müssen. Folglich ist die Vermittlung dieser Schlüsselkompetenz Ziel der Bildungssysteme, und sie findet ihre Verankerung in den Lehrplänen und Bildungsstandards (vgl. Artelt/Baumert/Julius-McElvany 2003). Ihre Bedeutsamkeit zeigt sich ebenfalls daran, dass selbstreguliertes Lernen neben den fachgebundenen Kompetenzen als fächerübergreifende (im Sinne von „in allen Fächern notwendige") Kompetenz bereits im ersten Zyklus der PISA-Studie als internationale Option erfasst wurde.

Selbstreguliertes Lernen kann einerseits als fächerübergreifende Fähigkeit angesehen werden, welche in unterschiedlichen Bereichen des Lernens notwendig ist. Gleichzeitig ist es aber insofern fachspezifisch, als dass das selbstregulierte Lernen in verschiedenen Bereichen an unterschiedliche Voraussetzungen und Anforderungen gebunden ist. Unter Berücksichtigung der Unterschiede in den Anforderungen beim selbstregulierten Lernen fokussiert das vorliegende Projekt exemplarisch auf das selbstregulierte Lernen aus Sachtexten und die dafür notwendigen Kompetenzen.

Zur Beschreibung des selbstregulierten Lernens wurde in den letzten 30 Jahren eine Vielzahl an Modellen entwickelt (vgl. z.B. Boekaerts 1999; Schreiber 1998; Winne/Hadwin 1998; Zimmerman 2000). Sie lassen sich in zwei Klassen unterteilen (vgl. Thillmann 2008), wobei sich die Zugehörigkeit eines Modells zu einer Klasse danach richtet, ob es die notwendigen Kompetenzen als Voraussetzung zum Lernen (*Komponentenmodelle*; vgl. Boekaerts 1999) oder die prozessualen Anforderungen während des Lernens (*Prozessmodelle*; vgl. Winne/Hadwin 1998; Zimmerman 2000) fokussiert.

Ziel der Komponentenmodelle ist es, die Gesamtheit der Teilkompetenzen zu beschreiben, welche notwendig sind, um den aufgabenbezogenen Lernanforderungen gerecht zu werden. Ein Beispiel hierfür ist das Sechs-Komponenten-Modell von Boekaerts (1999), welches sowohl kognitive als auch motivationale Kompetenzen für das Lernen definiert.

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: XY 1234/5-6) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen" (SPP 1293).

Kritisch zu betrachten ist hierbei jedoch die fehlende Beschreibung des genauen Zeitpunktes, an welchem die einzelnen Kompetenzen während des Lernens benötigt werden.

Prozessmodelle hingegen fokussieren auf die prozessualen Anforderungen des Lernens. Sie unterteilen das Lernen in Phasen und benennen für jede Phase Anforderungen, welche von Schülerinnen und Schülern erfüllt werden müssen. Ein prominentes Beispiel für Prozessmodelle ist das Phasenmodell von Zimmerman (2000). Zimmerman postuliert in diesem Modell einen zyklischen Lernprozess mit drei Phasen: Die *Forethought*-Phase zu Beginn des Lernens als Voraussetzung für den aktiven Lernprozess, die *Performance*-Phase während des aktiven Lernens sowie die *Self-reflection*-Phase am Ende eines jeden Lernprozesses zur Reflektion des vorangegangenen aktiven Lernens. Ein Problem der Prozessmodelle besteht jedoch darin, dass nicht direkt die Teilkompetenzen benannt werden, welche während der Phasen des Lernens benötigt werden, um den bestehenden Anforderungen gerecht zu werden.

Weder Komponenten- noch Prozessmodelle beschreiben umfassend, welche Kompetenzen in welcher Phase des Lernprozesses notwendig sind. Zudem existiert unserem Wissen nach noch kein Modell, das die Struktur dieser Teilkompetenzen und ihrer wechselseitigen Beziehungen angemessen abbildet. Deshalb wird im Projekt der Frage nachgegangen, wie sich die Struktur der relevanten Teilkompetenzen für ein erfolgreiches selbstreguliertes Lernen (aus Sachtexten) abbilden lässt. Zusätzlich werden – basierend auf der modellierten Kompetenzstruktur – Kompetenzniveaus für das selbstregulierte Lernen aus Sachtexten definiert.

Zur Modellierung der Kompetenzstruktur müssen in einem ersten Schritt die relevanten Teilkompetenzen des selbstregulierten Lernens aus Sachtexten bestimmt werden. Dafür wurden zunächst die Anforderungen identifiziert, für deren Bewältigung die Teilkompetenzen des selbstregulierten Lernens notwendig sind. Auf Basis einer Analyse beider Modellklassen wurden fünf zentrale Anforderungen beim selbstregulierten Lernen aus Sachtexten identifiziert (vgl. Wirth/Leutner 2008): (1) das Setzen von Zielen und Standards, (2) das Erstellen eines Handlungsplans, (3) das Beobachten des eigenen Lernvorgehens, (4) das Bewerten des eigenen Lernvorgehens sowie des Lernergebnisses und (5) das Reagieren im Falle von Diskrepanzen zwischen geplantem und beobachtetem Vorgehen bzw. zwischen geplantem und beobachtetem Lernergebnis. Für die Bewältigung dieser Anforderungen werden verschiedene Teilkompetenzen benötigt, welche zusammengefasst die Struktur der Selbstregulationskompetenz bilden. Einen Überblick über die bestehenden Anforderungen sowie die benötigten Teilkompetenzen während des selbstregulierten Lernens aus Sachtexten bietet Abbildung 1.

Zum *Setzen geeigneter Ziele und Standards* müssen Schülerinnen und Schüler in der Lage sein, (a) die Lernaufgabe in Bezug auf ihre Anforderungen und ihre Randbedingungen zu analysieren und unter Berücksichtigung des eigenen Wissens und der eigenen Fähigkeiten realistisch einzuschätzen, (b) ihr aufgabenrelevantes Vorwissen und ihre aufgabenrelevanten Fähigkeiten zu aktivieren und realistisch einzuschätzen und (c) auf dieser Basis geeignete, für erfolgreiches Lernen erforderliche und hilfreiche Ziele und Standards zu formulieren. Um darauf aufbauend ihre *Lernhandlung planen* zu können, müssen Schülerinnen und Schüler fähig sein, (d) aus dem ihnen verfügbaren Reper-

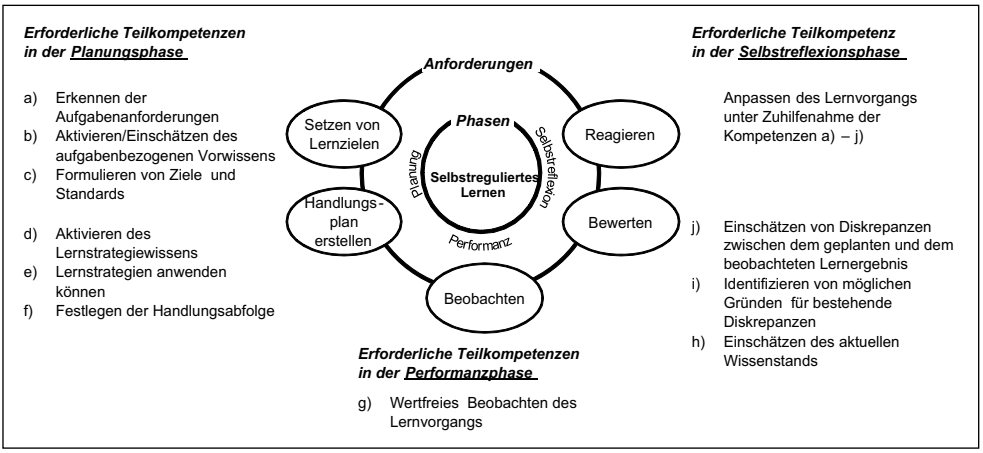


Abb. 1: Teilkompetenzen des selbstregulierten Lernens aus Sachtexten

toire an Lernstrategien die geeigneten auszusuchen, (e) die ausgesuchten Lernstrategien in der konkreten Lernsituation adäquat einzusetzen und (f) die Lernhandlungen im Sinne eines Handlungsplans in eine sinnvolle Reihenfolge zu bringen. Das *Beobachten* des eigenen Lernvorgehens dient der Feststellung des aktuellen Ist-Zustandes. Kompetent selbstregulierte Schülerinnen und Schüler führen diese Selbstbeobachtung kontinuierlich und in zeitlicher Nähe zum Lernen durch. Dabei sind sie fähig, (g) ihr Lernvorgehen akkurat und wertfrei zu beobachten und die spezifischen situationalen und persönlichen Bedingungen, unter denen das Lernen stattfindet, mit einzubeziehen. Die *Bewertung* des eigenen Lernvorgehens sowie des Lernergebnisses erfordert von den Schülerinnen und Schülern, (h) ihren Wissensstand nach dem Lernen einzuschätzen, (i) Diskrepanzen zwischen dem geplanten und beobachteten Lernvorgehen bzw. Lernergebnis objektiv, akkurat und umfassend einzuschätzen sowie, im Falle von entdeckten Diskrepanzen, (j) Gründe für deren Auftreten zu identifizieren. Das *Reagieren* auf Diskrepanzen erfordert wiederum eine oder mehrere der Kompetenzen (a) bis (j), wodurch der zyklische Charakter der Selbstregulation deutlich wird.

Das Projekt untersucht die Struktur der Fähigkeit zum selbstregulierten Lernen aus Sachtexten, die durch die Teilkompetenzen (a) bis (j) und ihre wechselseitigen Beziehungen gebildet wird. Im nachfolgenden Abschnitt werden das dabei verwendete Design sowie die Operationalisierungen der Teilkompetenzen dargestellt.

2. Projektdesign

Das Projekt ist zweischrittig angelegt. Zur Erfassung und Modellierung der oben beschriebenen Teilkompetenzen des selbstregulierten Lernens aus Sachtexten wurden im Zeitraum von April 2008 bis Dezember 2008 adäquate Testverfahren nach dem unter Kapitel 2.2 beschriebenen Prinzip entwickelt und im Rahmen von Pilotstudien evalu-

iert. In einem zweiten Schritt werden zurzeit die Teilkompetenzen von Schülerinnen und Schülern der neunten Jahrgangsstufe mit diesen Verfahren getestet, um basierend auf den gewonnenen Daten die Kompetenzstruktur beim Lernen aus Sachtexten modellieren und Kompetenzniveaus definieren zu können. Dabei bearbeiten die Schülerinnen und Schüler zunächst einen Sachtext mit dem Ziel, seinen Inhalt selbstreguliert zu erlernen. In einem Prä-Posttestdesign wird ihr jeweiliger Lernzuwachs ermittelt. Danach werden bei der Bearbeitung eines zweiten, vergleichbaren Sachtextes die Teilkompetenzen erfasst, die für ein erfolgreiches selbstreguliertes Lernen aus Sachtexten erforderlich sind. Diese Teilkompetenzen werden dann zueinander und zu dem Lernzuwachs aus dem ersten Sachtext in Beziehung gesetzt, um so die Struktur der Selbstregulationskompetenz abbilden zu können.

2.1 Konzeption der Sachtexte

Die Grundlage zur Erfassung und Modellierung der Kompetenzen bilden zwei anspruchsvolle Sachtexte aus dem naturwissenschaftlichen Bereich. Hierbei handelt es sich um einen Sachtext zum Thema „Blitze“ (vgl. Schmidt-Weigand 2006) sowie einen Sachtext zum Thema „Wasser“ (vgl. Leopold/den Elzen-Rump/Leutner 2006). In Zusammenarbeit mit Linguist/innen der Universität Duisburg-Essen wurden beide Sachtexte basierend auf Erkenntnissen der Verständlichkeitsforschung modifiziert (vgl. Christmann/Groeben 1999), um ein vergleichbares, das selbstregulierte Lernen stimulierendes Niveau zu gewährleisten.

2.2 Operationalisierung der Teilkompetenzen

Die Konstruktion der Testmaterialien zur Erfassung der jeweiligen Teilkompetenz basiert auf dem Prinzip metakognitiver Vergleichsprozesse (vgl. Wirth/Leutner 2008). Ansätze zum selbstregulierten Lernen postulieren unter dem Stichwort „Monitoring“, dass Schülerinnen und Schüler für ein erfolgreiches Lernergebnis während des Lernprozesses stets Vergleiche zwischen dem aktuellen und dem gewünschten Lernzustand bzw. Lernverhalten ziehen müssen (vgl. Butler/Winne 1995). Damit stellt die Fähigkeit, Vergleiche zu ziehen, eine zentrale Komponente der Selbstregulation dar, die sich in allen Teilkompetenzen widerspiegelt.

Zur Erfassung dieser Vergleichsprozesse wird im vorliegenden Projekt für jede der neun Teilkompetenzen jeweils ein Testpaar entwickelt. Jedes Testpaar besteht aus einem subjektiven Verhalten (z.B. Selbsteinschätzung des verfügbaren aufgabenrelevanten Vorwissens) sowie aus einem objektiven Kriterium (z.B. Leistung bei einem Testverfahren zur Erfassung des tatsächlich verfügbaren aufgabenrelevanten Vorwissens). Die Ausprägung der erfassten Teilkompetenz wird dann über den Vergleich zwischen subjektivem Verhalten und dem objektiven Kriterium ermittelt. Die Umsetzung dieses Prinzips ist exemplarisch für drei Teilkompetenzen dem folgenden Abschnitt zu entnehmen.

3. Ausgewählte Ergebnisse der Evaluationsstudien

Im Folgenden wird die Konzeption der Testverfahren basierend auf dem Prinzip metakognitiver Vergleichsprozesse für drei der insgesamt neun zu erfassenden Teilkompetenzen dargestellt, und ausgewählte Ergebnisse der entsprechenden Evaluationsstudien werden präsentiert. Die exemplarische Auswahl der Teilkompetenzen erfolgte in Anlehnung an die drei von Zimmerman (2000) postulierten Phasen. Die Formulierung geeigneter Ziele und Standards (Kompetenz c) ist eine zentrale Anforderung während der *Forethought*-Phase. Die Anwendung geeigneter Lernstrategien (Kompetenz e) ist während der *Performance*-Phase erforderlich. Zur Bewältigung der letzten Phase (*Self-reflection*-Phase) ist u.a. die Kompetenz (h) notwendig, welche die adäquate Einschätzung des eigenen Wissensstandes zur Bewertung des eigenen Lernerfolgs nach dem Lernen umfasst. Exemplarisch beschränken sich die folgenden Darstellungen auf die Bearbeitung des Sachtextes „Wasser“, wobei vergleichbare Ergebnisse auch bei der Bearbeitung des Sachtextes „Blitze“ erzielt wurden.

3.1 Ziele und Standards formulieren (Kompetenz c)

In Abhängigkeit von der subjektiven Einschätzung der Aufgabenanforderungen (Teilkompetenz a) und der Einschätzung des aufgabenrelevanten Vorwissens (Teilkompetenz b) besteht vor dem Lesen eines anspruchsvollen Sachtextes eine weitere zentrale Aufgabe darin zu entscheiden, auf welche inhaltlichen Aspekte des Textes beim Lernen speziell fokussiert werden soll (vgl. Schreiber 1998; Winne/Hadwin 1998). Kompetent selbstregulierte Schülerinnen und Schüler setzen sich hierbei vor dem Lernen Lernziele, die auf bestehende Wissenslücken ausgerichtet sind. Gute Lernziele zeichnen sich zudem dadurch aus, dass sie realistisch, aufgabenspezifisch, in überschaubarer Zeit erreichbar, konkret und herausfordernd (aber nicht überfordernd) sind und ihr Erreichen eine Belohnung für die Lernanstrengungen darstellt (vgl. Locke/Latham 1990).

Subjektives Verhalten. Zur subjektiven Erfassung der Kompetenz, sich adäquate Lernziele zu setzen, wird den Schülerinnen und Schülern eine Liste mit 19 Lernzielen vorgegeben. Die Schülerinnen und Schüler werden aufgefordert, Ziele auszuwählen, welche sie in einem vorgegebenen Zeitraum durch das Lesen des Sachtextes erreichen möchten. Zehn der Ziele thematisieren konkrete Inhalte des Sachtextes. Insofern stellen diese potenziell gute Lernziele dar, falls der jeweilige konkrete Inhalt vorab bei der Einschätzung des eigenen Vorwissens (Teilkompetenz b) als Wissenslücke identifiziert wurde. Die übrigen Lernziele der Lernzielliste repräsentieren globale, nur schwer umsetzbare Ziele.

Objektives Kriterium. Als objektives Kriterium für die Kompetenz, sich adäquate Lernziele zu setzen, wurde ein Kodiermanual entwickelt, welches die Qualität der einzelnen Lernziele jeweils in Bezug darauf quantifiziert, ob sie realistisch, aufgabenspezifisch, zeitlich erreichbar, konkret und herausfordernd formuliert sind. Bei der Kodierung wird zusätzlich für jede Schülerin und jeden Schüler sowohl die subjektive

Einschätzung der Aufgabenanforderungen als auch die subjektive Einschätzung des aufgabenbezogenen Vorwissens (bzw. der Wissenslücken) in die Auswertung mit einbezogen.

Kompetenzmaß. Als Gütekriterium der Kompetenz gilt die durch das Kodiermanual ermittelte Angemessenheit der durch die Schülerinnen und Schüler ausgewählten Art und Anzahl an Zielen.

Evaluation. Es besteht die Annahme, dass die Auswahl konkreter, aufgabenspezifischer Lernziele, die auf Wissenslücken ausgerichtet sind, in einem positiven Zusammenhang mit dem Wissenszuwachs beim Lesen eines Sachtextes steht. Im Gegensatz hierzu sollte die Auswahl globaler, nur schwer umsetzbarer Lernziele keinen lernförderlichen Einfluss ausüben. Im Rahmen der Evaluationsstudie mit 50 Gymnasiastinnen und Gymnasiasten konnte erwartungskonform ein signifikant positiver Zusammenhang zwischen der Auswahl konkreter Lernziele und dem Wissenszuwachs ($r = .37$, $p < .05$) aufgedeckt werden. Bei der Analyse des Zusammenhangs zwischen der Auswahl globaler, nur schwer umsetzbarer Lernziele und dem Wissenszuwachs ergab sich ebenfalls erwartungskonform ein negativer (jedoch statistisch nicht bedeutsamer) Zusammenhang ($r = -.22$, n.s.).

3.2 Lernstrategien adäquat einsetzen (Teilkompetenz e)

Während des Lesens eines Sachtextes ist der Einsatz von Lernstrategien zentral, um die relevanten Informationen des Textes zu organisieren, zu elaborieren und zu integrieren (vgl. Weinstein/Mayer 1986). Erfolgreiche selbstregulierte Schülerinnen und Schüler verfügen demnach über die Kompetenz, zuvor ausgewählte Lernstrategien unter Zuhilfenahme des eigenen Strategiewissens (Teilkompetenz d) so anzuwenden, dass beim Lesen des Textes ein maximaler Wissenszuwachs erzielt wird (vgl. Pressley/Borkowski/Schneider 1987).

Subjektives Verhalten. Zur subjektiven Erfassung der Kompetenz, Lernstrategien adäquat einzusetzen, werden den Schülerinnen und Schülern zwei Textpassagen des Sachtextes „Wasser“ vorgegeben mit der Instruktion, je eine von zwei Lernstrategien (*Textmarkieren* und *concept mapping*) auf eine Textpassage so anzuwenden, dass sie im Anschluss möglichst viel über die zentralen Aspekte des Sachtextes wissen.

Objektives Kriterium. Als objektives Kriterium für die Kompetenz, Lernstrategien adäquat einzusetzen, wurde ein Kodiermanual zur Bestimmung der Qualität der Strategieanwendungen entwickelt. Dieses Manual wurde auf der Basis der Textmarkierungen bzw. concept maps erstellt, die Expertinnen und Experten des selbstregulierten Lernens bei derselben Aufgabenstellung vorgenommen hatten.

Kompetenzmaß. Als Gütekriterium der Kompetenz, Lernstrategien adäquat einzusetzen, gilt die Übereinstimmung der Strategieanwendungen der Schülerinnen und Schüler mit denen der Expertinnen und Experten.

Evaluation. Die zwei Produkte der Strategieanwendung (Markierungen im Text, concept maps) wurden für die weiteren Analysen zu einem Lernstrategiescore zusam-

mengefasst. Es besteht die Annahme, dass Schülerinnen und Schüler durch die adäquate Anwendung von Lernstrategien (im Sinne eines qualitativ hochwertigen Einsatzes der Lernstrategien) eine tiefere Verarbeitung der Inhalte des Sachtextes erzielen, was sich folglich in einem höheren Wissenszuwachs widerspiegeln sollte (vgl. z.B. Leutner/Leopold 2006). Diese Annahme wurde im Rahmen einer Untersuchung mit 68 Gymnasiastinnen und Gymnasiasten überprüft. Hierbei erwies sich die Reliabilität des Lernstrategiescores als zufriedenstellend ($\alpha = .75$). Ein erwartungskonformer statistisch bedeutsamer positiver Zusammenhang zwischen dem Lernstrategiescore und dem Wissenszuwachs ($r = .39$, $p < .05$) ist ein erster Hinweis auf die Validität des Verfahrens.

3.3 Wissenstand nach dem Lernen einschätzen (Teilkompetenz h)

Die Kompetenz, den eigenen Wissensstand möglichst objektiv zu bewerten, ist eine zentrale Kompetenz im Anschluss an das Lernen. Sie dient unter anderem dem Aufdecken von Diskrepanzen, welche zwischen den vorab gesetzten Lernzielen und dem aktuellen Wissensstand noch bestehen (vgl. Winne/Hadwin 1998), sowie der Analyse, an welcher Stelle im Lernprozess Korrekturen erforderlich sind, um ihn zu optimieren (vgl. Schreiber 1998).

Subjektives Verhalten. Zur Erfassung der subjektiven Einschätzung des eigenen Wissenstands nach dem Lernen wurde ein Selbsteinschätzungsbogen mit zehn Items entwickelt. Die zehn Items repräsentieren zentrale Themen des Sachtextes. Die Aufgabe der Schülerinnen und Schüler besteht darin, auf einer sechsstufigen Likert-Skala einzuschätzen, inwiefern sie Fragen zu den Themen beantworten können.

Objektives Kriterium. Bei der Entwicklung des Testmaterials zur Erfassung des tatsächlich vorhandenen Wissens nach dem Lesen des Sachtextes wurde der von Leopold, den Elzen-Rump und Leutner (2006) konstruierte Wissenstest zum Sachtext „Wasser" den Veränderungen im modifizierten Sachtext angepasst. Der resultierende Wissenstest enthält 15 Fragen im geschlossenen Antwortformat sowie drei offene Fragen. Der Wissenstest behandelt hierbei die identischen zehn Themen wie der subjektive Einschätzungsbogen und wurde im Anschluss an die Selbsteinschätzung (subjektives Verhalten) durch die Schülerinnen und Schüler bearbeitet.

Kompetenzmaß. Als Gütekriterium der Kompetenz, den eigenen Wissenstand nach dem Lernen einzuschätzen, gilt die Übereinstimmung der subjektiven Einschätzungen mit den entsprechenden Antworten im Nachtest. Es resultierte ein Gesamtmaß für die Anzahl der Übereinstimmung zwischen subjektiver Einschätzung und tatsächlichem Wissen im Anschluss an das Lernen.

Evaluation. Die Evaluation des objektiven Nachtests ergab in einer Untersuchung mit 59 Gymnasiastinnen und Gymnasiasten einen zufriedenstellenden Reliabilitätskoeffizienten von $\alpha = .77$. Der Selbsteinschätzungsbogen erbrachte bei identischer Schülerstichprobe einen Reliabilitätskoeffizienten von $\alpha = .80$. Bezogen auf die Kompetenz, seinen eigenen Wissensstand nach dem Lernen adäquat einzuschätzen, besteht die Annahme, dass Schülerinnen und Schüler, die sich ihres Wissenstandes nach dem Lesen eines

Sachtextes bewusst sind, dieses metakognitive Wissen auch während des Lesens genutzt haben, um bestehende Wissenslücken zu füllen. Diese Kompetenz sollte sich folglich in einem höheren allgemeinen Wissenszuwachs niederschlagen. Zur Überprüfung der Annahme und Validierung des Übereinstimmungsmaßes wurde der Zusammenhang des Selbsteinschätzungsmaßes mit dem Wissenszuwachs bestimmt. Hierbei zeigte sich erwartungskonform ein statistisch bedeutsamer positiver Zusammenhang ($r = .48$, $p < .01$).

4. Ausblick

Seit dem Frühjahr 2009 werden im Rahmen der Hauptuntersuchung die evaluierten Testverfahren mit mehr als 500 Schülerinnen und Schülern am Ende der neunten Jahrgangsstufe eingesetzt. Basierend auf den so gewonnenen Daten wird im Anschluss die Kompetenzstruktur beim Lernen aus Sachtexten mit Hilfe von Strukturgleichungsmodellen modelliert. Ferner wird in einem weiteren Schritt angestrebt, Item-Response-Modelle zu prüfen und empirisch aufeinander aufbauende Kompetenzniveaus zu definieren.

5. Theoretischer und praktischer Nutzen

Im Rahmen der Forschung zum selbstregulierten Lernen aus Sachtexten liefert das vorliegende Projekt einen bedeutenden Beitrag zur allgemeinen Theoriebildung, indem es die Relevanz der verschiedenen Teilkompetenzen sowie die wechselseitigen Beziehungen zwischen den Teilkompetenzen aufdeckt. Die Definition von Kompetenzniveaus macht es darüber hinaus möglich, unterschiedliche Ausprägungen der Fähigkeit zum selbstregulierten Lernen aus Sachtexten inhaltlich zu beschreiben und damit die Fähigkeit selbst genauer zu definieren.

Für die Praxis bietet sich hieraus die Möglichkeit, differenzierte Fördermaßnahmen für die verschiedenen Teilkompetenzen zu entwickeln. Dies ermöglicht es, gezielt nur diejenigen Teilkompetenzen zu fördern, die bei den jeweiligen Schülerinnen und Schülern gering ausgeprägt sind, was im Vergleich zu derzeit gängigen globalen Fördermaßnahmen im Bereich des selbstregulierten Lernens ein vielversprechendes, ökonomisches Vorgehen darstellt.

Literatur

Artelt, C./Baumert, J./Julius-McElvany, N. (2003): Selbstreguliertes Lernen. Motivation und Strategien in den Ländern der Bundesrepublik Deutschland. In: Baumert, J./Artelt, C./Klieme, E./Neubrand, M./Prenzel, M./Schiefele, U./Schneider, W./Tillmann, K.-J./Weiß, M. (Hrsg.): PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland. Opladen: Leske+Budrich, S. 51–76.

Boekaerts, M. (1999): Self-regulated learning: Where we are today. In: International Journal of Educational Research 31, S. 445–457.

Butler, D.L./Winne, P.H. (1995): Feedback and self-regulated learning: A theoretical synthesis. In: Review of Educational Research 65, S. 245–281.

Christmann, U./Groeben, N. (1999): Psychologie des Lesens. In: Franzmann, B./Hasemann, K./Löffler, D./Schön, E. (Hrsg.): Handbuch Lesen. München: Saur, S. 145–223.

Leopold, C./den Elzen-Rump, V./Leutner, D. (2006): Selbstreguliertes Lernen aus Sachtexten. In: Prenzel, M./Allolio-Näcke, L. (Hrsg.): Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms. Münster: Waxmann, S. 268–290.

Leutner, D./Leopold, C. (2006): Selbstregulation beim Lernen aus Sachtexten. In: Mandl, H./Friedrich, H.F. (Hrsg.): Handbuch Lernstrategien. Göttingen: Hogrefe, S. 38–49.

Locke, E.A./Latham, G.P. (1990): A theory of goal setting and task performance. Englewood Cliffs, NJ: Prentice Hall.

Pressley, M./Borkowski, J.G./Schneider, W. (1987): Cognitive strategies. Good strategy users coordinate metacognition and knowledge. In: Vasta, R./Whitehurst, G. (Hrsg.): Annals of child development 4. Greenwich, CT: JAI Press, S. 80–129.

Schmidt-Weigand, F. (2006): Dynamic visualizations in multimedia learning: The influence of verbal explanations on visual attention, cognitive load and learning outcome. Dissertation, Justus-Liebig-Universität Gießen. http://geb.uni-giessen.de/geb/volltexte/2006/2699/ [20.07.2009].

Schreiber, B. (1998): Selbstreguliertes Lernen. Münster: Waxmann.

Thillmann, H. (2008): Selbstreguliertes Lernen durch Experimentieren. Von der Erfassung zur Förderung. Dissertation, Universität Duisburg-Essen, Fachbereich Bildungswissenschaften, Essen.

Weinstein, C.E./Mayer, R. (1986): The teaching of learning strategies. In: Wittrock, M.C. (Hrsg.): Handbook of research on teaching. New York: Macmillan, S. 315–327.

Winne, P.H./Hadwin, A.F. (1998): Studying as self-regulated learning. In: Hacker, D.J./Dunlosky, J./Graesser, A.C. (Hrsg.): Metacognition in educational theory and practice. Mahwah, NJ: Erlbaum, S. 277–304.

Wirth, J./Leutner, D. (2008): Self-regulated learning as a competence. Implications of theoretical models for assessment methods. In: Zeitschrift für Psychologie/Journal of Psychology 216, S. 102–110.

Zimmerman, B.J. (2000): Attaining self-regulation: A social cognitive perspective. In: Boekaerts, M./Pintrich, P.R./Zeidner, M. (Hrsg.): Handbook of self-regulation. San Diego, CA: Academic Press, S. 13–69.

Anschrift der Autor/innen

Dipl.-Psych. Melanie Schütte, Lehrstuhl für Lehr-Lernforschung, Institut für Erziehungswissenschaft, Ruhr-Universität Bochum, Postfach 102148, D-44780 Bochum
E-Mail: melanie.schuette@rub.de

Prof. Dr. Joachim Wirth, Lehrstuhl für Lehr-Lernforschung, Institut für Erziehungswissenschaft, Ruhr-Universität Bochum, Postfach 102148, D-44780 Bochum
E-Mail: joachim.wirth@rub.de

Prof. Dr. Detlev Leutner, Lehrstuhl für Lehr-Lernpsychologie, Fachbereich Bildungswissenschaften, Universität Duisburg-Essen, Postfach, D-45117 Essen
E-Mail: detlev.leutner@uni-due.de

Tobias Gschwendtner/Bernd Geißel/Reinhold Nickolaus

Modellierung beruflicher Fachkompetenz in der gewerblich-technischen Grundbildung

Projekt Berufspädagogik[1]

1. Vorbemerkungen

In der beruflichen Bildung ist der Anspruch weit verbreitet, berufliche Handlungskompetenz möglichst umfassend, d.h. auch unter Einschluss von Bereitschaften, Motivationen und Orientierungen zu modellieren, da nur so der Bezug zur beruflichen Performanz zu sichern sei. Dies spiegelt sich auch in den zahlreichen, in den beiden letzten Dekaden entstandenen hypothetischen Kompetenzstrukturmodellen, die z.B. aus Fach-, Human- und Sozialkompetenz zusammengesetzt sind (vgl. KMK 2007; Breuer 2006; Nickolaus 2008). Demgegenüber stehen bisher nur wenige Versuche, diese auf eine empirische Basis zu stellen. Ein Grund für diesen Kontrast ist sicherlich, dass berufliche Handlungskompetenz mannigfaltige handlungstheoretische Facetten berührt (vgl. Straka/Macke 2009), die eine Konstruktoperationalisierung schwierig werden lassen. Ein anderer Grund könnte die Erhebungssituation selbst sein: Der reale Arbeitsprozess als Ort beruflicher Performanz stößt an die Grenzen reliabler, objektiver und praktikabler Messungen (vgl. Gschwendtner/Abele/Nickolaus 2009). Umfassende Modellierungen werden wohl erst dann gelingen können, wenn dazu ein hinreichend theoretisches und empirisches Fundament innerhalb der relevanten Kompetenzdimensionen gelegt ist (vgl. Nickolaus 2008). Vor diesem Hintergrund beschränken wir uns in diesem Beitrag auf die Modellierung beruflicher Fachkompetenz als eine zentrale Facette beruflicher Handlungskompetenz. Die Daten zur beruflichen Fachkompetenz wurden in „simulierten“ Umwelten (paper-pencil Tests mit anwendungsorientierten, komplexen Aufgaben und computerbasierten Simulationen) für die Berufe Kfz-Mechatroniker/in und Elektroniker/in für Energie- und Gebäudetechnik im ersten Ausbildungsjahr (auch Grundbildung genannt) erhoben. Hinsichtlich der Wahl geeigneter Erhebungsinstrumente in der beruflichen Bildung ist auch die Frage interessant, inwiefern computersimulierte Umwelten valide Erhebungsarrangements für Leistungsabschätzungen in realen Arbeitsumwelten sein können (Befunde hierzu vgl. Gschwendtner/Abele/Nickolaus 2009; vgl. auch Abschnitt 5).

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: DFG Ni 606/3-1 (kooptiertes Projekt)) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

2. Forschungsstand

In den empirisch ausgerichteten Arbeiten zur Kompetenzmodellierung im beruflichen Bereich wird meist auf eine horizontale und vertikale Differenzierung von Wissen zurückgegriffen. Horizontale Differenzierungen segmentieren meist unterschiedliche Wissensformen in Abhängigkeit angenommener kognitiver Repräsentiertheit (entweder bezogen auf ein kognitives Wissensmodell oder curriculare Zuschnitte). Vertikale Differenzierungen sollen dann innerhalb horizontaler Kategorien den Status der semantischen Differenzierung bzw. kognitiven Aktivitäten hierarchisieren. Dabei dienen bspw. modifizierte Lernzielklassifikationen im Anschluss an Bloom u.a. (1973) oder Anderson/Krathwohl (2001) zur vertikalen und z.B. Fortmüllers (1996) Arbeit zur horizontalen Differenzierung. Beispielhaft dafür steht das der Testkonstruktion in der Untersuchung von Leistungen, Motivation und Einstellungen in Abschlussklassen beruflicher Schulen (ULME III) zugrunde liegende Klassifikationsraster, in dem drei Dimensionen (Faktenwissen, Konzeptwissen, prozedurales Wissen) und drei Niveaus (Reproduktion, Anwendung, Reflektieren und Bewerten) konzeptionell unterschieden werden (vgl. Brand/Hofmeister/Tramm 2005; Lehmann/Seeber 2007). In den auf Basis probabilistischer Testtheorie generierten empirischen Modellbildungen ergaben sich in ULME im gewerblich-technischen Bereich durchgängig eindimensionale Modelle des Fachwissens.[2] Befriedigende Niveaumodelle, die zur Erklärung der Anforderungen einen Beitrag leisten könnten, wurden im gewerblich-technischen Bereich in ULME nicht erzielt. Im kaufmännischen Bereich ergaben sich partiell Hinweise auf eine bessere Passung eines zweidimensionalen Modells mit den Dimensionen „betriebs- und volkswirtschaftliche sowie rechtliche Aspekte“ und dem „Rechnungswesen“ (vgl. Seeber 2008, S. 80ff.). Für die Niveaumodellierung erweisen sich bei Seeber (2008) das Konzeptwissen (Struktur und Zusammenhangswissen), das prozedurale Wissen, die Verknüpfung beider Komponenten und lernfeldübergreifende Aufgaben als bedeutsam. Für die Fachleistungstests bei Einzelhandelskaufleuten erweist sich die Notwendigkeit der Anwendung mathematischer Strukturen und Algorithmen auf ökonomische Zusammenhänge als stärkster Prädiktor der Itemschwierigkeit. Weitere Beiträge erbringen das fachspezifische Begriffswissen und die kognitive Durchdringung ökonomischer Sachverhalte (Seeber 2007). Winther und Achtenhagen (2008) gehen im kaufmännischen Bereich konzeptionell von zwei Dimensionen kaufmännischen Wissens (economic literacy und numeracy) aus, die sie als Facetten einer verstehensbasierten Kompetenz ausweisen (vgl. Winther/Achtenhagen 2009). Sie können empirisch zeigen, dass sich Fachwissen („verstehensbasierte Kompetenz“) und fachspezifische Problemlösefähigkeit („handlungsbasierte Kompetenz“) günstiger in einem zweidimensionalen Modell darstellen lassen. Die Niveaumodellierung kommt bei Winther (2008) durch die inhaltliche Komplexität und Ansprüche an die funktionale Modellierung zustande.

2 Einschränkend sei darauf hingewiesen, dass in ULME aufgrund der geringen Fallzahlen nicht für alle einbezogenen gewerblich-technischen Ausbildungsberufe Skalierungen vorgelegt werden konnten.

3. Forschungsdesign

3.1 Zielsetzung der Untersuchung

Die Auswertungen unseres Forschungsprojekts zielten bisher darauf, die prädiktive Kraft kognitiver und motivational-affektiver Merkmale (fachspezifisches Vorwissen, IQ, Lesefähigkeit, allgemein-mathematische Kompetenz, motivationale Variablen) und spezifischer Lernumgebungen (Ausbildungsformen (dual, vollzeitschulisch), betriebliche Ausbildungsqualitäten und schulische Lehrformen) auf die Entwicklung berufsfachlicher Kompetenzen im ersten Ausbildungsjahr zu untersuchen (s. hierzu u.a. Nickolaus/Gschwendtner/Geißel 2008).

In diesem Beitrag gehen wir nun zusätzlich folgenden Fragen nach:

1. Lassen sich ähnlich zu den oben referierten Studien empirisch unterscheidbare Kompetenzdimensionen in der gewerblich-technischen Grundbildung (erstes Ausbildungsjahr) bei den Berufen Kfz-Mechatroniker/in und Elektroniker/in für Energie- und Gebäudetechnik aufzeigen?
2. Verändern sich etwaige Dimensionalitäten über das erste Ausbildungsjahr hinweg?
3. Welche Schwierigkeitsmerkmale erweisen sich zur Beschreibung verschiedener Anforderungsniveaus in den genannten technischen Berufen als relevant?

3.2 Stichprobenzusammensetzung und verwendete Instrumente

Die Stichprobe setzt sich aus 203 Auszubildenden aus 9 Elektroniker/innen-Klassen (5 Klassen der einjährigen Berufsfachschule, 4 Teilzeitklassen) und 286 Auszubildenden aus 11 Kfz-Mechatroniker/innen-Klassen (7 Klassen der einjährigen Berufsfachschule, 4 Teilzeitklassen) zusammen.

Die Testkonstruktion folgt dem Gedanken, dass sich berufsfachliche Kompetenzen konzeptionell in drei Ebenen aufspannen lassen: Deklaratives Fachwissen, prozedurales Fachwissen und fachspezifische Problemlösefähigkeit (vgl. Knöll 2007). Deklaratives Fachwissen kann auf einem Kontinuum abgebildet werden, das sich zwischen der Reproduktion einfachster Sachverhalte und Begründungen/Beurteilungen innerhalb komplexer Zusammenhänge abspielt. Prozedurales Fachwissen zeigt sich in der Anwendung deklarativen Wissens. Von Problemlösen kann bei Aufgaben die Rede sein, die zwar berufstypische Aufgaben sind, deren Neuigkeits- und Komplexitätsgrad jedoch vielfältigere mentale Prozesse aktiviert als dies bei deklarativen oder prozeduralen Aufgaben der Fall ist. Deklaratives und prozedurales Fachwissen wurde mit paper-pencil Tests erhoben, fachspezifische Problemlösefähigkeit bei den Kfz-Mechatroniker/innen im gleichen Format, bei den Elektroniker/innen mithilfe von computerbasierten Simulationen. Der Fachwissenstest wurde in einer nahezu identischen Version zu Beginn und am Ende des ersten Ausbildungsjahres eingesetzt, der Test zur

Erfassung der fachspezifischen Problemlösefähigkeit einmalig am Ende des erstens Ausbildungsjahres.

Zur besseren Nachvollziehbarkeit, wie berufsfachliche Kompetenzen gemessen wurden, seien im Folgenden einige Ausschnitte aus den Instrumentarien und zwei Beispielitems vorgestellt.

Instrumente zur Erfassung deklarativen und prozeduralen Fachwissens

Die Tests für beide Berufe sind Eigenkonstruktionen im Anschluss an curriculare Analysen der Lernfelder 1 bis 4 der jeweiligen Rahmenlehrpläne und, soweit diese für die Grundstufe Relevanz annehmen, arbeitsprozessbezogene Merkmale von Facharbeit. Die endgültige Itemselektion erfolgte auf der Basis einer inhaltlichen Validierung durch Expert/innen der jeweiligen Berufe und durch Pilotierungen. Selektiert wurden bei den Kfz-Mechatroniker/innen 27 Items, davon 3 dichotome und 24 polytome (Cronbach's $\alpha = .84$), bei den Elektroniker/innen 14 dichotome und 7 polytome Items (Cronbach's $\alpha = .78$).

Ein Beispielitem aus dem Kraftfahrzeugmechatronikertest soll exemplarisch unseren Testzugang verdeutlichen. Den Schüler/innen wurde bspw. das untenstehende Motorschnittbild dargeboten (vgl. Abbildung 1), verbunden mit der Aufforderung, die Motorbauteile zu benennen. Im Anschluss daran wurde, stochastisch hiervon unabhängig, eine Frage nach dem Motortyp (Verbrennungs- und Einspritzverfahren) gestellt, wobei die Begründung dafür zu geben war, warum es sich um den vermuteten und nicht um einen anderen Motortyp handelt.[3]

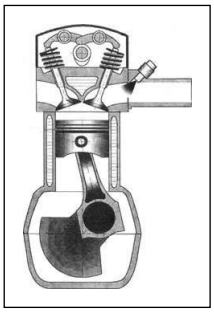


Abb. 1: Schnittbild eines 4-Takt-Ottomotors mit indirekter Einspritzung

3 Die Antwort erhält man nur durch ein wechselseitiges in Beziehung setzen von Informationen aus dem Bild, wobei eine Umdeutung einer „Fehlinformation" vorgenommen werden muss: Die Abbildung enthält keine Zündkerze, was als Zeichen für einen Dieselmotor gedeutet werden kann, wenn nicht die Einspritzdüse, wie in der Abbildung dargestellt, im Ansaugkanal sitzen würde. Das heißt zum einen, dass es sich um eine indirekte Einspritzung handelt und als Verbrennungsprinzip dabei nur der Ottomotor in Frage kommt, da Dieselmotoren entweder direkt in den Brennraum einspritzen oder indirekt in einen dem Brennraum nebengelagerten Zusatzraum (Vorkammer, Wirbelkammer) und eben nicht in den Ansaugkanal. Dass die Zündkerze nicht in der Abbildung vorkommt, liegt allein an der Art und Weise der Schnittbildung.

Instrumente zur Erfassung fachspezifischer Problemlösefähigkeit

In technischen Systemen geht der eigentlichen Fehlerbehebung als notwendige Bedingung die Fehleranalyse voraus. Dabei kann der Ausgangspunkt einer Fehleranalyse als ein Hypothesengenerieren über die Wahrscheinlichkeit einer Fehlerursache u.a. in Abhängigkeit vom Problemcharakter und dem Komplexitätsgrad des technischen Systems angesehen werden. Diese Hypothesen können primär aus dem Fehlerbild und der Kenntnis der Systemzusammenhänge abgeleitet werden. Im Anschluss daran sind meist eine messtechnische Eingrenzung des Fehlers und dann der Tausch des entsprechenden Bauteils oder der Baugruppe notwendig. Der gesamte Prozess kann als Ausdruck der Problemlösefähigkeit angesehen werden. In unserer Untersuchung beschränken wir uns zum einen aus forschungspragmatischen Gründen, jedoch auch aufgrund des erfolgskritischen Status solcher Diagnoseleistungen auf die Fehleranalyse und verwenden dieses Wort synonym zu Problemlösen. Bis auf den Reparaturaspekt integrierten wir sämtliche obige Handlungssystematiken resp. Fehlersuchstrategien in unsere Tests.

Der Test für die Kfz-Mechatroniker/innen umfasste sechs Fehlfunktionen eines Autos im Bereich der Beleuchtungsanlage (Cronbach's α = .50). Zum einen wurden die Auszubildenden anhand eines autospezifischen Stromlaufplans (hier sind sämtliche Verdrahtungen und Bauteile graphisch dargestellt) um ihre Einschätzung gebeten,

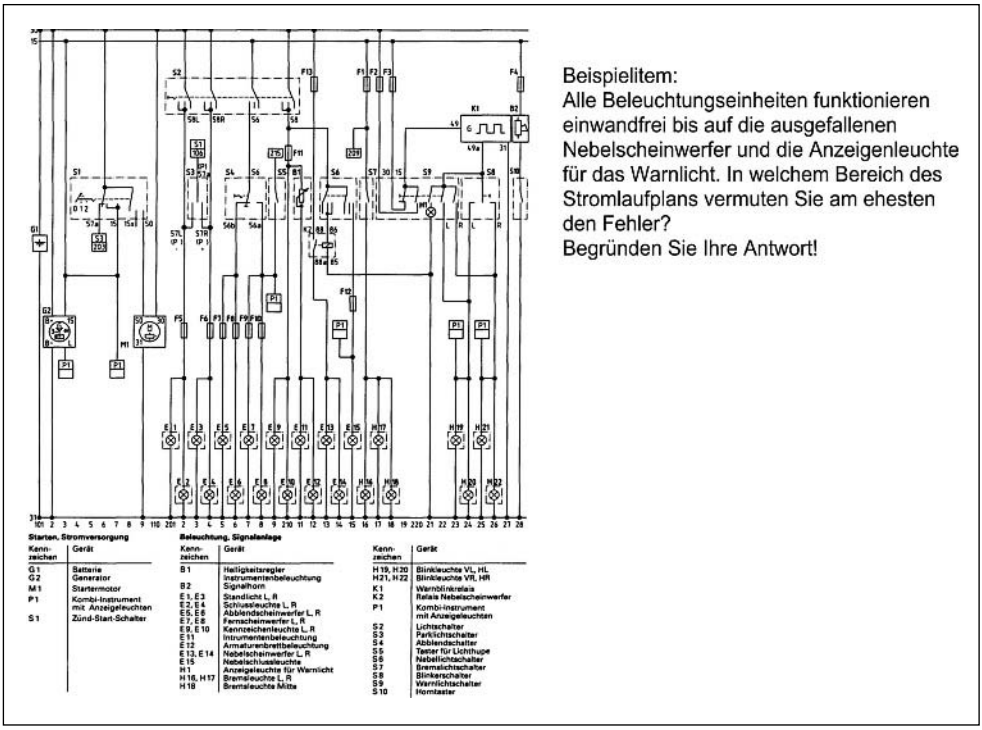


Abb. 2: Beispielitem aus dem Test zur Erfassung fachspezifischer Problemlösefähigkeit bei Kfz-Mechatroniker/innen

welche(s) Bauteil(e) am wahrscheinlichsten defekt sein könnte(n) (s. hierzu das Beispielitem und den Stromlaufplan in Abbildung 2). Ferner wurde analog zur Praxis um einen messtechnischen Vorschlag gebeten, wie der Fehler gefunden werden könnte, wozu Messstellen, Messinstrumente und erwartete Messwerte zu spezifizieren waren.

Bei den Elektroniker/innen waren die Messinstrumente zum Zeitpunkt der Datenerhebung aufgrund von Synergieeffekten aus Vorstudien etwas fortgeschrittener. Hier wurde die computerbasierte Simulation MILAS verwendet (vgl. Gschwendtner/Geißel/Nickolaus 2007), bei der mithilfe von authentischen Arbeitsaufträgen vier Fehlerfälle in zwei elektrotechnischen Systemen (Kochplatte und Wechselschaltung) interaktiv zu analysieren waren (Cronbach's α = .53) (s. zur Realisierung der Simulation Abbildung 3). Zusätzlich wurden auch Daten zur Messstrategie erhoben.

Die niedrigen Reliabilitäten der Tests zur Erfassung fachspezifischer Problemlösefähigkeit werden weniger über sehr niedrige oder gar negative Trennschärfen der einzelnen Fehlerfälle moderiert (part-whole Korrelationen bewegen sich bei den Kfz-Mechatroniker/innen zwischen r = .22 und r = .44 und bei den Elektroniker/innen zwischen r = .22 und r = .37) als vielmehr durch die relativ geringe Itemanzahl der Problemlösefähigkeitsskala und der relativ geringen Interkorrelationen der Fehlerfälle (r~.20; Ausnahme ist ein Korrelationswert von r = .46) und dies gleichgültig, ob mit Rangdaten (Spearman's Rho) oder binär kodierten Daten (Phi) ausgewertet wird.[4]

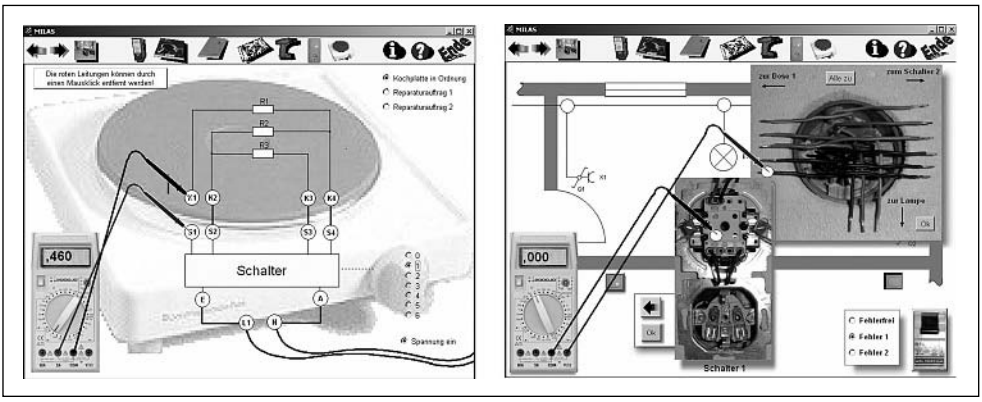


Abb. 3: Die Systeme Kochplatte (links) und Wechselschaltung (rechts) zur Erfassung fachspezifischer Problemlösefähigkeit bei Elektroniker/innen

4 Unreliable Problemlösefähigkeitsskalen sind ein nicht selten auftretender Befund (s. hierzu z.B. Süß 1996) und inhaltlich mitverursacht durch die Heterogenität dessen, was man unterschätzend im Singular Realität nennt. Wu (2004) macht in diesem Zusammenhang deutlich: „Uni-dimensionality is only a theoretical notion. In reality, there is no such thing as uni-dimensionality, only the degree of uni-dimensionality." Niedrige Interkorrelationen zwischen Problemfällen können zusätzlich auch eine Frage der verwendeten Korrelationsansätze sein. So zeigen tetrachorische Ansätze i.d. Regel höhere Zusammenhänge an als die hier verwendeten klassischen Verfahren (Rho, Phi).

4. Ergebnisse und Diskussion

4.1 Strukturmodellierungen bei den Kfz-Mechatroniker/innen

Im Anschluss an die analytische Wissensdimensionierung von Fortmüller (1996) und der curricular-inhaltlichen Ausrichtung der Handlungs- und Fachsystematiken können verschiedene kognitive Repräsentationsmodelle von Fachkompetenz vermutet werden:

- Ein 1-dimensionales Modell aller Fachwissensitems.
- Ein 2-dimensionales Modell im Anschluss an Fortmüller (1996). Dieses wäre ein Fachwissensstrukturmodell aus einer deklarativen und einer prozeduralen Dimension.
- Ein 2-dimensionales Modell im Anschluss an die Handlungs- und Fachsystematiken. Die eine Dimension wäre durch Items genuin fahrzeugmechanischen Inhalts gebildet, die andere durch elektrotechnische Inhalte.
- Ein 4-dimensionales Kombinationsmodell. Die vier Dimensionen würden sich durch Aufgaben abbilden lassen, die (1) deklarativ-fahrzeugmechanische, (2) deklarativ-elektrotechnische, (3) prozedural-fahrzeugmechanische und (4) prozedural-elektrotechnische Inhalte repräsentieren.

Die insgesamt 3 möglichen Mehrdimensionalitäten für den Wissenstest wurden zur eindimensionalen Skalierung vergleichend gerechnet. Alle Modellvergleiche auf Basis der Devianzunterschiede, Itemstatistiken, Skalierungen und Korrelationen wurden mit ConQuest (vgl. Wu/Adams/Wilson 1998) gerechnet.

Am Messzeitpunkt zum Ende der Grundbildung korrelieren die modellierten Mehrdimensionalitäten latent sehr hoch um $r = .9$. Die mehrdimensionalen Modellierungen besitzen zudem höhere Modellabweichungen (Devianzen) als die eindimensionale Modellierung. Die Unterschiede sind, geprüft mittels einer Chi-Quadrat-Verteilung der entsprechenden Freiheitsgrade, durchgängig signifikant ($p < .000$). Damit passt ein eindimensionales Fachkompetenzmodell besser zu den Daten als ein konkurrierendes mehrdimensionales Modell.

Am Messzeitpunkt zu Beginn der Ausbildung spricht die gleiche Devianzstatistik für eine signifikant ($p < .000$) günstigere Passung eines 2-dimensionalen Modells, das aus den Dimensionen fahrzeugmechanisches und elektrotechnisches Fachwissen gebildet wird. Die beiden Dimensionen korrelieren latent mit $r = .75$. Diese Befunde deuten darauf hin, dass es plausibel erscheint, dass sich in Folge berufsschulischer Lernwege mittels komplexer, thematisch vernetzter Lernfeldaufgaben die zu Ausbildungsbeginn vorherrschenden zwei Dimensionen während des ersten Ausbildungsjahres zu einer Dimension verdichten.

Auch wenn das Testformat zur Erfassung der fachspezifischen Problemlösefähigkeit ohne Animation und Simulation technischer Systeme auskam und auf ein paper-pencil Format zurückgegriffen wurde, konnte die latente Korrelation von $r = .76$ und die bessere Passung eines zweidimensionalen Modells zwischen Fachwissen und fachspezifi-

scher Problemlösefähigkeit zeigen, dass zwei separierbare Kompetenzfacetten erfasst wurden (vgl. dazu ausführlich Gschwendtner 2008). Jedoch war es aufgrund der geringen Anzahl von sechs Problemfällen nicht möglich, eine eigenständige und reliable Berichtsskala zu generieren.[5]

4.2 Strukturmodellierungen bei den Elektroniker/innen für Energie- und Gebäudetechnik

Als mögliche Dimensionen wurden hier vor dem Hintergrund der eingangs vorgestellten Befundlagen und des im Vergleich zu den Kfz-Mechatroniker/innen anderen Inhaltszuschnitts vergleichend gerechnet:

- Ein 1-dimensionales Modell aller Fachwissensitems.
- Analog zu den Kfz-Mechatroniker/innen ein 2-dimensionales Modell im Anschluss an Fortmüller (1996).
- Ein 2-dimensionales Modell, das die Kategorien Aufgaben mit und ohne mathematische Operationen beinhaltet.

Die Modelle wurden sequentiell gerechnet, wobei auch hier das eindimensionale Fachwissensmodell die beste Passung erzielte. Bei Einbezug der Daten zur fachspezifischen Problemlösefähigkeit deutet sich ebenso wie bei den Kfz-Mechatroniker/innen eine bessere Passung eines zweidimensionalen Modells, bestehend aus Fachwissen und fachspezifische Problemlösefähigkeit an.

4.3 Niveaumodellierungen in beiden Berufen

Im Anschluss an eigene (vgl. Gschwendtner/Geißel/Nickolaus 2007; Nickolaus/Gschwendtner/Knöll 2006) und andere Vorarbeiten (vgl. Hartig 2007; Hartig/Jude 2007; Seeber 2007; Seeber 2008) wurden die Items post-hoc mit folgenden Schwierigkeitsmerkmalen bewertet:

- Vertrautheit aus der Sekundarstufe 1
- Hinweisgüte des Tabellenbuches[6]

---

5 Bestandteil des Fachwissenstests waren auch elektrotechnische Inhalte, die im Test zur fachspezifischen Problemlösefähigkeit benötigt wurden. Latente Korrelationen stellen messfehlerbereinigte Zusammenhänge dar. Hierbei wird der Zusammenhang um die Messungenauigkeit (Unreliabilität) der korrelierten Tests korrigiert (vgl. Lord/Novick 1968). Somit können auch Korrelationen unreliabler Skalen vorgenommen werden.

6 Das Tabellenbuch stellt in der gewerblich-technischen Ausbildung ein Hilfsmittel im Sinne eines Nachschlagewerks von Formeln, Kennwerten und Informationen zur Funktionsweise ganzer Baugruppen dar. Der Einsatz ist gewöhnlich auch bei Klassenarbeiten und bei den Ab-

- Bloomsche Taxonomie
- „Wissensvernetztheit“: Struktureigenschaften des erforderlichen Wissens (Einzelheiten, Zusammenhänge, Systemwissen)
- Anzahl der notwendigen Lösungsschritte
- Modellierungsnotwendigkeit: Müssen (Teil-)Funktionen technischer Elemente erschlossen werden?
- Wissensart (deklarativ/prozedural)

Die Niveaubildung wurde im Anschluss an Hartig (vgl. 2007) durchgeführt. Dabei wurden die Aufgaben einer Analyse mittels der oben genannten Kriterien unterzogen. Für die z.T. dummykodierten schwierigkeitsbestimmenden Aufgabenmerkmale wurde anschließend eine Regressionsanalyse vorgenommen, die nach dem Verfahren der schrittweisen Regression (abhängige Variable Itemschwierigkeit) bei einem Signifikanzniveau von $p < .05$ ausgeführt wurde. Im Falle der Kfz-Mechatroniker/innen wurde die Bloomsche Taxonomie (44,9% Varianzaufklärung), die Wissensvernetztheit (5,3%) und die Vertrautheit aus der Sekundarstufe 1 (2,1%), im Falle der Elektroniker/innen die Hinweisgüte des Tabellenbuches (39,6%) und die Bloomsche Taxonomie (14,8%) in die Modellbildung aufgenommen. Substantielle Korrelationen mit den Itemschwierigkeiten weisen jedoch auch weitere Merkmale auf, wie im Falle der Elektroniker/innen z.B. die Anzahl der Lösungsschritte, Modellierungsnotwendigkeiten und die Vertrautheit aus der Sekundarstufe 1.

Bemerkenswert ist der Befund, dass, ähnlich wie in ULME III, die oberen Niveaus, die zugleich das curriculare Anspruchsniveau repräsentierten, nur von einem sehr kleinen Anteil der Auszubildenden erreicht werden (vgl. ausführlicher in Geißel 2008; Gschwendtner 2008; Nickolaus/Gschwendtner/Geißel 2008).

5. Zusammenfassung und Ausblick

Wir konnten erstens zeigen, dass Fachwissen bei beiden gewerblich-technischen Berufen am Ende der Grundstufe nicht weiter in Wissensformen oder curriculare Inhalte subdimensioniert werden kann, wie dies in anderen Studien anhand anderer Berufe dargestellt werden konnte. Wir konnten zweitens zeigen, dass Fachwissen und dessen Anwendung, die fachspezifische Problemlösefähigkeit zwar relativ hoch miteinander korrelieren, diese Konstrukte jedoch als eigenständig aufzufassen sind und sich damit in der Tat zwei Facetten beruflicher Fachkompetenz darstellen lassen. Wir konnten drittens zeigen, dass sich bei den Kraftfahrzeugmechatroniker/innen die am Anfang der Ausbildung noch bestehenden zwei Wissensdimensionen (fahrzeugmechanisches und elektro-

---

schlussprüfungen (Teil 1 und Teil 2) zur Nutzung erlaubt. Es ist davon auszugehen, dass das Tabellenbuch in Abhängigkeit der Rezeption bzw. Rezeptionskompetenzen einen nicht unerheblichen Beitrag zur Lösungswahrscheinlichkeit erbringt. Dies erklärt den differenzierten Zugriff innerhalb dreier Unterkategorien des Grades der zur Informationsaufnahme aus dem Tabellenbuch notwendigen symbolischen Enkodierung (explizit und ohne symbolische Rekodierung rezipierbar, explizit nicht wörtlich kodiert, implizit kodiert).

technisches Fachwissen) während des ersten Ausbildungsjahres zu einer Dimension verdichten, was als „kognitive Integrationsleistung" der Ausbildung verstanden werden kann. Viertens konnten wir zeigen, dass die einbezogenen Schwierigkeitsmerkmale immerhin ca. 50% der Itemschwierigkeitsvarianz erklären können. Gleichzeitig sind wir jedoch von einer dezidierten Beschreibung, gleichsam einer präzisen und interindividuell nutzbaren Anleitung für Itementwicklungen, noch ein Stück entfernt. Hierfür dient u.a. auch das bereits angelaufene Folgeprojekt (DFG Ni-606/6-1). Dabei werden in Orientierung an den gewonnenen Erkenntnissen zu den Schwierigkeitsmerkmalen systematisch Testversionen zur Erfassung des Fachwissens und der fachspezifischen Problemlösefähigkeit am Ausbildungsende entwickelt und damit u.a. der Ausbildungserfolg modelliert. Ergänzend wird der Frage nachgegangen, welche Zusammenhänge zwischen allgemeiner (dynamischer) Problemlösefähigkeit (mit dem MicroDYN-Ansatz (vgl. Greiff/Funke 2009) erhoben) und fachspezifischer Problemlösefähigkeit bestehen. Verwiesen sei abschließend auf eine im Kfz-Bereich durchgeführte Studie, die klären sollte, inwieweit die Testleistungen von Auszubildenden in realen mit jenen in computersimulierten Arbeitsproben zusammenhängen. Die Ergebnisse der Studie (vgl. Gschwendtner/Abele/Nickolaus 2009) weisen die Simulationsaufgaben als valide aus, d.h. der Aufgabenlösung realer und simulierter Arbeitsproben unterliegt die gleiche latente Fähigkeitsstruktur und ergänzend dazu unterscheiden sich die Itemschwierigkeiten zwischen den Testmodi nur minimal. Die Erweiterung dieses Testansatzes ist ebenso Gegenstand des Folgeprojekts und weiterer Publikationen.

Literatur

Anderson, L.W./Krathwohl, D.R. (2001): A Taxonomy for Learning, Teaching and Assessing. New York u.a.: Longman.

Bloom, B.S./Engelhart, M.D./Furst, E.J./Hill, W.H./Krathwohl, D.R. (1973): Taxonomie von Lernzielen im kognitiven Bereich. Weinheim/Basel: Beltz.

Brand, W./Hofmeister, W./Tramm, T. (2005): Auf dem Weg zu einem Kompetenzstufenmodell für die berufliche Bildung – Erfahrung aus dem Projekt ULME. In: bwp@ Berufs- und Wirtschaftpädagogik 8. http://www.bwpat.de/ausgabe8/brand_etal_bwpat8.shtml. [14.04.2009].

Breuer, K. (2006): Kompetenzdiagnostik in der beruflichen Bildung – eine Zwischenbilanz. In: Zeitschrift für Berufs- und Wirtschaftspädagogik 102, S. 194–210.

Fortmüller, R. (1996): Wissenschaftsorientierung und Praxisbezug als komplementäre Prinzipien lernpsychologisch fundierter Lehr-Lern-Arrangements. In: Fortmüller, R./Aff, J. (Hrsg.): Wissenschaftsorientierung und Praxisbezug in der Didaktik der Ökonomie. Festschrift Wilfried Schneider. Wien: Manz Schulbuch, S. 372–400.

Geißel, B. (2008): Ein Kompetenzmodell für die elektrotechnische Grundbildung: Kriteriumsorientierte Interpretation von Leistungsdaten. In: Nickolaus, R./Schanz, H. (Hrsg.) (2008): Didaktik der gewerblich-technischen Berufsbildung. Baltmannsweiler: Schneider Hohengehren, S. 121–142.

Greiff, S./Funke, J. (2009). Measuring complex problem solving – The MicroDYN approach. In: Scheuermann, F. (Hrsg.): The transition to computer-based assessment – Lessons learned from large-scale surveys and implications for testing. Luxembourg: Office for Official Publications of the European Communities. http://www.psychologie.uni-heidelberg.de/ae/allg/mitarb/jf/Greiff&Funke_2009_LuxPaper.pdf [03.05.2010].

Gschwendtner, T. (2008): Ein Kompetenzmodell für die kraftfahrzeugtechnische Grundbildung. In: Nickolaus, R./Schanz, H. (Hrsg.) (2008): Didaktik der gewerblich-technischen Berufsbildung. Baltmannsweiler: Schneider Hohengehren, S. 103–120.

Gschwendtner, T./Abele, S./Nickolaus, R. (2009): Computersimulierte Arbeitsproben: Eine Validierungsstudie am Beispiel der Fehlerdiagnoseleistungen von Kfz-Mechatronikern. In: Zeitschrift für Berufs- und Wirtschaftspädagogik 105, S. 557–578.

Gschwendtner, T./Geißel, B./Nickolaus, R. (2007): Förderung und Entwicklung der Fehleranalysefähigkeit in der Grundstufe der elektrotechnischen Ausbildung. In: bwp@ Berufs- und Wirtschaftpädagogik 13. http://www.bwpat.de/ausgabe13/gschwendtner_etal_bwpat13.pdf. [01.04.2009].

Hartig, J. (2007): Skalierung und Definition von Kompetenzniveaus. In: Beck, B./Klieme, E. (Hrsg.): Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie. Weinheim/Basel: Beltz, S. 83–99.

Hartig, J./Jude, N. (2007): Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In: Hartig, J./Klieme E. (Hrsg.): Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung. Bonn (Bildungsforschung: Bd. 20), S. 17–36.

KMK-Sekretariat der Kultusministerkonferenz. (Hrsg.) (2007): Referat Berufliche Bildung und Weiterbildung. Handreichung für die Erarbeitung von Rahmenlehrplänen der Kultusministerkonferenz für den berufsbezogenen Unterricht in der Berufsschule und ihre Abstimmung mit Ausbildungsordnungen des Bundes für anerkannte Ausbildungsberufe. http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2007/2007_09_01-Handreich-Rlpl-Berufsschule.pdf. [15.08.2009].

Knöll, B. (2007): Differenzielle Effekte von methodischen Entscheidungen und Organisationsformen beruflicher Grundbildung auf die Kompetenz- und Motivationsentwicklung in der gewerblich-technischen Erstausbildung. Eine empirische Untersuchung in der Grundausbildung von Elektroinstallateuren. Aachen: Shaker, Stuttgart, Univ., Diss. (Stuttgarter Beiträge zur Berufs- und Wirtschaftspädagogik; Bd. 30).

Lehmann, R./Seeber, S. (2007): Untersuchungen von Leistungen, Motivation und Einstellungen der Schülerinnen und Schüler in den Abschlussklassen der Berufsschlussklassen der Berufsschulen (ULME III). Hamburg: Behörde für Bildung und Sport.

Lord, F.M./Novick, M.R. (1968): Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Nickolaus, R. (2008): Vorstellungen zur Modellierung beruflicher Handlungskompetenz und erste Versuche zu ihrer empirischen Prüfung. In: Nickolaus, R./Schanz, H. (Hrsg.) (2008): Didaktik der gewerblich-technischen Berufsbildung. Baltmannsweiler: Schneider Hohengehren, S. 87–102.

Nickolaus, R./Gschwendtner, T./Geißel, B. (2008): Entwicklung und Modellierung beruflicher Fachkompetenz in der gewerblich-technischen Grundbildung. In: Zeitschrift für Berufs- und Wirtschaftspädagogik 104, S. 48–73.

Nickolaus, R./Gschwendtner, T./Knöll, B. (2006): Handlungsorientierte Unterrichtskonzepte als Schlüssel zur Bewältigung problemhaltiger Aufgaben? In: Minnameier, G./Wuttke, E. (Hrsg.): Berufs- und wirtschaftspädagogische Grundlagenforschung. Festschrift für Klaus Beck. Frankfurt a.M. u.a.: Peter Lang, S. 209–224.

Nickolaus, R./Schanz, H. (Hrsg.) (2008): Didaktik der gewerblich-technischen Berufsbildung. Baltmannsweiler: Schneider Hohengehren.

Seeber, S. (2007): Zur Anforderungsstruktur eines Fachleistungstests für Auszubildende des Berufs Einzelhandelskaufmann/Einzelhandelskauffrau. In: Münck, D./Van Buer, J./Breuer, K./Deißinger, T. (Hrsg.): Hundert Jahre kaufmännische Ausbildung in Berlin. Berlin: Opladen, S. 184–193.

Seeber, S. (2008): Ansätze zur Modellierung beruflicher Fachkompetenz in kaufmännischen Ausbildungsberufen. In: Zeitschrift für Berufs- und Wirtschaftspädagogik 104, S. 74–97.
Straka, G.A./Macke, G. (2009): Neue Einsichten in Lehren, Lernen und Kompetenz. Bremen: Institut Technik und Bildung (ITB), Universität Bremen (ITB Forschungsberichte 40/2009). http://elib.suub.uni-bremen.de/ip/docs/00010417.pdf [10.03.2010].
Süß, H.-M. (1996): Intelligenz, Wissen und Problemlösen. Kognitive Voraussetzungen für erfolgreiches Handeln bei computersimulierten Problemen. Göttingen u.a.: Hogrefe.
Winther, E. (2008): Kompetenztests für die kaufmännische Erstausbildung. Vortrag anlässlich der AEPF-Tagung in Kiel.
Winther, E./Achtenhagen, F. (2008): Kompetenzstrukturmodell für die kaufmännische Bildung. In: Zeitschrift für Berufs- und Wirtschaftspädagogik 104, S. 511–538.
Winther, E./Achtenhagen, F. (2009): Skalen und Stufen kaufmännischer Kompetenz. In: Zeitschrift für Berufs- und Wirtschaftspädagogik 105, S. 521–556.
Wu, M.L. (2004): Item Response Modelling with ConQuest. MPI Summer School. (Vortrag).
Wu, M.L./Adams, R.J./Wilson, M.R. (1998): ACER ConQuest. Generalised Item Response Modelling Software. Melbourne: Acer Press.

Anschrift der Autoren

Dipl.-Gwl. Tobias Gschwendtner, Universität Stuttgart, Institut für Erziehungswissenschaft und Psychologie, Abteilung Berufs-, Wirtschafts- und Technikpädagogik, Geschwister-Scholl-Straße 24D, D-70174 Stuttgart
E-Mail: gschwendtner@bwt.uni-stuttgart.de

Prof. Dr. phil. Bernd Geißel, Pädagogische Hochschule Ludwigsburg, Institut für Naturwissenschaften und Technik, Abteilung Technik, Reuteallee 46, D-71634 Ludwigsburg
E-Mail: geissel@ph-ludwigsburg.de

Prof. Dr. phil. habil. Reinhold Nickolaus, Universität Stuttgart, Institut für Erziehungswissenschaft und Psychologie, Abteilung Berufs-, Wirtschafts- und Technikpädagogik, Geschwister-Scholl-Straße 24D, D-70174 Stuttgart
E-Mail: nickolaus@bwt.uni-stuttgart.de

Franziska Perels

Modellierung und Messung fächerübergreifender Kompetenzen und ihre Bedeutung für die Bildungsforschung

Kritische Reflexion der Projektbeiträge

1. Einführung

Zielsetzung dieses Artikels ist es, die Beiträge der Projekte aus der Domäne „Fächerübergreifende Kompetenzen" des DFG-Schwerpunktprogramms hinsichtlich ihrer Relevanz für die Bildungsforschung zu kommentieren.

Um eine Einordnung der vorliegenden Beiträge in die (empirische) Bildungsforschung zu ermöglichen, ist es notwendig zu klären, welche Zielsetzungen und Inhalte mit diesem Forschungsbereich verbunden sind. Folgt man der gängigen Definition des Deutschen Bildungsrats, so wird Bildungsforschung definiert als die „Untersuchung der Voraussetzungen und Möglichkeiten von Bildungs- und Erziehungsprozessen im institutionellen und gesellschaftlichen Kontext" (Deutscher Bildungsrat 1974, S. 16). Prenzel (2006) wird bei seiner Beschreibung dessen, was Bildungsforschung ausmacht, noch konkreter: „Ihr Gegenstand umfasst Voraussetzungen, Prozesse und Ergebnisse von Bildung über die Lebensspanne, und zwar innerhalb wie außerhalb von (Bildungs-)Institutionen und im gesellschaftlichen Kontext. Ihr Anliegen ist es, die Bildungswirklichkeit zu verstehen und zu verbessern; …" (S. 73). Später im gleichen Aufsatz betont er weiterhin die Bedeutung der empirischen Methoden und macht deutlich, dass sich Bildungsforschung auf das aktuelle Bildungsgeschehen beziehen und auf dem neusten Stand der Methodenentwicklung empirisch fundierte Ergebnisse liefern solle. Des Weiteren wird übereinstimmend festgestellt (vgl. z.B. Schaffert/Schmid 2004 oder Tippelt/Schmid 2009), dass Bildungsforschung interdisziplinär angelegt sein sollte und neben Erziehungswissenschaft und Psychologie unter anderem auch Erkenntnisse und Methoden aus der Didaktik, Soziologie, Betriebswirtschaft und Philosophie integrieren sollte (vgl. z.B. Rindermann 2003). Dabei sollten qualitative und quantitative Herangehensweisen kombiniert werden, um verschiedene Perspektiven bezogen auf den Untersuchungsgegenstand einbringen zu können.

Die Diskussion der Beiträge des Schwerpunktprogramms aus der Domäne „Fächerübergreifende Kompetenzen" soll nun genau anhand dieser Aspekte der (empirischen) Bildungsforschung erfolgen. Dabei liegt der Fokus (a) auf der Relevanz der Ergebnisse für das aktuelle Bildungsgeschehen über die Lebensspanne im Sinne des Verstehens und Förderns von Bildungsprozessen und (b) auf der Bedeutung der verwendeten Methoden für die Bereitstellung empirisch fundierter Ergebnisse.

2. Einordnung der Beiträge und Bedeutung der Ergebnisse für die Bildungsforschung

Die sechs Beiträge aus der Domäne „Fächerübergreifende Kompetenzen" spannen entsprechend der Interdisziplinarität der Bildungsforschung den Bogen von der (Pädagogischen) Psychologie über die Lehr- Lernforschung und Didaktik bis zur Erziehungswissenschaft und Berufspädagogik. In den Projekten werden dabei sowohl qualitative als auch quantitative Verfahren angewendet. Dadurch werden sie der Vielseitigkeit des Forschungsgebiets gerecht und bilden eine gute Grundlage der Modellierung und Messung übergreifender Kompetenzen. Entsprechend der Zielsetzung des Schwerpunktprogramms beschäftigen sich alle Projekte mit der Beschreibung, Modellierung und Messung von Kompetenzen. Damit analysieren sie einen Kernbereich der Bildungsforschung mit einer Schlüsselfunktion für die Optimierung von Bildungsprozessen und die Weiterentwicklung des Bildungswesens (Klieme/Leutner 2006).

Betrachtet man die Beiträge dieser Domäne vor dem Hintergrund der *Relevanz der Ergebnisse für das aktuelle Bildungsgeschehen* genauer, so lassen sich grob zwei Bereiche unterscheiden. Während im Projekt *Berufspädagogik* sowie in einem Teil auch im Projekt *Problemlösen* der Schwerpunkt auf der Modellierung beruflicher Fachkompetenzen bzw. fachbezogener Problemlösekompetenzen liegt, sind in den anderen Projekten eher fächerübergreifende Kompetenzen im eigentlichen Sinne im Fokus der Betrachtung. Dabei wird zum einen Bezug genommen auf die fächerübergreifende Problemlösekompetenz, bei der wiederum zwischen analytischem Problemlösen (Projekt *Problemlösen*) und Modellen komplexen, dynamischen Problemlösens (Projekt *Dynamisches Problemlösen*) unterschieden wird. Auch der Bereich der Metakognition wird durch zwei Beiträge repräsentiert. Während das Projekt *EWIKO* sich mit dem metakognitiven Wissen in der Sekundarstufe beschäftigt, wird im Projekt *Bewertungskompetenz* auf eine Teilkompetenz des Modells für Bewertungskompetenz im Kontext Nachhaltiger Entwicklung fokussiert und die Teilkompetenz „Generieren und Reflektieren von Sachinformationen" empirisch überprüft. Schließlich bezieht sich das Projekt *Selbstregulationskompetenz* auf eine weitere Facette übergreifender Kompetenzen und entwickelt und evaluiert ein Kompetenzstrukturmodell zur Selbstregulation beim Lernen aus Sachtexten.

In der Zusammenschau der verschiedenen Forschungsthemen wird die große thematische Bandbreite der Arbeiten deutlich, die ein weites Spektrum überfachlicher Kompetenzen deutlich machen. Dabei sind die beschriebenen Kompetenzen von unterschiedlicher Bedeutsamkeit für die Bildungsforschung. So liegt die Modellierung von Selbstregulationskompetenz im Zentrum bildungswissenschaftlicher Untersuchungen (siehe Projekt *Selbstregulationskompetenz*). Eine Vielzahl von Studien belegen mittlerweile, dass eine hohe Selbstregulation mit hoher akademischer Leistung einhergeht (vgl. z.B. Fuchs u.a. 2003; Hidi/Ainley 2008). Selbstreguliert lernen zu können, stellt also eine wesentliche Voraussetzung für den Lernerfolg dar (vgl. Dembo/Eaton 2000). Aufgrund der Bedeutung dieser Kompetenz für „Bildungs- und Erziehungsprozesse" wird diese übergreifende Kompetenz auch in den internationalen Leistungsvergleichsstudien fokussiert (z.B. TIMSS oder PISA).

Von ähnlicher Zentralität für die Bildungsforschung ist auch die Problemlösekompetenz (siehe Projekte *Problemlösen* und *Dynamisches Problemlösen*). Aufgrund der breiten Forschungserfahrung zu ihrer Präzisierung und Operationalisierung, die zu einer besseren Klärung des Konstrukts beigetragen hat, wurde diese Kompetenz als übergreifende Fähigkeit in den PISA-Studien geprüft. Aus den Forschungsergebnissen zum Problemlösen wird deutlich, dass die Fähigkeit, Probleme zu lösen, eine wichtige Qualifikation in verschiedenen Lernbereichen bzw. Schulfächern ist.

Eng mit der Selbstregulation verknüpft ist die metakognitive Kompetenz bzw. das metakognitive Wissen, das durch zwei Projekte der Domäne abgebildet wird (Projekte *Bewertungskompetenz* und *EWIKO*). Auch wenn die Bedeutung dieser Kompetenz für schulisches Lernen und akademische Leistung nachweisbar ist, ist dieser Kompetenzbereich bezogen auf die Bildungsforschung eher randständig und wird (z.B. in den internationalen Schulleistungsstudien) zumeist unter übergreifenden Konzepten wie Selbstregulation subsumiert.

Ähnlich verhält es sich auch mit der Modellierung beruflicher Fachkompetenz, die für den Bereich der Berufspädagogik von Bedeutung ist, in ihrem Kern jedoch nicht unbedingt die zentralen Themen der empirischen Bildungsforschung berührt (siehe Projekt *Berufspädagogik*).

Betrachtet man die in den Projekten *verwendeten Methoden*, so zeigt sich eine ähnlich große Vielfalt wie bei den Disziplinen, die an den Projekten beteiligt sind, die entsprechend der Forderung Prenzels (2006) auf dem neusten Stand der Methodenentwicklung empirisch fundierte Ergebnisse liefern. Dabei wird der Bogen gespannt von qualitativen Verfahren (Qualitative Inhaltsanalyse nach Mayring im Projekt *Bewertungskompetenz*) über Methoden der klassischen Testtheorie (z.B. Berechnung von Beurteilerübereinstimmungen: Projekte *Problemlösen* bzw. *Selbstregulationskompetenz* und Berechnung der internen Konsistenz: Projekt *EWIKO*) sowie klassischen Evaluationsmethoden (z.B. Unterschiedstestung über *t*-Tests, verteilungsfreie Verfahren oder Varianzanalysen: Projekte *Problemlösen*, *EWIKO*, *Dynamisches Problemlösen*; Zusammenhangsanalysen z.B. Projekt *Selbstregulationskompetenz*) bis hin zu im Antrag zur Errichtung des Schwerpunktprogramms besonders hervorgehobenen Methoden der probabilistischen Testtheorie (z.B. Projekt EWIKO) und Strukturmodellierungen (z.B. Projekt *Berufspädagogik*). Diese sehr unterschiedlichen Verfahren sind den verschiedenen Fragestellungen geschuldet und machen deutlich, dass es gerade im Bereich der Modellierung und Messung von Kompetenzen innerhalb der Bildungsforschung notwendig ist, ein geeignetes Repertoire an empirischen Verfahren einzubeziehen. Zieht man jedoch in Betracht, dass die psychometrische Modellierung von Kompetenzen vor allem auf der Basis der Item-Response-Theorie erfolgen sollte, wären für die zukünftige Entwicklung der beschriebenen Projekte eine stärkerer Fokussierung auf diese Verfahren wünschenswert.

Orientiert man sich an den Darstellungen im Rahmenantrag des Schwerpunktprogramms, so zeigt sich innerhalb der Projekte zum Teil eine klare Strukturierung des Vorgehens entsprechend der Bereiche (1) Kompetenzmodelle; (2) Psychometrische Modelle; (3) Messkonzepte sowie (4) Nutzung von Diagnostik und Assessment. Zum Teil ist diese Strukturierung nicht auf den ersten Blick erkennbar, wenn z.B. Instrumente ent-

wickelt werden, bevor eine ausreichende theoretische und empirische Kompetenzmodellierung erfolgt ist.

3. Zusammenfassung und Fazit

Ingesamt ist festzuhalten, dass mit den Projekten in der Domäne „Fächerübergreifende Kompetenzen" des Schwerpunktprogramms Fragestellungen untersucht werden, die Relevanz für die Bildungsforschung insofern haben, als die behandelten Inhalte Bezüge zu den zentralen Themen der empirischen Bildungsforschung aufweisen und somit das zentrale Anliegen dieser Forschungsrichtung bedienen. Auch im Hinblick auf die verwendeten Methoden zeigt sich eine große Vielfalt, die auf die Fragestellungen und spezifischen Ziele der Projekte ausgerichtet sind und adaptiv entwickelt bzw. angewendet werden. Zusammenfassend kann also festgehalten werden, dass die Beiträge sowohl von theoretischer (im Sinne des oben erwähnten „Verstehens" von Bildungswirklichkeit) als auch von praktischer (im Sinne der „Verbesserung" der Bildungswirklichkeit) Relevanz für die empirische Bildungsforschung anzusehen sind.

Literatur

Dembo, M.H./Eaton, M.J. (2000): Self-regulation of academic learning in middle-level schools. In: The Elementary School Journal 100, S. 473–490.

Deutscher Bildungsrat (1974): Empfehlungen der Bildungskommission. Zu Neuordnung der Sekundarstufe II, 38. Sitzung der Bildungskommission, 13./14.02.74 in Bonn. Stuttgart.

Fuchs, L.S./Fuchs, D./Prentice, K./Burch, M./Hamlett, C.L./Owen, R./Schroeter, K. (2003): Enhancing third-grade students' mathematical problem solving with self-regulated learning strategies. Journal of Educational Psychology 95, H. 2, S. 306–315.

Hidi, S./Ainley, M. (2008): Interest and self-regulation: Relationships between two variables that influence learning. In: Schunk, D.H./Zimmerman, B.J. (Hrsg.): Motivation and self-regulated learning: Theory, research, and applications. Mahwah: Lawrence Erlbaum Associates Publishers, S. 77–109.

Klieme, E./Leutner, D. (2006): Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Überarbeitete Fassung des Antrags an die DFG auf Einrichtung eines Schwerpunktprogramms. http://kompetenzmodelle.dipf.de/pdf/rahmenantrag [05.08.2010].

Prenzel, M. (2006): Bildungsforschung zwischen Pädagogischer Psychologie und Erziehungswissenschaft. In: Merkens, H. (Hrsg.): Erziehungswissenschaft und Bildungsforschung. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 69–79.

Rindermann, H. (2003): Bildungsforschung. http://www-e.uni-magdeburg.de/methpsy/hr/Lehrmaterialien/bilduV.pdf [01.10.2004].

Schaffert, S./Schmidt, B. (2004): Inhalt und Konzeption der „bildungsforschung". In: bildungsforschung Jahrgang 1. http://www.bildungsforschung.org/Archiv/2004-01/einfuehrung [01.09.2009].

Tippelt, R./Schmidt, B. (2009): Einleitung des Herausgebers. In: Tippelt, R./Schmidt. B. (Hrsg.): Handbuch Bildungsforschung. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 9–19.

Anschrift der Autorin

Franziska Perels, Universität des Saarlandes, Campus A 5.4, D-66123 Saarbrücken
E-Mail: f.perels@mx.uni-saarland.de

Simone Bruder/Julia Klug/Silke Hertel/Bernhard Schmitz

Modellierung der Beratungskompetenz von Lehrkräften

Projekt Beratungskompetenz[1]

1. Fragestellung und theoretischer Ansatz

In der aktuellen pädagogisch-psychologischen Literatur, wie auch in den von der Kultusministerkonferenz herausgegebenen Standards für die Lehrerbildung, wird die Beratung von Eltern und Schüler/innen neben dem Unterrichten, Erziehen und Beurteilen als Kernaufgabe von Lehrkräften aufgeführt (vgl. Grewe 2005; KMK 2004; Schnebel 2007; Landesinstitut für Schule und Weiterbildung 1998). Mit der Formulierung eines entsprechenden Beratungsauftrags an Lehrkräfte wird das Ziel verfolgt, in Beratungsgesprächen eine Problemlösung für spezifische Fälle zu erarbeiten.

Schnebel (2007) weist darauf hin, dass die Anzahl an Situationen, in denen Beratung nötig ist, ansteigt. Gerade die Lernberatung gewinnt vor dem Hintergrund aktueller pädagogischer Innovationen und einem veränderten Verständnis von Lernen an Bedeutung. Lehrkräfte sind oft die ersten Ansprechpartner/innen der Eltern, wenn Schüler/innen Probleme mit dem Lernen haben. Sie erleben das Kind in der Schule und können einschätzen, wo Defizite vorliegen (vgl. Wild/Hofer 2002). Dies erfordert diagnostische Kompetenz von Lehrkräften, welche wiederum insbesondere für die Beratung relevant ist – denn ohne eine Feststellung der Lernvoraussetzungen und Lernprozesse ist eine individuelle Förderung kaum möglich. Die Analyse der Lernvoraussetzungen ist auch Grundlage jeder Lernberatung (vgl. KMK 2004).

Der Stellenwert von Beratungsaufgaben im Berufsalltag von Lehrkräften wird auch darin deutlich, dass die Beratung in aktuellen Modellen zur professionellen Handlungskompetenz von Lehrkräften integriert ist (vgl. z.B. Baumert/Kunter 2006). Das Beratungswissen ist hier neben dem *Fachwissen,* dem *fachdidaktischen Wissen*, dem *pädagogischen Wissen und* dem *Organisations- und Interaktionswissen* explizit als Wissensbereich aufgeführt. Dennoch wird die Beratungsarbeit in der Lehreraus- und -weiterbildung oft nur sehr wenig thematisiert. Meist sind Alltags- und Berufserfahrungen die

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: SCHM 1538/5-1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“.

Grundlage für Elterngespräche, selten basieren sie auf einer professionellen Beratungskompetenz der Lehrer (vgl. Landesinstitut für Schule und Weiterbildung 1998). Viele Lehrkräfte fühlen sich auf Beratungsgespräche nicht gut vorbereitet und dementsprechend unsicher in den Gesprächen (vgl. Hertel 2009). Dabei führt gerade die Situation, dass Lehrkräfte zugleich Berater/innen als auch Fach- und Klassenlehrer/innen (also Beurteiler) sind, häufig zu einer Unklarheit/Diffusion in der Lehrerrolle. Dies macht Beratung in der Schule besonders komplex (vgl. Landesinstitut für Schule und Weiterbildung 1989).

Aktuelle Studien betonen die Wichtigkeit der Beratungsarbeit von Lehrkräften und zielen mit Hilfe von Trainingsmaßnahmen auf die Förderung der Beratungskompetenz ab (vgl. Hertel 2009; Aich 2006). Dennoch fehlt bislang eine empirisch und theoretisch verankerte Definition des Konstrukts der Beratungskompetenz (vgl. Strasser/Gruber 2003; Hertel 2009). Insbesondere fehlen domänenspezifische Kompetenzstrukturmodelle, die eine profunde Basis zur Messung und auch zur gezielten Förderung der Beratungskompetenz leisten können (vgl. z.B. Hertel 2009). An dieser Stelle setzt unser Projekt zur Modellierung der Beratungskompetenz von Lehrkräften an.

Ziel unseres Projektes ist, das Konstrukt „Beratungskompetenz von Lehrkräften" zu modellieren und dabei sowohl die Domänenspezifität als auch die Entwicklung der Beratungskompetenz von Lehrkräften zu betrachten. Daher greifen wir die Idee eines domänenspezifischen Modellierungsansatzes auf und beziehen uns bei unseren Modellierungen auf die Domäne der Lernberatung, da diese ein zentrales Beratungsfeld für Lehrer/innen ist. Es sollen sowohl kognitive als auch handlungsbezogene Kompetenzbereiche einbezogen werden. Weiterhin soll der Einfluss von verschiedenen Prädiktoren wie Wissen, Selbstwirksamkeit, Teilnahme an Fortbildungen und Berufserfahrung auf die Beratungskompetenz überprüft werden, um auch hieraus Informationen für weitere Fördermaßnahmen ableiten zu können.

Folgende Fragestellungen lassen sich aus den Zielen der Studie ableiten:

1. Lässt sich ein aus der Theorie und Empirie entwickeltes Modell der Beratungskompetenz von Lehrkräften mit Hilfe von Strukturgleichungsmodellen empirisch nachweisen?
   Wir legen unserem Modell dabei eine fünfdimensionale Struktur der Beratungskompetenz von Lehrkräften zugrunde.

2. Haben Berufserfahrung, Teilnahme an Fortbildungen sowie Wissen im Bereich von Beratung und selbstreguliertem Lernen einen Einfluss auf die Beratungskompetenz?
   Bezogen auf diese Fragestellung ist es Ziel der Studie, zu überprüfen, welche Aspekte auf den Erwerb von Beratungskompetenz einen Einfluss nehmen. Es wird angenommen, dass das Wissen über Beratung und selbstreguliertes Lernen, die Teilnahme an Fortbildungen zu Beratung, die Berufserfahrung und auch die beratungsspezifische Selbstwirksamkeit einen Einfluss auf die Beratungskompetenz haben.

Diese Fragestellung ist insbesondere im Hinblick auf weitere Studien zentral, in denen die Beratungskompetenz über die Berufslaufbahn modelliert werden soll.

Aufbauend auf den Ergebnissen dieser ersten Studie soll dann in folgenden Studien die Struktur von Beratungskompetenz sowie die Zusammenhänge mit der Beratungsleistung für unterschiedliche Expertisestufen in der Berufslaufbahn von Lehrpersonen untersucht werden. Darauf aufbauend soll im Sinne des Aptitude-Treatment-Interaktion-Ansatzes eine Kompetenzdiagnostik-basierte Empfehlung von Fortbildungen gegeben werden können.

2. Methodisches Vorgehen

Die Formulierung der Kompetenzdimensionen erfolgte auf Basis der aktuellen Literatur zum Thema Beratung (vgl. Hertel 2009; Strasser/Gruber 2003; Schwarzer/Buchwald 2006; West/Cannon 1988). So beschreiben Strasser und Gruber (2003) Beratungskompetenz als fachliches Wissen um Sachverhalte und um die Wirksamkeit von Maßnahmen, welches auf der Grundlage personaler Ressourcen und reflektierter Erfahrung erlaubt, Wissen situationsangemessen und effektiv anzuwenden, was dann zu beraterischem Erfolg, also dem Erreichen der im Beratungsprozess gesetzten Ziele führt. Schwarzer und Buchwald (2006) beschreiben neben dem Fachwissen und den personalen Ressourcen zusätzlich noch vier weitere Dimensionen der Beratungskompetenz: die soziale Kompetenz, die Berater-Skills, die Bewältigungskompetenz und die Prozesskompetenz.

West und Cannon (1988) haben in einer umfassenden Studie Beratungsexperten gefragt, welche Bereiche der Beratung sie als zentral für den Erfolg ansehen. Hier wurden die zwischenmenschliche Kommunikation, die Berücksichtigung der Gleichstellung und Wertvorstellungen, persönliche Merkmale der Beraterin/des Beraters, gemeinschaftliches Problemlösen und die Evaluation der Effektivität der Beratung genannt.

Das Modell von Hertel (2009) integriert die Modelle von Strasser und Gruber (2003) und Schwarzer und Buchwald (2006). Es wurden dabei in einer empirischen Studie die Zusammenhänge folgender Kompetenzdimensionen überprüft: personale Ressourcen, soziale Kooperationskompetenz, Berater-Skills und pädagogisches Wissen, Prozesskompetenz sowie Bewältigungskompetenz. Hertel (2009) testete dabei dieses fünfdimensionale Modell mittels konfirmatorischer Faktorenanalysen gegen ein eindimensionales Modell. Dabei zeigte sich, dass ein mehrdimensionales Modell besser zur Beschreibung der Beratungskompetenz geeignet ist als ein eindimensionales.

Aufbauend auf der Literatur und den Ergebnissen der Studie von Hertel (2009) wird ein Modell mit den fünf Dimensionen *Berater-Skills*, *Ressourcen- und Lösungsorientierung*, *Pädagogisches Wissen und Diagnostizieren*, *Kooperation* sowie *Bewältigung* postuliert. Die *Berater-Skills* umfassen die Bereiche des adäquaten Gesprächsaufbaus und des Einsatzes von Gesprächsstrategien (aktives Zuhören, Paraphrasieren). Die *Ressourcen- und Lösungsorientierung* beinhalten die Ziel-, Lösungs- und Ressourcenorientierung in einer Beratung, das *Pädagogische Wissen und Diagnostizieren* das di-

agnostische Handeln (Problemdefinition, Ursachensuche) und das Strategiewissen (Lernstrategien). Die *Kooperation* umfasst das kooperative Handeln im Umgang mit den Ratsuchenden und die *Bewältigung* die Fähigkeit, in Elterngesprächen mit Kritik und schwierigen Beratungssituationen umgehen zu können. Es kann davon ausgegangen werden, dass es für den Erwerb von Beratungskompetenz wichtig ist, dass die Lehrkräfte zunächst Wissen über *Berater-Skills* und *Ressourcen- und Lösungsorientierung* haben. Die *Ressourcen- und Lösungsorientierung* und auch die *Kooperation* umfassen zudem eine systemische Sichtweise auf das Problem und die Lösung. Das *Diagnostizieren* im Beratungsprozess beschäftigt sich mit dem Problem, hierbei geht es um eine Definition und Ursachensuche. Als Variable der Person der Beraterin/des Beraters geht die *Bewältigung*, d.h. der Umgang mit schwierigen Beratungssituationen, in das Modell ein.

Das Kompetenzmodell wird mit Hilfe von konfirmatorischen Faktorenanalysen empirisch überprüft. Die Kompetenzstruktur/Kompetenzdimensionen werden sowohl auf Basis eines Fragebogens zur Selbsteinschätzung als auch auf Basis eines handlungsnahen Fallszenarios überprüft.

2.1 Instrumente

Die Messung der Beratungskompetenz erfolgte multimethodal mit den drei folgenden Instrumenten:

1. Selbsteinschätzung (Fragen zu allen Skalen der fünf Dimensionen)
2. Fallszenario (Fragen zu allen Skalen der fünf Dimensionen)
3. Wissenstest (Multiple-Choice-Test zur Beratung und zum selbstregulierten Lernen)

Der Selbsteinschätzungsteil bestand aus 64 Fragen zu den fünf Dimensionen und aus 38 Fragen zu Einstellungen und Überzeugungen zu dem Thema Beratung (Selbstwirksamkeit in Bezug auf Beratung, Empathiefähigkeit, Motivation zur Beratung und Interesse an Beratung sowie Erfahrung mit Beratung). Es wird angenommen, dass diese Variablen wichtige Prädiktoren für die Beratungskompetenz sind.

Das Fallszenario bestand aus einer ausführlichen Beschreibung eines Falls mit 10 offenen Fragen, in denen nach dem Beratungshandeln gefragt wurde. Die einzelnen Fragen wurden entsprechend der fünf Dimensionen der Beratungskompetenz konstruiert. Zur Auswertung wurde ein ausführliches Handbuch zur Beurteilung der offenen Antworten entwickelt, in dem die genaue Durchführung der Bewertung der Antworten festgehalten war.

Der Wissenstest bestand aus 8 Fragen zur Beratung und 9 Fragen zum selbstregulierten Lernen. Er war in einem Multiple-Choice Format konstruiert, sodass es pro Frage vier Antwortmöglichkeiten gab.

2.2 Stichprobe

Die Messung der Beratungskompetenz wurde an 141 Gymnasiallehrkräften durchgeführt. Von 20 Schulen, die angefragt wurden, erklärten sich fünf Schulen aus verschiedenen hessischen Landkreisen bereit, an der Studie teilzunehmen.

Von den 141 Lehrkräften mussten 16 Personen wegen zu vieler fehlender Angaben aus den Analysen ausgeschlossen werden, sodass die Gesamtstichprobe aus 125 Lehrkräften bestand. Die deskriptiven Daten der Lehrkräftestichprobe können Tabelle 2 entnommen werden.

N		**125**
Geschlecht	Männlich	36 (31,0%)
	Weiblich	87 (46,0%)
Jahre im Schuldienst	1–10 Jahre	68 (54,4%)
	11–20 Jahre	23 (18,4%)
	21–30 Jahre	19 (15,2%)
	> 31 Jahre	13 (10,4)
An Fortbildungen zu dem Thema Beratung teilgenommen?	Ja	31 (24,8%)
	Nein	90 (72,0%)
	keine Angaben	4 (3,2%)
Erreichte Punkte im Wissenstest	0–3	4 (3,2%)
	4–7	18 (14,4%)
	8–11	73 (58,4%)
	12–15	30 (24,0%)

Tab. 1: Kennwerte der Lehrerstichprobe

3. Ergebnisse

3.1 Ergebnisse Fragestellung 1

Die erste Fragestellung zielte auf die domänenspezifische Modellierung der Beratungskompetenz. Das postulierte fünfdimensionale Modell wurde mittels konfirmatorischer Faktorenanalysen für die Selbsteinschätzung und für das Fallszenario überprüft. Es ergaben sich sowohl für die Selbsteinschätzung ($Chi^2 = 29.46$, $df = 21$ ($p = .1$); CFI = .99; RMSEA = .06; SRMR = .04) als auch für das Fallszenario ($Chi^2 = 25.32$, $df = 34$ ($p = .86$); CFI = 1.00; RMSEA = .00; SRMR = .05) sehr gute Modellpassungen für das

fünfdimensionale Modell. Für die Seite der Selbsteinschätzung konnte das Modell bestätigt werden, alle Skalen laden hypothesenkonform auf der entsprechenden Dimension. Auf Seiten des Fallszenarios konnte die postulierte Faktorenstruktur nicht gefunden werden. In der letztlich herangezogenen Modellversion lädt die Lösungsorientierung auf der Dimension *Pädagogisches Wissen und Diagnostizieren*; auf der Dimension *Berater-Skills* laden nur die Gesprächsstrukturierung und das Paraphrasieren, das aktive Zuhören jedoch nicht. Es zeigt sich auch kein Zusammenhang zwischen dem aktiven Zuhören und einer anderen Dimension der Beratungskompetenz. Abbildung 1 stellt die Ergebnisse der konfirmatorischen Faktorenanalyse für die Selbsteinschätzung (Fragebogen zur Beratungskompetenz), Abbildung 2 für das Fallszenario dar.

Anschließend wurde die Beratungskompetenz (gemessen mit dem Fallszenario) mittels eines Strukturgleichungsmodells durch die selbsteingeschätzte Beratungskompetenz validiert. Auch in diesem Modell wurden die fünf Dimensionen als Indikatoren für die Beratungskompetenz verwendet.

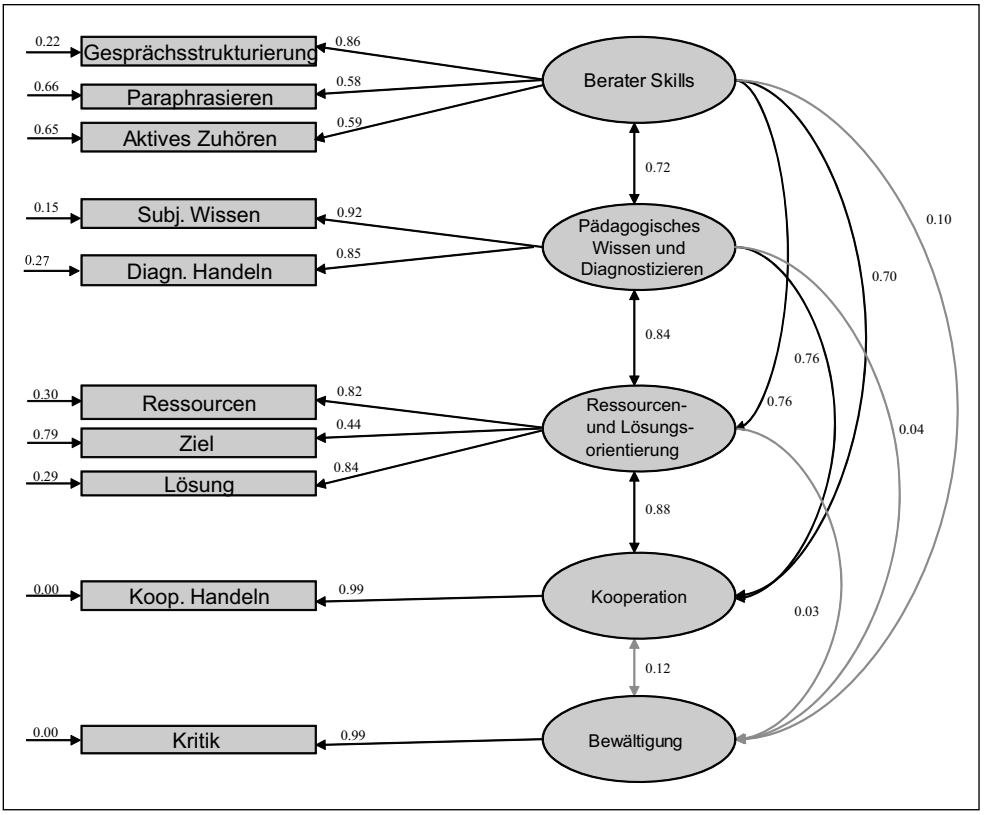


Abb. 1: Konfirmatorische Faktorenanalyse des fünfdimensionalen Modells der Beratungskompetenz auf Basis der Selbsteinschätzung

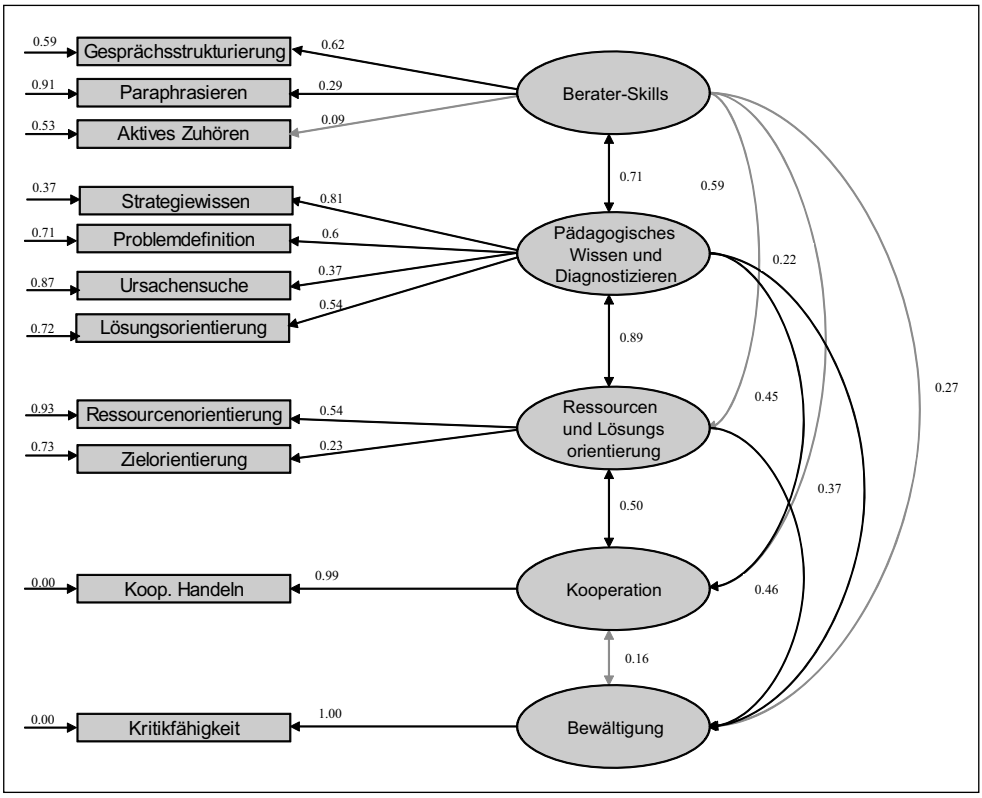


Abb. 2: Konfirmatorische Faktorenanalyse des fünfdimensionalen Modells der Beratungskompetenz auf Basis des Fallszenarios

Als zusätzliche Prädiktoren wurden die beratungsbezogene Selbstwirksamkeitserwartung, das Wissen über Beratung und Lernstrategien, die Berufserfahrung und die Teilnahme an Fortbildungen mit in das Modell einbezogen. Die Selbstwirksamkeit, die typischerweise durch Items zur Selbsteinschätzung erfasst wird, wurde als Prädiktor für die Selbsteinschätzung miteinbezogen. Die Berufserfahrung, das Wissen und die Teilnahme an Fortbildungen gingen als Prädiktoren für das Fallszenario in das Modell mit ein, da hier ein direkter Bezug zur Handlung zu erwarten ist. Um die Variablenanzahl zu reduzieren, gingen in die Berechnung des Strukturmodells für beide Instrumente nur die Dimensionsmittelwerte ein. Die Kennwerte weisen auf einen guten Modellfit hin ($Chi^2 = 94.84$, $df = 64$ ($p = .01$); CFI = .94; RMSEA = .06; SRMR = .07). Lediglich die Dimension *Bewältigung* zeigt keinen Zusammenhang zu der latenten Variable Beratungskompetenz auf der Selbsteinschätzungsseite. Abbildung 3 stellt das Modell zur Validierung des Fallszenarios durch die Selbsteinschätzung dar. Es zeigt sich, dass die Ergebnisse im Fallszenario signifikant durch die Selbsteinschätzung vorhergesagt werden. Entgegen den Erwartungen erreichten die Lehrkräfte mit weniger Berufserfahrung signifikant bessere Werte im Fallszenario als Lehrkräfte mit mehr Berufserfahrung. Das

Wissen, die Teilnahme an Fortbildungen und auch die Selbstwirksamkeit bezüglich der eigenen Beratungsleistungen erweisen sich wie angenommen als signifikante Prädiktoren für die Beratungskompetenz.

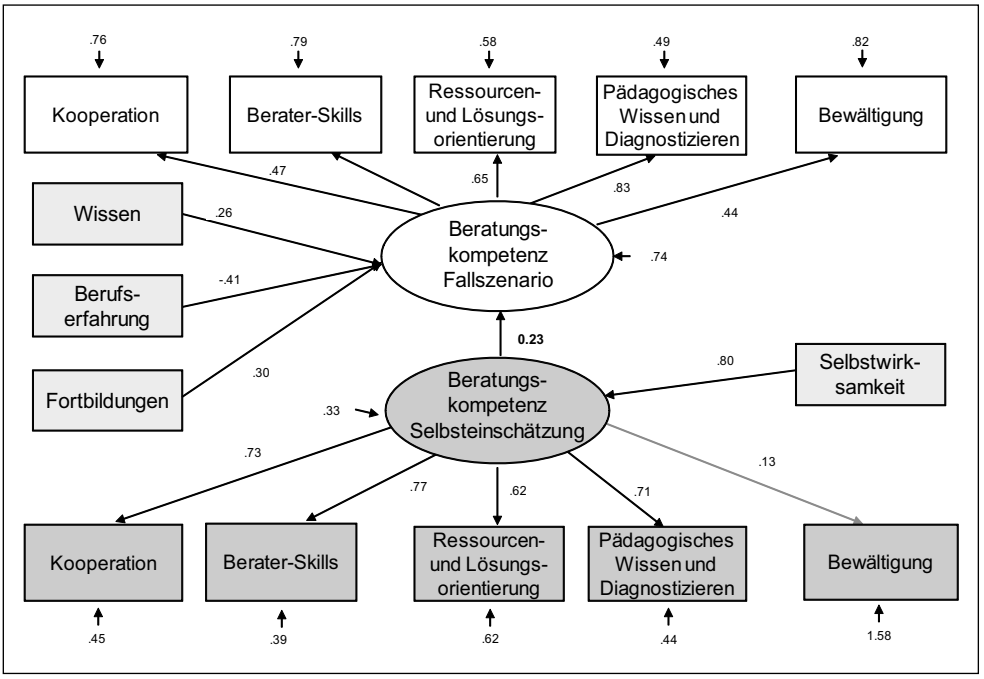


Abb. 3: Strukturgleichungsmodell zur Vorhersage der Beratungskompetenz gemessen mit dem Fallszenario durch die selbst eingeschätzte Beratungskompetenz

3.2 Ergebnisse Fragestellung 2

Zur Überprüfung des Einflusses der Berufserfahrung, des Wissens, der Selbstwirksamkeit und auch der bisherigen Teilnahme an Fortbildungen zu Beratung auf die Beratungskompetenz gemessen mit dem Fallszenario wurden multiple Regressionsanalysen mit den genannten Variablen als Prädiktoren durchgeführt. Als Kriterien gingen die einzelnen Kompetenzdimensionen und der Gesamtwert im Fallszenario in die Analysen ein. Tabelle 3 zeigt die ß-Gewichte, die Varianzaufklärung und die Signifikanzen für die einzelnen Kriteriumsvariablen. Signifikante Korrelationen bestehen zwischen 1) den Prädiktoren Berufserfahrung und Selbstwirksamkeit ($r = .23$, $p = .01$) und 2) den Prädiktoren Teilnahme an Fortbildungen und Wissen ($r = .23$, $p = .05$).

Die Prädiktorvariable Berufserfahrung sagt die Beratungskompetenz für das *Fallszenario-Gesamt*, die *Berater-Skills*, die *Ressourcen- und Lösungsorientierung*, die *Bewältigung* und das *Pädagogische Wissen und Diagnostizieren* vorher. Lehrer/innen mit

weniger Berufserfahrung weisen bezüglich dieser Kompetenzdimensionen signifikant bessere Werte im Fallszenario auf als Lehrkräfte mit mehr Berufserfahrung.

Das Wissen über Beratung und selbstreguliertes Lernen beeinflusst die Beratungskompetenz bezüglich des Gesamtwertes im Fallszenario, der *Berater-Skills*, der *Kooperation* und dem *Pädagogischen Wissen und Diagnostizieren*. Lehrkräfte, die höhere Werte im Wissenstest erzielt haben, erreichten auch auf diesen Dimensionen des Fallszenarios höhere Werte.

Bezüglich der Selbstwirksamkeit ergeben sich signifikante Einflüsse auf den Gesamtwert im Fallszenario, auf die Dimension *Ressourcen- und Lösungsorientierung* und auf die Dimension *Pädagogisches Wissen und Diagnostizieren*. Lehrer/innen, die eine höhere Selbstwirksamkeit angaben, erzielten hier bessere Werte.

Für die Teilnahme an Fortbildungen zeigt sich ein mäßiger Effekt bezüglich der Dimension Berater-Skills. Lehrkräfte, die an einer Fortbildung zu Beratung teilgenommen haben, weisen hier höhere Werte auf der Dimension *Berater-Skills* auf.

Kriteriums-variablen	**Fall-szenario Gesamt**	**Berater-Skills**	**Koope-ration**	**Ressourcen- & Lösungs-orientierung**	**Bewäl-tigung**	**Päd. Wissen & Diagnosti-zieren**
R^2	0.15*	0.17**	0.08	0.10*	.09#	0.18**
Prädiktoren				**β**		
Berufs-erfahrung	−.32***	−.32***	−.14	−.20*	−.31***	−.34***
Wissen	.15#	.19*	.23**	.10	.04	.21**
Selbstwirk-samkeit	.22*	.06	.10	.23**	.08	.21**
Teilnahme Fort-bildungen	.03	.15#	−.06	.11	.04	−.11

#$p < .10$; *$p < .05$; **$p < .01$; *** $p < .001$

Tab. 2: Regressionsanalysen

4. Diskussion und Erkenntnisgewinn

Die Ergebnisse der Studie zeigen, dass sich bei der untersuchten Stichprobe von 125 Lehrkräften das postulierte fünfdimensionale Modell gut zur Beschreibung des Konstrukts der Beratungskompetenz eignet. Die Modellfits zeigen sowohl für das Fallszenario als auch für die Selbsteinschätzungen sehr gute Werte. Auf Seiten des Fallszenarios ergaben sich bei der Modellierung Schwierigkeiten bezüglich der Lösungsorientierung und des aktiven Zuhörens. Die Lösungsorientierung lädt auf der Dimension *Pädagogisches Wissen und Diagnostizieren* und nicht wie erwartet auf der Dimension *Ressourcen- und Lösungsorientierung*. Dies kann daran liegen, dass die inhaltliche Frage zur Lösungsorientierung sehr nah an der Skala Ursachensuche (Dimension *Pädagogisches Wissen und Diagnostizieren*) angeschlossen war, sodass es für die Lehrkräfte nahe lag, hier ähnliche Antworten zu geben. In der weiterführenden Studie wird dies berücksichtigt und die Instrumente werden optimiert, sodass eine klare Trennung zwischen diesen beiden Skalen vorliegt. Bezüglich des aktiven Zuhörens kann angenommen werden, dass dieses nicht auf der Dimension *Berater-Skills* lädt und auch mit keiner anderen Skala korreliert, weil sich aktives Zuhören methodisch durch das Fallszenario nur schwer erfassen lässt. Es wird bei der Überarbeitung darauf geachtet, dass diese Gesprächsstrategie genauer erfasst wird.

Bezüglich des Gesamtmodells zeigt sich, dass auf Seiten der Selbsteinschätzung die *Bewältigung* nicht auf der latenten Variable Beratungskompetenz lädt. Eine mögliche Erklärung hierfür kann die geringere Reliabilität auf der Skala Kritikfähigkeit in der Haupterhebung sein. Das Wissen und die Teilnahme an Fortbildungen und auch die Selbstwirksamkeit können im Gesamtmodell als zentrale Prädiktoren für eine erfolgreiche Beratung angesehen werden. Für die Berufserfahrung ergab sich in dieser Studie ein Effekt entgegengesetzt der Erwartungen derart, dass jüngere Lehrkräfte signifikant bessere Werte im Fallszenario erzielten. Dies kann aus mehreren Perspektiven diskutiert werden. Da sich die Ausbildungssituation verbessert hat und das Thema der Beratung und Gesprächsführung immer mehr in die Studienseminare integriert wird (vgl. Hertel/Bruder/Schmitz 2009), kann es sein, dass die jüngeren Lehrkräfte bereits besser in Beratung ausgebildet sind als ältere Lehrkräfte. Weiterhin kann auch eine fehlende Motivation beim Ausfüllen des Fallszenarios nicht ausgeschlossen werden, d.h. dass jüngere Lehrkräfte das Szenario mit dem offenen Antwortformat gewissenhafter ausgefüllt haben könnten. In folgenden Studien soll überprüft werden, ob sich dieses Ergebnis replizieren lässt. Für die Weiterentwicklung des Modells, insbesondere hinsichtlich der Modellierung der Beratungskompetenz über die Berufslaufbahn, sollte das Fallszenario daher unbedingt um ein geschlossenes Antwortformat, wie es in Situational-Judgement-Tests vorliegt, erweitert werden, sodass Motivationseffekte ausgeschlossen werden können.

Die Resultate der Regressionsanalysen untermauern die Ergebnisse der Modellierung insbesondere bezüglich der Berufserfahrung. Es zeigt sich auch hier, dass Lehrkräfte mit weniger Berufserfahrung in fast allen Bereichen den Lehrkräften mit mehr Berufserfahrung überlegen sind. Hier gilt es in folgenden Studien zu überprüfen, ob

weitere Variablen wie z.B. die reflektierte Erfahrung oder die Fortbildungsbereitschaft einen Einfluss auf diesen Effekt nehmen. Hinsichtlich des Wissens zeigt sich, dass sich dieses in den Dimensionen *Berater-Skills*, *Pädagogisches Wissen und Diagnostizieren* und *Kooperation* niederschlägt. Interessant ist das Ergebnis, dass Lehrkräfte, die eine höhere beratungsspezifische Selbstwirksamkeit haben, insbesondere bezüglich der *Ressourcen- und Lösungsorientierung* profitieren. Wer also über mehr Selbstwirksamkeit verfügt, denkt ressourcen-, ziel- und lösungsorientierter. Bei der Teilnahme an Fortbildungen konnte durch die Analysen gezeigt werden, dass diese einen wichtigen Beitrag zur Verbesserung der Gesprächsführungsstrategien leisten. Teilnehmer/innen, die angaben, an Fortbildungen teilgenommen zu haben, erreichten signifikant bessere Ergebnisse in diesem Bereich. Dies untermauert die Annahme, dass die Beratungskompetenz gefördert werden kann und auch weiterhin gefördert werden sollte. Eine Integration in die Lehreraus- und -weiterbildung sollte weiter voran getrieben werden. Weiterführende Studien sollten die reflektierte Erfahrung mit einbeziehen, weil nicht nur die Erfahrung allein, sondern gerade die Reflexion über gemachte Erfahrungen zentral für den Erwerb von Kompetenzen ist. Eine grundlegende Wissensbasis dient dabei allerdings als Ausgangspunkt für die Entwicklung von Routinen und hilft, die eigene Praxis zu reflektieren (vgl. Strasser 2006). Dies zeigt sich auch in den Ergebnissen zum Wissenstest. Wissen über Beratung und domänenbezogen über selbstreguliertes Lernen, scheint eine notwendige Vorraussetzung für die Beratungskompetenz zu sein.

Wissen ist als notwendige Grundlage für den Erwerb von Handlungskompetenzen zu sehen, die sich im weiteren Verlauf durch Übung und Reflexion herausbilden und festigen (vgl. Strasser/Gruber 2003; Vonken 2005; Strasser 2006).

In einer nächsten Studie soll aufbauend auf den Erkenntnissen dieser Studie die Struktur von Beratungskompetenz sowie die Zusammenhänge mit der Beratungsleistung für unterschiedliche Expertisestufen in der Berufslaufbahn von Lehrpersonen untersucht werden. Die bisherigen Ergebnisse deuten darauf hin, dass eine Modellierung der Beratungskompetenz über die Berufslaufbahn von Lehrkräften einen vielversprechenden Ansatz darstellt.

Literatur

Aich, G. (2006): Kompetente Lehrer: Ein Konzept zur Verbesserung der Konflikt- und Kommunikationsfähigkeit. Hohengehren: Schneider.

Baumert, J./Kunter, M. (2006): Stichwort: Professionelle Kompetenz von Lehrkräften. In: Zeitschrift für Erziehungswissenschaft 9, H. 4, S. 469–520.

Grewe, N. (2005): Der Beratungsalltag des Lehrers. Anlässe – Erfahrungen – Hilfen. In: Pädagogik 6, S. 10–13.

Hertel, S. (2009): Beratungskompetenz von Lehrern. Kompetenzdiagnostik, Kompetenzförderung und Kompetenzmodellierung. Münster: Waxmann.

Hertel, S./Bruder, S./Schmitz, B. (2009): Beratungs- und Gesprächsführungskompetenz von Lehrkräften. In: Zlatkin-Troitschanskaia, O./Beck, K./Sembill, D./Nickolaus, R./Mulder, R. (Hrsg.): Lehrprofessionalität – Bedingungen, Genese, Wirkungen und ihre Messung. Weinheim u.a.: Beltz, S. 117–129.

Kultusministerkonferenz (KMK) (2004): Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004. http://www.kmk.org/filead min/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf [01.09.2009].

Landesinstitut für Schule und Weiterbildung (^{3}1998): Fachgutachten: Beratung in der Schule und im Schulsystem. Ergebnisse einer Überprüfung und Anregungen zur weiteren Entwicklung. Bönen: Verlag für Schule und Weiterbildung, Druck Verlag Kettler.

Schnebel, S. (2007): Professionell beraten. Beratungskompetenz in der Schule. Weinheim: Beltz.

Schwarzer, C./Buchwald, P. (2001): Beratung. In: Krapp, A./Weidenmann, B. (Hrsg.): Pädagogische Psychologie. Weinheim u.a.: Beltz, S. 565–600.

Strasser, J. (2006): Erfahrung und Wissen in der Beratung. Theoretische und empirische Analysen zum Entstehen professionellen Wissens in der Erziehungsberatung. Göttingen: Cuvillier.

Strasser, J./Gruber, H. (2003): Kompetenzerwerb in der Beratung: Eine kritische Analyse des Forschungsstands. In: Psychologie in Erziehung und Unterricht 50, S. 381–399.

Vonken, W. (2005): Handlung und Kompetenz. Wiesbaden: VS Verlag für Sozialwissenschaften.

West, J.F./Cannon, G.S. (1988): Essential Collaborative Consultation Competencies for Regular and Special Educators. In: Journal of Learning Disabilities 21, H. 1, S. 56–63.

Wild, E./Hofer, M. (2002): Familie mit Schulkindern. In: Hofer, M./Wild, E./Noack, P. (Hrsg.): Lehrbuch Familienbeziehungen. Eltern und Kinder in der Entwicklung. Göttingen: Hogrefe, S. 216–240.

Anschrift der Autor/innen

Dipl.-Psych. Simone Bruder, Technische Universität Darmstadt, Institut für Psychologie, Alexanderstraße 10, D-64283 Darmstadt
E-Mail: bruder@psychologie.tu-darmstadt.de

Dipl.-Psych. Julia Klug, Technische Universität Darmstadt, Alexanderstraße 10, Institut für Psychologie, D-64283 Darmstadt
E-Mail: klug@psychologie.tu-darmstadt.de

Dr. Dipl.-Psych. Silke Hertel, Deutsches Institut für Internationale Pädagogische Forschung, Schlossstraße 29, D-60486 Frankfurt
E-Mail: hertel@dipf.de

Prof. Dr. Bernhard Schmitz, Technische Universität Darmstadt, Institut für Psychologie, Alexanderstraße 10, D-64283 Darmstadt
E-Mail: schmitz@psychologie.tu-darmstadt.de

Cornelia Gräsel/Sabine Krolak-Schwerdt/Ines Nölle/Thomas Hörstermann

Diagnostische Kompetenz von Grundschullehrkräften bei der Erstellung der Übergangsempfehlung

Eine Analyse aus der Perspektive der sozialen Urteilsbildung

Projekt Diagnostische Kompetenz[1]

1. Der Übergang am Ende der Grundschulzeit auf die weiterführende Schule

Die Wahl der weiterführenden Schulform am Ende der Grundschulzeit stellt im deutschen Bildungssystem eine wichtige Entscheidung für die schulische und berufliche Zukunft der Schülerinnen und Schüler[2] dar. Bei dieser folgenreichen Entscheidung kommt nicht nur dem elterlichen Bildungswunsch, sondern auch der von den Grundschullehrkräften erteilten Übergangsempfehlung eine bedeutende Rolle zu. Zum einen wird die Übergangsempfehlung der Lehrer häufig von den Eltern übernommen (vgl. Stubbe/Bos 2008), zum anderen ist die Lehrerempfehlung in einigen Bundesländern bindend und kann nur in Ausnahmefällen umgangen werden. Dementsprechend ist die Erstellung einer angemessenen Übergangsempfehlung eine verantwortungsvolle Aufgabe der Grundschullehrkräfte, die hohe diagnostische Kompetenz erfordert.

Die Bildungsforschung untersuchte bereits in den 70er und 80er Jahren, welche Faktoren die Übergangsentscheidungen am Ende der Grundschulzeit beeinflussen und inwieweit diese Entscheidungen zur Entstehung von sozialen Disparitäten beitragen (vgl. z.B. Becker 2004). Mit der Thematisierung der im Kontext der internationalen Schulleistungsstudien erneut aufgezeigten Chancenungleichheit im deutschen Bildungssystem rückte in den letzen Jahren wieder verstärkt der Übergang von der Grundschule in die weiterführende Schule in den Blickpunkt der Forschung (vgl. zum Überblick Maaz u.a. 2006). Vergleichsweise viele der anschließenden Forschungsarbeiten befassten sich mit den sozialschichtspezifischen elterlichen Bildungsaspirationen und deren Beitrag zur sozialen Selektivität am Übergang zur weiterführenden Schule (vgl. z.B. Becker 2004). Nur wenige Studien untersuchten dagegen bisher die Lehrerentscheidungen (vgl. z.B. Bos u.a. 2004; Ditton/Krüsken 2006). Übereinstimmend zeigen die verschiedenen Studien, dass sich Lehrkräfte bei der Übergangsentscheidung primär an Leistungsmerk-

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: GR 1883/5-1; KR 2162/4-1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

2 Aus Gründen der besseren Lesbarkeit und der erforderlichen Kürze wird im Folgenden nur die männliche Form verwendet, wobei die weiblichen Personen mitgemeint sind.

malen orientieren, insbesondere an den Noten in den Kernfächern. Darüber hinaus belegen diese Studien, dass Merkmale der sozialen Herkunft, wie z.B. der sozioökonomische Status, die Lehrerempfehlung der Lehrkräfte beeinflussen.

In diesen Studien wurde hauptsächlich geprüft, welche Merkmale des Schülers und des Elternhauses mit der erteilten Übergangsempfehlung zusammenhängen. Hierbei wurden häufig die Merkmale der Schüler und deren Eltern mittels einer Schüler- und/oder Elternbefragung oder Tests erfasst und anschließend auf systematische Zusammenhänge mit der Lehrerempfehlung hin untersucht (vgl. z.B. Bos u.a. 2004). Welche *Merkmale aus Sicht der Lehrkräfte* bei der Erstellung der Übergangsempfehlung eine bedeutende Rolle spielen, kann aus den bisherigen Befunden nur indirekt geschlossen werden (vgl. van Ophuysen 2006). Weiterhin gibt es kaum Theorien und Befunde dazu, wie Lehrkräfte zu ihren Übergangsempfehlungen gelangen, d.h. wie sie die einzelnen Schülerinformationen zu einer Entscheidung bezüglich einer angemessenen Schullaufbahnempfehlung zusammenführen. Die Analyse der Perspektive der Lehrkräfte und deren Entscheidungsfindung bei der Erstellung der Übergangsempfehlung stehen im Vordergrund unseres Projektes. Im hier vorliegenden Artikel wird einerseits ein Überblick über das Gesamtprojekt gegeben. Andererseits stellen wir die zentralen Befunde der Vorstudien aus der bisherigen Projektarbeit vor.

2. Theoretischer Hintergrund des Projektes

Zur Analyse der Entscheidungsprozesse der Lehrkräfte bei der Erstellung der Übergangsempfehlung werden im Projekt sozial kognitive Modelle der Urteilsbildung und Entscheidungsfindung herangezogen. Insbesondere orientieren wir uns an den theoretischen Annahmen dualer Prozessmodelle und übertragen sie auf den Kontext der Übergangsentscheidung (vgl. z.B. Fiske/Neuberg 1990; Chen/Chaiken 1999). Diese Modelle gehen davon aus, dass Urteile und Entscheidungen mittels einer informationsintegrierenden oder einer heuristischen Informationsverarbeitungsstrategie getroffen werden können. Diese beiden Verarbeitungsarten zeichnen sich sowohl durch eine unterschiedliche Informationssuche als auch durch eine unterschiedliche Integration der Information in der Urteilsbildung aus. Die informationsintegrierende Verarbeitung ist durch eine aufwändige Informationssuche gekennzeichnet, in der alle entscheidungsrelevanten Informationen berücksichtigt sowie systematisch gewichtet und integriert werden. Im Gegensatz dazu ist die heuristische Verarbeitung durch die Verwendung von vereinfachten Entscheidungsregeln, sogenannten kognitiven Heuristiken charakterisiert (vgl. Tversky/Kahneman 1974). In der Entscheidungsfindung wird nur knappste Information berücksichtigt. Heuristische Informationsverarbeitung verläuft daher mit geringem kognitivem Aufwand und effizient.

Die Verwendung beider Verarbeitungsarten wurde bereits im schulischen Kontext bei der Beurteilung von Schülern aufgezeigt (vgl. u.a. Krolak-Schwerdt/Rummer 2005). Unser Ziel ist es, die Verarbeitungsprozesse im Kontext der Übergangsentscheidung zu untersuchen. Dabei gehen wir davon aus, dass die aufwändige, integrative Verarbeitung

aller relevanten Informationen für diese komplexe und verantwortungsvolle Aufgabe die angemessene Strategie der Entscheidungsfindung darstellt und hoher diagnostischer Kompetenz entspricht. Im Gegensatz dazu führt die heuristische Verarbeitung aufgrund ihrer begrenzten Informationssuche und verkürzten Entscheidungsfindung weniger verlässlich zu guten Beurteilungen. Dementsprechend wird in unserem Projekt die diagnostische Kompetenz von Grundschullehrkräften bei der Erstellung der Übergangsempfehlung als die Fähigkeit verstanden, alle entscheidungsrelevanten Informationen über den zu beurteilenden Schüler zu berücksichtigen und diese in einer Übergangsempfehlung zu integrieren.

Befunde aus der sozialen Kognitionsforschung belegen, dass sowohl die Entscheidungsverantwortung, also das Ausmaß, in dem der Entscheidungsträger für seine Entscheidung in der Verantwortung steht, als auch die Falltypikalität (Klarheit des Falls) die Wahl der Entscheidungsstrategie beeinflusst. Bei geringer Entscheidungsverantwortung und hoher Falltypikalität ist die Verarbeitung heuristisch, bei hoher Verantwortung oder niedriger Typikalität des zu beurteilenden Falls wird die informationsintegrierende Verarbeitung begünstigt (vgl. Fiske/Neuberg, 1990; Tetlock 1992). Da sowohl Falltypikalität als auch Entscheidungsverantwortung bei der Erstellung der Übergangsempfehlung eine entscheidende Rolle spielen, berücksichtigen wir in unserem Projekt ebenfalls deren Einfluss auf den Entscheidungsprozess der Lehrkräfte. Die zentrale Hypothese des Projektes ist, dass die informationsintegrierende Verarbeitung eher dann auftritt, wenn die Informationen über die Schüler widersprüchlich sind (geringe Falltypikalität, z.B. divergierende Noten) oder die Lehrperson in der Verantwortung für ihre Entscheidung steht, d.h. eine bindende Übergangsentscheidung trifft.

3. Fragestellungen und methodisches Vorgehen des Projektes

Folgende zentrale Fragestellungen werden im Projekt untersucht:

(a) Welchen Einfluss üben die untersuchten Faktoren Entscheidungsverantwortung und Falltypikalität auf die Informationssuche und den Entscheidungsprozess der Lehrkräfte aus (Frage nach Moderatoren der Informationsverarbeitung)?
(b) Sind die Entscheidungsprozesse, die sich bei der Bearbeitung von experimentell präsentierten Fällen finden lassen, auch im Kontext realer Übergangsempfehlungen aufzeigbar (Frage nach der ökologischen Validität von Fällen)?
(c) Führt die informationsintegrierende Verarbeitung tatsächlich zu einer diagnostisch kompetenten Entscheidung (Frage nach der prognostischen Validität)?

Zur Beantwortung der Fragestellung (a) werden im Projekt drei experimentelle Untersuchungen durchgeführt. Die Untersuchungsteilnehmer sind erfahrene Grundschullehrkräfte aus Nordrhein-Westfalen und Rheinland-Pfalz. Das erste Experiment, das derzeit durchgeführt wird, untersucht speziell die Informationssuche während des Entscheidungsprozesses in Abhängigkeit der Entscheidungsverantwortung und Typikalität.

Hierzu wird auf das *Mouselab-Design* (vgl. Johnson u.a. 1986) zurückgegriffen. In diesem Design werden den Grundschullehrkräften auf einem Computerbildschirm Schülerinformationen in Form von aufdeckbaren Feldern präsentiert. Per Mausklick können die Felder geöffnet und die einzelnen Informationen gelesen werden. Abschließend wird die Lehrkraft gebeten, eine Übergangsempfehlung zu erteilen. Die Verantwortung der Entscheidung wird hierbei, wie auch in den anderen Experimenten, durch Instruktion, ein empfehlendes oder verbindliches Übergangsurteil abzugeben, variiert. Im zweiten Experiment untersuchen wir die Integration der Schülerinformationen zu einer Entscheidung. Die Aufgabe der Lehrkräfte besteht darin, auf Basis von schriftlichen Schülerbeschreibungen (Fallvignetten) mit hoher bzw. niedriger Falltypikalität eine Übergangsempfehlung zu treffen. Der Entscheidungsprozess wird regressionsanalytisch modelliert, um die Einflüsse der einzelnen Informationen aus den Schülerbeschreibungen auf die Lehrerentscheidung abzubilden (vgl. Dawes/Corrigan 1974). Die dritte experimentelle Untersuchung repliziert die vorangegangene Untersuchung in Kombination mit der Methode des *lauten Denkens*. Die Lehrkräfte sind bei dieser Methode aufgefordert, ihre Gedanken während der Verarbeitung der Schülerbeschreibung zu verbalisieren.

Die Befunde dieser drei Experimente ermöglichen es, die Informationssuche und den Entscheidungsprozess unter unterschiedlicher Entscheidungsverantwortung und Falltypikalität zu modellieren und die Einflüsse dieser Faktoren zu analysieren.

Um die Fragestellungen (b) und (c) zu verfolgen, wird derzeit eine Validierungsstudie durchgeführt. In der ersten Erhebung der Validierungsstudie schätzen Klassenlehrer aus Nordrhein-Westfalen und Rheinland-Pfalz die Schüler ihrer eigenen vierten Klasse in einer Reihe von übergangsrelevanten Eigenschaften anonymisiert ein. Weiterhin geben die Lehrkräfte an, auf welche Schulform sie die Schüler empfohlen haben, wie leicht ihnen diese Entscheidung fiel und inwieweit sie sich sicher sind, dass die empfohlene Schulform für den jeweiligen Schüler angemessen ist. Neben den Einschätzungen ihrer eigenen Schüler werden den Lehrkräften in einer zweiten Erhebung die in den experimentellen Untersuchungen verwendeten Schülerbeschreibungen präsentiert. Anschließend werden die Lehrkräfte gebeten, für diese Schüler Übergangsempfehlungen zu erteilen. Regressionsanalytisch wird für die Experimentaldaten und die realen Übergangsempfehlungen modelliert, welche Eigenschaften jeweils die Empfehlungen der Lehrkraft beeinflussen. Anschließend wird verglichen, ob sich die derart extrahierten Eigenschaften aus beiden Datenanalysen decken. Um die prognostische Validität der Übergangsempfehlungen zu untersuchen, werden die Eltern der beurteilten realen Schüler im Frühjahr 2010 danach befragt, auf welcher Schulform sich die Schüler ein halbes Jahr nach dem Schulübergang befinden und wie sie sich dort leistungsbezogen, sozial und motivational entwickeln.

Zur Vorbereitung der Hauptuntersuchungen werden Vorstudien durchgeführt. Ziel der Vorstudien ist es, diejenigen Eigenschaften von Schülern zu identifizieren, die von den Grundschullehrkräften als relevant für die Erstellung der Übergangsempfehlung erachtet werden. Diese relevanten Eigenschaften dienen als Grundlage zur Konstruktion der Untersuchungsmaterialien der Hauptuntersuchungen (u.a. die Schülerbeschreibungen). Abbildung 1 gibt eine Übersicht über die einzelnen Schritte und Untersuchungen des Projektes.

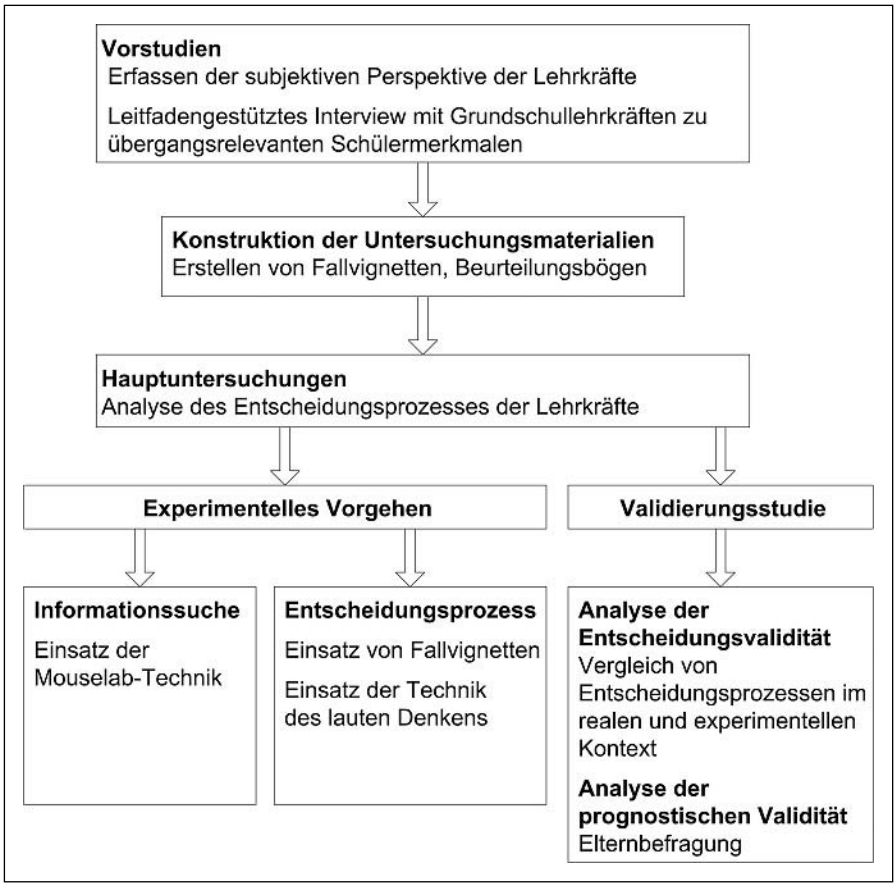


Abb.1: Übersicht über das Projekt

4. Erste Ergebnisse

Im Folgenden werden die Ergebnisse der Voruntersuchungen und die zur Erstellung der Untersuchungsmaterialien durchgeführten Schritte näher dargestellt.

4.1 Die subjektive Perspektive der Lehrkräfte

Zur Identifizierung der entscheidungsrelevanten Merkmale nahmen in der ersten Vorstudie 28 Grundschullehrkräfte aus Nordrhein-Westfalen und 24 Lehrkräfte aus Rheinland-Pfalz an einem halbstandardisierten Interview teil (vgl. Nölle u.a. 2009). Die Lehrkräfte verfügten bei einem mittleren Alter von 46.0 Jahren (SD = 11.8) über eine mittlere Diensterfahrung von 20.3 Jahren (SD = 12.3). Zur Anwerbung der Lehrkräfte wurden Schulen in nordrheinwestfälischen und rheinland-pfälzischen Städten und Kreisen zu-

fällig anhand von Schullisten ausgewählt und sowohl schriftlich als auch telefonisch kontaktiert. Die berücksichtigten Städte und Kreise umfassten in beiden Bundesländern sowohl städtisch als auch ländlich geprägte Gebiete. Nach Einwilligung der Schulleitung wurden die einzelnen Lehrkräfte der Schulen um Teilnahme gebeten. Die ca. 30-minütige Befragungen fanden in den Räumlichkeiten der Schulen statt.

Die Lehrkräfte wurden gebeten, aus den Bereichen der Schulleistung, des sozialen Umfeldes sowie des Lern- und Arbeitsverhaltens diejenigen Merkmale zu nennen, welche von ihnen als übergangsrelevant erachtet werden. Anschließend wurde ihnen die Möglichkeit geboten, weitere relevante Eigenschaften zu nennen.

Aus den Angaben der Lehrkräfte wurden jene Aussagen extrahiert, die übergangsrelevante Eigenschaften thematisieren. Diese Aussagen wurden von zwei unabhängigen Auswertern kategorisiert. Für die Kategorisierung ergab sich eine als befriedigend anzusehende Beurteilerübereinstimmung von $\kappa = .82$. Anschließend wurden die Nennungshäufigkeiten der einzelnen Eigenschaften bestimmt. Tabelle 1 führt diejenigen Eigenschaften auf, welche von mindestens 40% der teilnehmenden Lehrkräfte als relevant genannt wurden (vgl. ebd.).

absolute Häufigkeit	**genannte Eigenschaft**
43	Noten in den Hauptfächern
42	Entwicklung der Leistung über die Zeit, Intelligenz, Selbstständigkeit
34	verfügbare Unterstützung der Eltern
29	Ängstlichkeit, Ausdauer/Fleiß
27	Beteiligung im Unterricht
25	Noten allgemein
24	Kooperation
23	Vergleichsarbeiten, Interesse/Motivation
22	Leistungsbereitschaft, Kopfnoten

Anmerkung: Aufgeführt sind nur Eigenschaften, die mindestens 40% der Teilnehmer nannten.

Tab. 1: Nennungshäufigkeiten der als relevant erachteten Eigenschaften (N = 52) (vgl. Nölle u.a. 2009)

Die Analyse der Häufigkeiten, mit denen die Eigenschaften genannt wurden, ergab, dass die Lehrkräfte vor allem die Leistung der Schüler (z.B. Hauptfachnoten) und ihr Arbeits- und Sozialverhalten (z.B. Selbstständigkeit, Kooperation) als relevant erachten. Weiterhin berücksichtigten sie die längerfristige Leistungsentwicklung der Schüler. Aus dem sozialen Umfeld schätzten die Lehrkräfte nur die von den Eltern verfügbare Unterstützung als übergangsrelevante Information ein. Anhand eines Vergleichs von Lehrpersonen aus zwei Bundesländern mit unterschiedlich verbindlichen Übergangsregelungen

wurde zusätzlich untersucht, ob Lehrkräfte in Nordrhein-Westfalen (bindender Charakter der Empfehlung) andere Übergangskriterien berücksichtigen als Lehrkräfte aus Rheinland-Pfalz (empfehlender Charakter). Die Ergebnisse zeigen, dass die als relevant eingeschätzten Merkmale sich bei den Lehrkräften aus beiden Bundesländern nur geringfügig unterscheiden; die wenigen Unterschiede sind aber konsistent mit den Vorgaben, die in den jeweiligen Bundesländern bestehen (vgl. ebd.). So berücksichtigen die Lehrkräfte in Nordrhein-Westfalen zusätzlich eher das Sozialverhalten der Schüler bei der Übergangsentscheidung, wie es auch die Richtlinien und Lehrpläne für die Grundschulen des Landes vorgeben (vgl. Ministerium für Schule und Weiterbildung Nordrhein-Westfalen 2008a, b).

4.2 Konstruktion der Schülerbeschreibungen

Ausgehend von den Ergebnissen der ersten Voruntersuchung und den Befunden bestehender Übergangsstudien wurde für die zweite Vorstudie ein Set von übergangsrelevanten Schülereigenschaften zusammengestellt und durch Aussagesätze in verschiedenen Ausprägungen beschrieben (z.B. sehr hohe bis sehr geringe Selbstständigkeit). Diese Sätze wurden in einer zweiten Voruntersuchung einer neu akquirierten Stichprobe von 18 Grundschullehrkräften mit der Bitte vorgelegt, diese Aussagen hinsichtlich ihrer Relevanz für die Übergangsempfehlung zu bewerten. Zur Anwerbung wurden hier ebenfalls auf Basis von Schullisten zufällig ausgewählte Schulen telefonisch kontaktiert. Nach Einwilligung der Schulleitung erhielten diese Schulen die entsprechenden Fragebögen auf dem Postweg. Die Einschätzung erfolgte auf einer Skala von 1 (keine Relevanz) bis 5 (sehr hohe Relevanz). Die Analyse der Bewertungen ergab eine durchgehend hohe bis sehr hohe Relevanz der Aussagen, die Leistung und Arbeitsverhalten betrafen. Die mittleren Bewertungen variierten zwischen 3.3 für die Leistung in Vergleichsarbeiten und 4.5 für Ausdauer und Fleiß des Schülers. Von den Beschreibungen des sozialen Umfeldes wurde nur die verfügbare Unterstützung durch die Eltern als relevant eingestuft.

Insgesamt replizierten die Einschätzungen der Lehrkräfte die in der ersten Vorstudie aufgezeigte Perspektive der Lehrkräfte. Um diese Übereinstimmung empirisch quantifizieren zu können, wurde für jede Lehrkraft die punkt-biseriale Korrelation zwischen ihren Bewertungen und der Relevanz bzw. Irrelevanz der beschriebenen Eigenschaft laut der in der ersten Vorstudie aufgezeigten Perspektive der Lehrkräfte berechnet. Die mittlere Korrelation entsprach nach Fishers-Z-Transformation einem Wert von $\bar{r} = .80$.

Die Beschreibungssätze wurden anschließend zu Schülerbeschreibungen zusammengeführt, in denen jede Eigenschaft durch einen der erstellten Aussagesätze charakterisiert wurde. Des Weiteren wurden die Beschreibungen um Aussagen ergänzt, welche übergangsirrelevante Eigenschaften thematisieren (z.B. Freizeitverhalten des Schülers), um ein möglichst umfassendes Bild des Schülers zu vermitteln. Die Ausprägungen der charakterisierten Eigenschaften wurden so variiert, dass sich die Beschreibungen sowohl in ihrem durchschnittlichen Leistungsniveau als auch in der Typikalität des Falles unterscheiden.

5. Diskussion

Insgesamt ergibt sich in den Vorstudien eine Übereinstimmung zwischen der subjektiven Perspektive der Lehrkräfte, bestehenden Übergangsstudien und den gesetzlichen Vorgaben zur Erstellung der Übergangsempfehlung. So soll gemäß des Schulgesetzes des Landes Nordrhein-Westfalen (vgl. Ministerium für Schule und Weiterbildung Nordrhein-Westfalen 2008b) „… auf der Grundlage des Leistungsstands, der Lernentwicklung und der Fähigkeiten der Schülerin oder des Schülers eine zu begründende Empfehlung" (S. 4) getroffen werden. Laut Schulordnung für die öffentlichen Grundschulen des Landes Rheinland-Pfalz stellen die Leistung sowie das allgemeine Arbeitsverhalten des Schülers die Kriterien der Übergangsempfehlung dar, in welcher ebenfalls die Entwicklung des Schülers zu berücksichtigen ist (vgl. Ministerium für Bildung, Wissenschaft, Jugend & Kultur Rheinland-Pfalz 2008). Allerdings werden von den Lehrkräften auch Merkmale genannt, die im Hinblick auf die soziale Selektivität eher problematisch sind, z.B. die Unterstützung im Elternhaus. Hier stellt sich die Frage, ob diese Unterstützung – zumal für die zukünftige Entwicklung – von den Lehrpersonen wirklich eingeschätzt werden kann bzw. inwieweit bei der Einschätzung der häuslichen Unterstützung soziale Stereotype zum Tragen kommen. Auch die starke Berücksichtigung von Verhaltensmerkmalen (Ängstlichkeit, Ausdauer/Fleiß, Beteiligung im Unterricht) lässt die Frage aufkommen, inwieweit soziale Stereotype bei der Entscheidung zum Tragen kommen.

Gegen unser bisheriges methodisches Vorgehen kann eingewendet werden, dass die Ergebnisse eine Tendenz zur sozialen Erwünschtheit haben. Inwieweit dies der Fall ist, können wir in den nachfolgenden Untersuchungen genauer analysieren. Zum einen erlaubt die Verwendung der *Mouselab-Technik* zu sehen, welche Informationen die Lehrpersonen für die Entscheidung tatsächlich heranziehen. Zum anderen kann die Modellierung des Entscheidungsprozesses bei den Fällen und bei den eigenen Schülern Aufschluss darüber geben, inwieweit die bisher gefundenen Antwortmuster durch soziale Erwünschtheit verzerrt wurden.

6. Erkenntnisgewinn und Ausblick

Bestehende Übergangsstudien konzentrierten sich vor allem auf die Frage, welche Schülereigenschaften die Übergangsempfehlung erklären. Diese lassen jedoch die Frage nach dem Entscheidungsprozess bei den Lehrkräften offen, der der Übergangsempfehlung zugrunde liegt. Aus den Erkenntnissen dieses Projektes, unter welchen Rahmenbedingungen verschiedene Entscheidungsstrategien gewählt werden, lassen sich Implikationen darüber ableiten, wie im schulischen Alltag diagnostisch kompetente Entscheidungsprozesse begünstigt werden können.

Die Möglichkeit, aus den Erkenntnissen dieses Projektes Interventionsmaßnahmen zur Verbesserung des Entscheidungsprozesses abzuleiten, ist jedoch von der zentralen Frage nach der *ökologischen Validität* der gewonnenen Erkenntnisse abhängig. Können die im experimentellen Rahmen gewonnenen Erkenntnisse des Projektes auf den Schul-

alltag übertragen werden? Obwohl die durchgeführte Validierungsstudie des ersten Projektzeitraums zu dieser Frage erste Indizien liefert, stellt sich die Frage, ob die reduzierten und abstrakten Informationen künstlicher Fälle wirklich zu ähnlichen Entscheidungen führen wie reale und den Lehrpersonen bereits seit längerem bekannte Kinder in der eigenen Klasse. Diese Frage ist nicht zuletzt deswegen von großer Bedeutung, da zahlreiche Studien zur Erfassung von (diagnostischer) Kompetenz Fallvignetten unter der Annahme verwenden, damit reale Entscheidungssituationen abbilden zu können. Zudem ist die Frage der ökologischen Validität für die Entwicklung von Trainings- und Interventionsmaßnahmen zentral. Daher steht die ökologische Validität in der zweiten Phase des Projektes im Vordergrund.

Literatur

Becker, R. (2004): Soziale Ungleichheit von Bildungschancen. In: Becker, R./Lauterbach, W. (Hrsg.): Bildung als Privileg? Erklärungen und Befunde zu den Ursachen der Bildungsungleichheit. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 161–195.

Bos, W./Voss, A./Lankes, E.-M./Schwippert, K./Thiel, O. (2004): Schullaufbahnempfehlungen von Lehrkräften für Kinder am Ende der vierten Jahrgangsstufe. In: Bos,W./Lankes, E.-M./Prenzel, M./Schwippert, K./Valtin, R./Walther, G. (Hrsg.): IGLU. Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich. Münster: Waxmann, S. 191–228.

Chen, S./Chaiken, S. (1999): The heuristic-systematic model in its broader context. In: Chaiken, S./Trope, Y. (Hrsg.): Dual process theories in social psychology. New York: Guilford, S. 73–96.

Dawes, R.M./Corrigan, B. (1974): Linear models in decision making. In: Psychological Bulletin 81, S. 95–106.

Ditton, H./Krüsken, J. (2006): Der Übergang von der Grundschule in die Sekundarstufe I. In: Zeitschrift für Erziehungswissenschaft 9, S. 348–371.

Fiske, S.T./Neuberg, S.L. (1990): A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In: Zanna, M.P. (Hrsg.): Advances in experimental social psychology Vol. 23. New York, NY: Academic Press, S. 1–74.

Johnson, E.J./Payne, J.W./Schkade, D.A./Bettman, J.R. (1986): Monitoring information processing and decisions: The mouselab system. Unveröffentlichtes Manuskript. Center for Decision Studies, Fuqua School of Business, Duke University.

Krolak-Schwerdt, S./Rummer, R. (2005): Der Einfluss von Expertise auf den Prozess der schulischen Leistungsbeurteilung. In: Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie 33, S. 205–213.

Maaz, K./Hausen, C./McElvany, N./Baumert, J. (2006): Theoretische Konzepte und ihre Anwendung in der empirischen Forschung beim Übergang in die Sekundarstufe. In: Zeitschrift für Erziehungswissenschaft 9, S. 299–328.

Ministerium für Schule und Weiterbildung Nordrhein-Westfalen (2008a): Richtlinien und Lehrpläne für die Grundschule in Nordrhein-Westfalen. http://www.ritterbach.de/lp_online/2012%20Inhalt.pdf [03.08.2009].

Ministerium für Schule und Weiterbildung Nordrhein-Westfalen (2008b): Schulgesetz des Landes Nordrheinwestfalen (Schulgesetz NRW – SchulG). http://www.schulministerium.nrw.de/BP/Schulrecht/Gesetze/SchulG_Info/Schulgesetz.pdf [23.05.2009].

Ministerium für Bildung, Wissenschaft, Jugend und Kultur Rheinland-Pfalz (2008): Schulordnung für die öffentlichen Grundschulen. http://grundschule.bildung-rp.de/fileadmin/user_upload/grundschule.bildung-rp.de/Downloads/Amtliches/Neue_Grundschulordnung_08/GSO-Text.pdf [23.05.2009].

Nölle, I./Hörstermann, T./Krolak-Schwerdt, S./Gräsel, C. (2009): Relevante diagnostische Informationen bei der Übergangsempfehlung – Die Perspektive der Lehrkräfte. In: Unterrichtswissenschaft, S. 294–310.

Stubbe, T./Bos, W. (2008): Schullaufbahnempfehlungen von Lehrkräften und Schullaufbahnentscheidungen von Eltern am Ende der vierten Jahrgangsstufe. In: Empirische Pädagogik 22, H. 1, S. 49–63.

Tetlock, P.E. (1992): The impact of accountability on judgment and choice: Toward a social contingency model. In: Zanna, M.P (Hrsg.): Advances in Experimental Social Psychology Vol. 25. San Diego: Academic Press, S. 331–376.

Tversky, A./Kahneman, D. (1974): Judgment under uncertainty: Heuristics and biases. In: Science 185, S. 1124–1131.

van Ophuysen, S. (2006): Vergleich diagnostischer Entscheidungen von Novizen und Experten am Beispiel der Schullaufbahnempfehlung. In: Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie 38, H. 4, S. 154–161.

Anschrift der Autor/innen

Prof. Dr. Cornelia Gräsel, Bergische Universität Wuppertal, Institut für Bildungsforschung in der School of Education, Gaußstr. 20, D-42119 Wuppertal
E-Mail: graesel@uni-wuppertal.de

Dipl.-Psych. Ines Nölle, Bergische Universität Wuppertal, Institut für Bildungsforschung in der School of Education, Gaußstr. 20, D-42119 Wuppertal
E-Mail: ines.noelle@uni-wuppertal.de

Prof. Dr. Sabine Krolak-Schwerdt, University of Luxembourg, Faculty of Humanities, Arts and Educational Sciences, B.P. 2, L-7201 Walferdang
E-Mail: sabine.krolak@uni.lu

Dipl.-Psych. Thomas Hörstermann, University of Luxembourg, Faculty of Humanities, Arts and Educational Sciences, B.P. 2, L-7201 Walferdang
E-Mail: thomas.hoerstermann@uni.lu

Tina Seidel/Geraldine Blomberg/Kathleen Stürmer

„Observer" – Validierung eines videobasierten Instruments zur Erfassung der professionellen Wahrnehmung von Unterricht

Projekt OBSERVE[1]

1. Einführung

Im Projekt *Observe* wird die Struktur und Entwicklung einer professionellen Wahrnehmung von Unterricht bei Lehramtsstudierenden untersucht. Anlass des Projekts sind aktuell diskutierte Anforderungen an die Lehrerbildung, welche sich vor dem Hintergrund der Neuausrichtung (Modularisierung) der universitären Lehrerbildung stellen (Baumert/Kunter 2006). Im Zentrum dieser Veränderungen steht die Frage, welche Kompetenzen Lehrpersonen entwickeln sollten, um typische berufliche Anforderungen professionell zu bewältigen (Koster u.a. 2005). Die Entwicklung einer professionellen Wahrnehmung stellt einen wesentlichen Bestandteil von Lehrerexpertise dar (Sherin 2002). Die professionelle Wahrnehmung beschreibt die Art und Weise, wie Lehrpersonen Ereignisse und Situationen professionstypisch beobachten und interpretieren. Die Phase der Wahrnehmung kann als Anker zur Erfassung professioneller Kompetenzen gesetzt werden, da bereits in dieser Phase ein Teil der professionellen Anforderungsbewältigung erreicht wird (Bromme 1992).

Ein zentrales Ziel des Projekts Observe stellt die Entwicklung eines videobasierten, standardisierten Diagnoseinstruments dar. Im Rahmen der ersten Förderphase des Schwerpunktprogramms sind nach dem ersten Projektjahr die Instrumentenentwicklung und die Prüfung der Validität abgeschlossen. Diese Befunde zeigen, dass es mit dem *Observer* weitgehend gelungen ist, ein inhaltlich valides Instrument für die standardisierte Erfassung professioneller Unterrichtswahrnehmung zu entwickeln.

2. Theoretischer Ansatz und Fragestellungen

2.1 Professionelle Wahrnehmung von Unterricht

Als theoretischer Hintergrund beziehen wir uns auf das Konzept der professionellen Wahrnehmung als Bestandteil von Lehrerexpertise (Goodwin 1994; Sherin 2002).

1 Diese Veröffentlichung wurde ermöglicht durch eine Sachbeihilfe der Deutschen Forschungsgemeinschaft (Kennz.: SE 1297/2-1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen" (SPP 1293).

Professionelle Wahrnehmung beinhaltet wissensbasierte Prozesse der Aufmerksamkeitssteuerung und Informationsverarbeitung (van Es/Sherin 2008). Professionelle Wahrnehmung wird entsprechend in zwei Komponenten unterteilt: (1) noticing – Identifikation relevanter Situationen und Ereignisse im Unterrichtsgeschehen; (2) knowledge-based reasoning – wissensgesteuerte Verarbeitung identifizierter Situationen und Ereignisse.

„Noticing" – Wahrnehmung von Unterrichtskomponenten

„Noticing" bezieht sich auf die wissensgesteuerte Identifikation von Situationen und Ereignissen im Unterricht, die aus einer professionellen Sicht entscheidend für den Erfolg von Unterrichtshandlungen sind. Vor dem Hintergrund der empirischen Unterrichtsforschung lassen sich eine Reihe von Situationen und Ereignissen identifizieren, die Lernprozesse auf Seiten der Schülerinnen und Schüler unterstützen und somit „erfolgreiche" Komponenten im Unterricht kennzeichnen (Seidel u.a. 2006; Seidel/Shavelson 2007). Zu diesen Komponenten zählen unter anderem:

- die Bereitstellung von Struktur, Zielklarheit und Transparenz (Zielorientierung)
- die Begleitung des Lernens durch die Lehrpersonen und das Überwachen von Lernprozessen und -entwicklungen (Lernbegleitung)
- die Bereitstellung eines unterstützenden und positiven Lernklimas (Lernatmosphäre)

Diese Komponenten dienen als wesentliche Grundlage dafür, welche Situationen und Ereignisse von Lehrpersonen bei der Beobachtung von Unterricht identifiziert und herausgestellt werden. Die drei Komponenten sind inhaltlich in den Bereich pädagogisch-psychologischen Wissens (Shulman 1987) einzuordnen, stellen aber auch aus fächerspezifischer Sicht wesentliche Bestandteile von Unterricht dar.

Knowledge-based reasoning – Wissensgesteuerte Verarbeitung von Unterricht

Professionelle Wahrnehmung beinhaltet darüber hinaus Elemente einer systematischen Beobachtung, die den Einbezug theoretischen Wissens voraussetzt (Borko 2004; Sherin 2007; van Es/Sherin 2002). Nach dem gegenwärtigen Stand der Forschung lässt sich die wissensgesteuerte Verarbeitung durch drei qualitativ unterschiedliche Ebenen kennzeichnen (Berliner 1987,1991; Sherin/van Es 2009; van Es 2009):

- Komponenten eines lernwirksamen Unterrichts auf der Basis theoretischen Wissens differenziert *zu beschreiben*,
- Unterrichtssituationen auf der Basis wissenschaftlicher Theorien und Befunde *zu erklären*,
- Wirkungen von Unterrichtssituationen auf weitere Lehr-Lern-Prozesse *vorherzusagen*.

Befunde der Expertiseforschung zeigen, dass Noviz/innen im Lehrberuf Unterrichtssituationen vorwiegend beschreiben. Die Beschreibungen fallen häufig aufgrund feh-

lenden theoretischen Wissens undifferenziert und „naiv“ aus. Außerdem tendieren Noviz/innen dazu, Situationen zu übergeneralisieren (Berliner 1987,1991). Im Verlauf der Berufsbiografie wird theoretisches Wissen über Lehr-Lern-Prozesse zunehmend genutzt, um Situationen und Ereignisse im Unterricht systematisch einzuordnen und zu Prognosen über weitere Verläufe zu gelangen. Dementsprechend sind Expertinnen und Experten im Lehrberuf häufiger in der Lage, auf der Ebene des Erklärens und Vorhersagens zu operieren (Seidel/Prenzel 2007).

2.2 Methoden zur Erfassung professioneller Wahrnehmung

Leider fehlt es bisher an validen, standardisierten Messinstrumenten, die in der Lage sind, die Struktur und die Entwicklung einer solchen professionellen Wahrnehmung systematisch abzubilden. Im Bereich standardisierter Verfahren werden in der Lehrerforschung häufig weiche Instrumente (z.B. Fragebogenverfahren, berufsbiographische Daten) eingesetzt (Frey 2006). Diese Verfahren haben den Nachteil, dass sie auf subjektiven Selbsteinschätzungen der Befragten beruhen und losgelöst vom Kontext des Unterrichtsgeschehens erfolgen. Im Bereich der Forschung zur professionellen Wahrnehmung wird gegenwärtig überwiegend auf qualitative, in den Kontext des Unterrichtens eingebettete, Zugänge zurückgegriffen. Entwicklungen von Wahrnehmungsprozessen werden zum Beispiel in Videoclubs (van Es/Sherin 2008) untersucht, in denen eine Gruppe von Lehrpersonen über einen längeren Zeitraum Unterrichtsvideos gemeinschaftlich beobachtet und interpretiert. Entwicklungen der professionellen Wahrnehmung werden dann auf der Gruppenebene beschrieben. Rückschlüsse auf die Entwicklung der professionellen Wahrnehmung auf individueller Ebene sind so nur schwer möglich.

2.3 Fragestellungen

Ziel des Projekts *Observe* ist es, ein valides und standardisiertes Instrument zur Erfassung professioneller Wahrnehmung zu entwickeln. Als Zielgruppe dienen Studierende des Lehramts, da wir annehmen, dass die professionelle Wahrnehmung bereits zu einem frühen Zeitpunkt der Berufsbiographie ausgebildet werden kann und professionelle Wahrnehmung spätere Handlungskompetenzen im Unterricht vorbereitet. Folgende Fragestellungen werden bearbeitet:

(1) Gelingt es durch die Kombination von videobasierten Unterrichtsausschnitten und Einschätzverfahren im Ratingformat ein standardisiertes und dennoch kontextualisiertes Instrument zu entwickeln (Inhaltsvalidierung)?
(2) Kann die wissensgesteuerte Informationsverarbeitung von Unterricht (Beschreiben, Erklären und Vorhersagen) als Teil professioneller Unterrichtswahrnehmung abgebildet werden (Konstruktvalidierung)?

Nach einem Jahr Projektarbeit kann durch die Ergebnisse der Pilotierungsstudie die erste Fragestellung in Bezug auf die Entwicklung eines inhaltlich validen standardisierten Erhebungsinstrumentes beantwortet werden.

3. Methode: Instrumentenentwicklung und Pilotierungsstudie

3.1 Instrument: Observer

Vor dem Hintergrund der theoretischen Annahmen ergeben sich für die Entwicklung eines Instruments zentrale Anforderungen. Eine erste Anforderung liegt darin, die professionelle Unterrichtswahrnehmung kontextualisiert zu erfassen. Dafür stellt nach dem aktuellen Forschungsstand der Einsatz videografierter Unterrichtsaufzeichnungen, die situiert und authentisch Anforderungsbeispiele abbilden, einen geeigneten Zugang dar (Darling-Hammond 2006; Reusser 2005).

Auswahl der Videoclips
Bei der Entwicklung des standardisierten Instruments sollte die Identifikation von lernwirksamen Unterrichtskomponenten (noticing) über die gezielte Auswahl von Videoclips vordefiniert werden. Das bedeutet, dass solche Unterrichtssequenzen ausgewählt wurden, in denen Komponenten der Zielorientierung, Lernbegleitung und Lernatmosphäre identifiziert werden können. Zur Auswahl der Videoclips erfolgte die Sichtung von Unterrichtsaufzeichnungen aus deutschsprachigen Ländern (z.B. Reusser 2005–2009), die Inhalte aus unterschiedlichen Fächern (Mathematik/Naturwissenschaft; Gesellschaftswissenschaft/Sprache) und verschiedene Anforderungssituationen (Erarbeiten, Üben) abbilden. Über den gesamten Sichtungsprozess wurden aus ursprünglich 86 Unterrichtsaufzeichnungen zwölf Clips ausgewählt: Physik (2), Mathematik (4), Geschichte (4), Französisch (1) und Englisch (1). In der Sichtung des Videomaterials wurde deutlich, dass es aufgrund der Komplexität von Unterricht sehr schwierig ist, Sequenzen zu finden, die eindeutig nur eine Unterrichtskomponente abbilden. Aufgrund dieser Schwierigkeit wurde ein Vorgehen gewählt, bei dem ein Clip für jeweils zwei der drei Komponenten steht (in unterschiedlichen Kombinationen von Zielorientierung, Lernbegleitung und Lernatmosphäre). Die Unterrichtsaufzeichnungen stammen alle aus den Jahrgangsstufen der 8./9. Klasse der Sekundarstufe.

Entwicklung standardisierter Ratingformate zu den Videoclips
Eine zweite Anforderung liegt in der Entwicklung standardisierter Ratingformate zu den Videoclips, durch welche die Dimensionen der wissensgesteuerten Informationsverarbeitung – knowledge-based reasoning – abgebildet werden. Die Items repräsentieren für jeden Videoclip die drei theoretisch angenommenen Dimensionen (Beschreiben, Erklären, Vorhersagen) durch je drei standardisierte Rating-Items. Die Items wurden aus bestehenden Beobachtungssystemen aus nationalen Videostudien in den Fächern Mathematik, Englisch und Physik übernommen, adaptiert und zum Teil neu entwickelt. Auf

der Dimension des Beschreibens zielen die Items auf die differenzierte Beobachtung zentraler Merkmale der entsprechenden Unterrichtskomponente (z.B. im Bereich der Zielorientierung an der Beobachtung, ob Lehr-Lern-Ziele explizit thematisiert werden). Auf der Dimension des Erklärens orientieren sich die Items an der Beziehung zwischen beobachteten Merkmalen und möglichen Wahrnehmungen auf Seiten der Schülerinnen und Schüler (z.B. ob sich Schülerinnen und Schüler auf der Basis geklärter Lehr-Lern-Ziele in ihrer Kompetenz unterstützt fühlen können). Auf der Dimension des Vorhersagens sollen Auswirkungen (z.B. auf die Lernmotivation, die kognitive Aktivität oder die emotionale Befindlichkeit) eingeschätzt werden. Grundlage für die Bezugnahme von beobachteten Merkmalen mit Erklärungen und Auswirkungen bildet das Angebot-Nutzung-Modell der Unterrichtsforschung mit einem Schwerpunkt auf der Integration der Selbstbestimmungstheorie der Motivation (Seidel 2003; Seidel u.a. 2006).

Integration in computerbasiertes Instrument

Abschließend wurde das Instrument *Observer* in ein computer- und onlinebasiertes Anwendertool integriert. Die Teilnehmerinnen und Teilnehmer werden Seite für Seite durch das (selbsterklärende) Instrument geführt (Abbildung 1). Zu Beginn erfolgen eine theoretische Einführung zu den drei Unterrichtskomponenten sowie Hinweise zur praktischen Handhabung des Instruments. Daraufhin werden sechs Clips präsentiert, welche

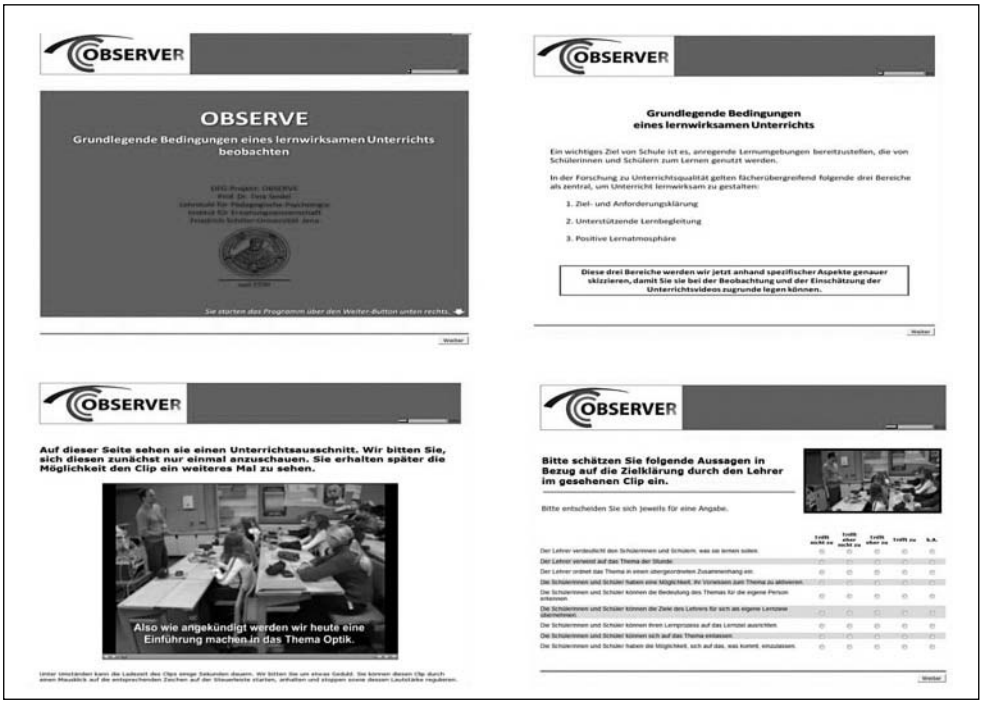


Abb. 1: Bild 1: Startseite, Bild 2: Theoretische Einführung, Bild 3: Präsentation Videoclip, Bild 4: Rating-Items professionelle Wahrnehmung.

jeweils in zugehörige Items eingebettet sind. Zu jedem Clip erhalten die Teilnehmerinnen und Teilnehmer eine Instruktion mit den wichtigsten Hintergrundinformationen zur Unterrichtssequenz. Im Anschluss erhalten sie die Möglichkeit, sich den Clip anzusehen und ihre Perspektiven, ihre Emotionen und mögliche Zuordnungen zu den Unterrichtskomponenten einzuschätzen (Prozessdaten). Anschließend wird der Clip ein zweites Mal präsentiert und anhand von Rating-Items zu den zwei jeweils zugeordneten Unterrichtskomponenten in den Dimensionen Beschreiben, Erklären, Vorhersagen eingeschätzt. Abschließend füllen die Teilnehmerinnen und Teilnehmer einen Evaluationsfragebogen aus. In dem Evaluationsbogen werden die Teilnehmerinnen und Teilnehmer gebeten, allgemeine Aussagen bezüglich des Gesamteindrucks zum Instrument auf einer 4-stufigen Skala einzuschätzen (1 = trifft zu, 4 = trifft nicht zu). Der Evaluationsbogen wurde im Rahmen des vorangegangen Projekts *Lernen aus Unterrichtsvideos* (LUV) entwickelt und skaliert (Seidel/Prenzel 2007). Die gesamte Bearbeitungszeit des Instruments *Observer* beträgt durchschnittlich 90 Minuten, sodass es im Rahmen üblicher Lehrveranstaltungszeiten in der Hochschule eingesetzt werden kann.

3.2 Pilotierungsstudie

Ziel der Pilotierungsstudie ist es, die Gültigkeit des Instruments *Observer* in Bezug auf die standardisierte und kontextualisierte Erfassung von professioneller Unterrichtswahrnehmung zu überprüfen. Im Fokus der Untersuchung steht, ob (a) die Videoclips situierte und authentische Anforderungsbeispiele abbilden, (b) die standardisierten Ratingformate als valide eingestuft werden und (c) welchen Gesamteindruck das Instrument vermittelt.

In der Pilotierungsstudie wurde der *Observer* von 40 Lehramtsstudierenden (davon 24 weiblich und 16 männlich, durchschnittlich im 6./7. Semester, unterschiedliche Fächerkombinationen) und sechs nationalen Expertinnen und Experten der Lehrer- und Unterrichtsforschung bearbeitet.[2] Um Positionseffekte der Clips innerhalb des Instruments zu kontrollieren, wurde den Teilnehmerinnen und Teilnehmern zufällig eine von zwei Parallelversionen des Instruments *Observer* (mit je 6 Clips) zugewiesen. Die Lehramtsstudierenden wurden gebeten, bei der Bearbeitung in Einzelsitzungen „laut zu denken". Zusätzlich wurde die Bearbeitung der Lehramtsstudierenden protokolliert und auf Video aufgezeichnet. Die sechs Expertinnen und Experten der Lehrerbildung bearbeiteten den *Observer* über einen Online-Zugang. Zur Prüfung der inhaltlichen Validität des Instruments werden bezüglich der Auswahl der Videoclips und der standardisierten Ratingformate die Angaben von drei internen Expertinnen und Experten der Projektgruppe, den externen nationalen Expertinnen und Experten sowie die Angaben der Lehramtsstudierenden herangezogen.

2 Wir bedanken uns bei den Lehramtsstudierenden der Friedrich-Schiller-Universität Jena sowie bei den nationalen Expert/innen der Lehrer- und Unterrichtsforschung für ihre Teilnahme an der Pilotierungsstudie.

4. Ergebnisse: Inhaltliche Validität des Instruments *Observer*

4.1 Validität der Videoclips

Die Auswahl der Unterrichtsausschnitte wurde über mehrere Validierungsschritte (extern, intern) realisiert: Erstens wurden formale Kriterien wie die Verständlichkeit, die Aufnahmequalität und die Authentizität der Videos überprüft. Zweitens wurde getestet, ob die ausgewählten Unterrichtssequenzen repräsentativ für die Bereiche Zielorientierung, Lernbegleitung und Lernatmosphäre sind. Dazu gaben die internen Expertinnen und Experten im Projektteam Einschätzungen ab, welche zwei Unterrichtskomponenten ein ausgewählter Videoclip repräsentiert. Die sechs nationalen externen Expertinnen und Experten schätzten (ohne Kenntnis der Vorauswahl) ebenfalls ein, für welche Komponenten der entsprechende Videoclip steht. Den externen Expertinnen und Experten war es dabei freigestellt, für wie viele Bereiche sie einen entsprechenden Clip als besonders repräsentativ betrachten. Als Maß für eine Übereinstimmung wurde – entsprechend der gängigen Verfahrensweise bei Videostudien – berechnet, in wie viel Prozent der Fälle die externen Expertinnen und Experten den Clip als repräsentativ für einen entsprechenden Bereich von Unterricht einschätzten. Im Bereich Zielorientierung stimmen im Durchschnitt 66,3% Prozent, im Bereich der Lernbegleitung 48,8% Prozent und im Bereich der Lernatmosphäre 66,3% Prozent der Expertinnen und Experten zu, dass die Clips den Bereich in besonderem Maße repräsentieren. Für die beiden Bereiche der Zielorientierung und der Lernatmosphäre sind die Übereinstimmungen durchaus zufriedenstellend. Im Bereich der Lernbegleitung wird deutlich, dass es sich hier um einen Merkmalsbereich von Unterricht handelt, der in fast allen Unterrichtssituationen vorkommt, aber nicht unbedingt besonders salient wird. Aus diesem Grund sind die Übereinstimmungen zwischen den Expertinnen und Experten hier geringer. Trotzdem gehen wir davon aus, dass in der Mehrzahl der Fälle bzw. Clips entsprechende Aspekte der Lernbegleitung (z.B. Fragen und Reaktionen der Lehrpersonen) deutlich werden. Insgesamt konnte über den Validierungsprozess sichergestellt werden, dass die für das Instrument ausgewählten Videoclips repräsentative Situationen im Unterricht für die Bereiche der Zielorientierung (mit Schwerpunkt Ziel- und Anforderungsklärung), Lernbegleitung (mit Schwerpunkt Fragen und Feedback der Lehrpersonen) und Lernatmosphäre (mit Schwerpunkt Humor und Ernstnehmen der Schülerinnen und Schüler) darstellen.

Im Rahmen der weiteren Evaluation des Instruments wurde geprüft, ob die Videoclips insgesamt als authentisch und für den Unterricht relevant eingeschätzt werden. Dafür wurden die Angaben der Lehramtsstudierenden und der Expertinnen und Experten aus dem Evaluationsbogen zugrunde gelegt. In allen zu evaluierenden Bereichen bewerteten die beiden Teilnehmergruppen im oberen Drittel der Items. Tabelle 1 veranschaulicht die Einschätzungen der Teilnehmergruppen bezüglich der Auswahl der Clips.

Die Clips fand ich …	**Lehramtsstudierende** ***M* (*SD*)**	**Expert/innen** ***M* (*SD*)**
ergiebig	3.34 (0.62)	3.33 (0.58)
aussagekräftig	3.37 (0.69)	3.25 (0.50)
zu kurz	2.55 (0.99)	2.50 (1.29)
untypisch	1.63 (0.79)	1.75 (0.96)
authentisch	3.45 (0.67)	3.75 (0.50)
interessant	3.76 (0.59)	3.50 (0.58)
abwechslungsreich	3.63 (0.64)	4.00 (0.00)

Anmerkung: Skala ‚1' trifft nicht zu bis ‚4' trifft zu

Tab. 1: Einschätzung der zwei Teilnehmergruppen zur Auswahl der Clips

4.2 Validität der standardisierten Ratingformate

Für die weitere Prüfung des Kompetenzmodells der professionellen Unterrichtswahrnehmung werden Ratingeinschätzungen von Expertinnen und Experten als „Schablone" zugrunde gelegt und die Einschätzungen von Proband/innen mit den Experteneinschätzungen verglichen (vgl. Seidel/Prenzel 2007). Ein zentraler Schritt in der Festlegung der Expertenratings ist die Prüfung der Unabhängigkeit der Einschätzungen. Dazu schätzten drei interne Expertinnen und Experten im Team alle Items zu den zwölf Videoclips einzeln und unabhängig voneinander ein. Die Übereinstimmung als durchschnittliches Cohen's Kappa der einzelnen Expertinnen und Experten beträgt $\kappa = 0.79$. Damit darf die Festlegung der Experteneinschätzungen als objektiv und reliabel eingestuft werden. Für die weiteren Skalierungen werden Differenzwerte zwischen Probanden- und Expertenurteilen gebildet.

Darüber hinaus zeigen die summativen Ergebnisse aus dem Evaluationsbogen eine positive Akzeptanz der standardisierten Items bei beiden Teilnehmergruppen. Ähnlich der Einschätzungen zu den Clips wurden Fragen wie: „Die Beantwortung der Fragen fand ich anregend" größtenteils im oberen Drittel der Skala eingeschätzt.

Die Fragen zu den Clips fand ich …	**Lehramtsstudierende** ***M* (*SD*)**	**Expert/innen** ***M* (*SD*)**
schwierig	1.84 (0.72)	2.25 (1.26)
interessant	3.13 (0.74)	3.50 (0.58)
angemessen	3.24 (0.68)	3.25 (0.50)

Die Fragen zu den Clips fand ich …	**Lehramtsstudierende** ***M* (*SD*)**	**Expert/innen** ***M* (*SD*)**
unpassend	1.66 (0.67)	2.00 (1.16)
zu umfangreich	2.13 (0.94)	1.75 (0.96)
abwechslungsreich	2.35 (0.95)	2.80 (1.30)
Die Beantwortung der Fragen fand ich …		
inhaltlich schwierig	1.95 (0.69)	2.50 (1.00)
anstrengend	1.95 (0.77)	2.00 (0.82)
aufwendig	1.97 (0.89)	2.00 (0.82)
herausfordernd	2.45 (0.86)	2.50 (1.29)
anregend	3.21 (0.74)	3.60 (0.55)

Anmerkung: Skala ‚1' trifft nicht zu bis ‚4' trifft zu

Tab. 2: Einschätzungen der zwei Teilnehmergruppen zu den Items

4.3 Gesamteindruck des Instruments

Insgesamt deuten die Ergebnisse – wie in Tabelle 3 exemplarisch dargestellt – auf eine positive Bewertung des Instruments hin, insbesondere in Bezug auf seine Möglichkeiten für das systematische Beobachten von Unterricht.

Itemtext	**Lehramts-studierende** ***M* (*SD*)**	**Nationale Expert/innen** ***M* (*SD*)**
„Die selbstständige Arbeit mit dem Programm hat geholfen, Details im Unterrichtsgeschehen zu erkennen, die ich sonst nicht wahrgenommen hätte."	3.03 (0.75)	3.25 (0.50)
„Das Programm ist ein geeignetes Mittel, um Unterricht zu analysieren."	3.29 (0.69)	3.50 (0.58)
„Die Videos sind eine geeignete Grundlage, um über Unterricht zu diskutieren."	3.79 (0.47)	3.75 (0.50)
„Die Arbeit mit dem Programm OBSERVER hat meine Aufmerksamkeit für verschiedene Perspektiven im Unterricht erhöht."	3.37 (0.68)	3.40 (0.89)

Anmerkung: Skala ‚1' trifft nicht zu bis ‚4' trifft zu

Tab. 3: Einschätzungen der Teilnehmergruppen zum Gesamteindruck des Instruments

5. Zusammenfassung und Ausblick

Innerhalb eines Projektjahres konnte ein inhaltlich valides Instrument entwickelt werden, welches die bisher eingesetzten Instrumente in der Lehrerforschung und der Forschung zur professionellen Wahrnehmung deutlich erweitert (Frey 2006). Dies zeigt sich in vertretbaren Übereinstimmungen zwischen Expertinnen und Experten in der Auswahl von Videoclips, die bestimmte Unterrichtskomponenten repräsentieren sollen. Die Abweichung der Expertenurteile bei der Zuordnung einzelner Unterrichtsausschnitte zu lernwirksamen Unterrichtskomponenten verdeutlicht aber auch, wie schwierig es ist, die Komplexität von Unterricht in einem standardisierten Instrument abzubilden. Außerdem sind sich Lehramtsstudierende wie nationale Expertinnen und Experten der Lehrer- und Unterrichtsforschung einig, dass die ausgewählten Videoclips Unterricht authentisch abbilden und lernrelevante Situationen und Ereignisse im Unterricht darstellen. Ähnlich positiv werden die standardisierten Ratingformate aufgenommen. In Bezug auf die Festlegung von Experteneinschätzungen zeigen sich hohe Übereinstimmungen zwischen Expertinnen und Experten, sodass die Ausprägungen der in den Videoclips sichtbaren Elemente des Beschreibens, Erklärens und Vorhersagens aus dieser Perspektive zuverlässig eingestuft werden.

Neben der Prüfung der inhaltlichen Validität zeigten die Ergebnisse der Pilotierungsstudie eine besonders positive Aufnahme des Instruments bei den Lehramtsstudierenden. In den Kommentaren der „Laut-Denken-Protokolle" wurde von den Studierenden vielfach herausgestellt, dass sie die Möglichkeit wertschätzen, ihr Wissen anhand von Unterrichtssituationen (-videos) anzuwenden. Weiter schätzen die Studierenden die Struktur, welche ihnen durch das standardisierte Format bereitgestellt wird. Dies deutet wiederum darauf hin, dass Noviz/innen im Lehrberuf sowie Lehramtsstudierende in spezieller Weise von diesem Instrument profitieren könnten.

Der nächste Schritt im Projekt ist nun, das angenommene Kompetenzmodell in den drei Bereichen des Beschreibens, Erklärens und Vorhersagens als Elemente einer wissensgesteuerten Informationsverarbeitung in der Hauptuntersuchung an $N = 150$ Lehramtsstudierenden zu prüfen. Dafür werden die Expertenratings als Bezugsnorm eingesetzt und Maße für den Unterschied zwischen Studierenden- und Expertenmeinung gebildet (vgl. Seidel/Prenzel 2007).

Literatur

Baumert, J./Kunter, M. (2006): Stichwort: Professionelle Kompetenz von Lehrkräften. In: Zeitschrift für Erziehungswissenschaft 9, H. 4, S. 469–520.

Berliner, D.C. (1987): Der Experte im Lehrerberuf: Forschungsstrategien und Ergebnisse. In: Unterrichtswissenschaft 15, S. 295–305.

Berliner, D.C. (1991): Perceptions of student behavior as a function of expertise. In: Journal of Classroom Interaction 26, H. 1, S. 1–8.

Borko, H. (2004): Professional development and teacher learning: mapping the train. In: Educational Researcher 33, H. 8, S. 3–15.

Bromme, R. (1992): Der Lehrer als Experte. Bern: Hans Huber.

Darling-Hammond, L. (2006): Assessing teacher education – The usefulness of multiple measures for assessing program outcomes. In: Journal of Teacher Education 57, H. 2, S. 120–138.

Frey, A. (2006): Methoden und Instrumente zur Diagnose beruflicher Kompetenzen von Lehrkräften – eine erste Standortbestimmung zu bereits publizierten Instrumenten. In: Allemann-Ghionda, C. (Hrsg.): Kompetenzen und Kompetenzentwicklung von Lehrerinnen und Lehrern. 51. Beiheft der Zeitschrift für Pädagogik. Weinheim/Basel: Beltz, S. 30–46.

Goodwin, C. (1994): Professional Vision. In: American Anthropologist 96, H. 3, S. 606–633.

Koster, B./Brekelmans, M./Korthagen, F./Wubbels, T. (2005): Quality requirements for teacher educators. In: Teaching and Teacher Education 21, S. 157–176.

Reusser, K. (2005): Situiertes Lernen mit Unterrichtsvideos. In: Journal für Lehrerinnen- und Lehrerbildung 2, S. 8–18.

Seidel, T. (2003): Lehr-Lernskripts im Unterricht. Münster: Waxmann.

Seidel, T./Prenzel, M./Rimmele, R./Schwindt, K./Kobarg, M./Herweg, C./Dalehefte, I.M. (2006): Unterrichtsmuster und ihre Wirkungen. Eine Videostudie im Physikunterricht. In: Prenzel, M./Allolio-Naecke, L. (Hrsg.): Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms. Münster: Waxmann, S. 100–124.

Seidel, T./Prenzel, M. (2007): Wie Lehrpersonen Unterricht wahrnehmen und einschätzen – Erfassung pädagogisch-psychologischer Kompetenzen bei Lehrpersonen mit Hilfe von Videosequenzen. In: Prenzel, M./Gogolin, I./Krüger, H.-H. (Hrsg.): Kompetenzdiagnostik. Zeitschrift für Erziehungswissenschaft, Sonderheft 8. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 201–218.

Seidel, T./Shavelson, R.J. (2007): Teaching effectiveness research in the past decade: Role of theory and research design in disentangling metaanalysis results. In: Review of Educational Research 77, S. 454–499.

Shulman, L.S. (1987): Knowledge and Teaching: Foundations of the New Reform. In: Harvard Educational Review 57, S. 1–22.

Sherin, M.G. (2002): When teaching becomes learning. In: Cognition and Instruction 20, H. 2, S. 119–150.

Sherin, M.G. (2007): The development of teachers' professional vision in video clubs. In: Goldman, R./Pea, R./Barron B./Derry, S.J. (Hrsg.): Video research in the learning sciences. Mahwah, N.J.: Lawrence Erlbaum, S. 383–395.

Sherin, M.G./van Es, E.A. (2009): Effects of video club participation on teachers' professional vision. In: Journal of Teacher Education 60, S. 20–37.

van Es, E.A. (2009): Participants' roles in the context of a video club. In: Journal of the Learning Sciences 18, H. 1, S. 100–137.

van Es, E.A./Sherin, M.G. (2002): Learning to notice: scaffolding new teachers' interpretations of classroom interactions. In: Journal of Technology and Teacher Education 10, H. 4, S. 571–596.

van Es, E.A./Sherin, M.G. (2008): Mathematics teachers' „learning to notice" in the context of a video club. In: Teaching and Teacher Education 24, S. 244–276.

Anschrift der Autorinnen

Tina Seidel, Friedl Schöller-Stiftungslehrstuhl für Unterrichts- und Hochschulforschung, TUM School of Education, Technische Universität München, Schellingstr. 33, D-80799 München
E-Mail: tina.seidel@tum.de

Geraldine Blomberg, Friedl Schöller-Stiftungslehrstuhl für Unterrichts- und Hochschulforschung, TUM School of Education, Technische Universität München, Schellingstr. 33, D-80799 München
E-Mail: geraldine.blomberg@tum.de

Kathleen Stürmer, Friedl Schöller-Stiftungslehrstuhl für Unterrichts- und Hochschulforschung, TUM School of Education, Technische Universität München, Schellingstr. 33, D-80799 München
E-Mail: kathleen.stuermer@tum.de

Mareike Kunter

Modellierung von Lehrerkompetenzen

Kommentierung der Projektdarstellungen

Das DFG-Schwerpunktprogramm (SPP) „Kompetenzmodelle" hat die Entwicklung und Prüfung theoretischer Kompetenzmodelle, psychometrischer Modelle und Testverfahren zum Ziel. Gemeinsame Grundlage aller Projekte ist ein Verständnis von Kompetenz als einer kognitiven Leistungsdisposition, die die Bewältigung von spezifischen Anforderungssituationen in umschriebenen Domänen ermöglicht. Dabei ist eine Kernannahme, dass diese Dispositionen prinzipiell entwicklungsfähig sind (Klieme/Hartig/Rauch 2008; Klieme/Leutner 2006). Folglich gilt es, handlungsrelevante und veränderbare Wissens- und Könnensmerkmale zu beschreiben und zu erfassen, um somit auch Ansatzpunkte zur Verbesserung von Bildungsprozessen zu gewinnen.

Im SPP liegen drei Projekte vor, die diesen Kompetenzgedanken auf Lehrkräfte übertragen.[1] Damit wird eine veränderte Sicht der schulischen Lehr-Lernprozesse deutlich: Die Erforschung und Verbesserung von Bildungsqualität scheint unvollständig, wenn nicht auch Lehrkräfte als im Bildungswesen aktiv handelnde Personen statt als reine „Input-Variablen" gesehen werden (vgl. Baumert/Kunter 2006; Lipowsky 2006). Denn auch sie müssen komplexe Anforderungen in spezifischen Situationen bewältigen, sei es die Unterrichtssituation selbst oder außerunterrichtliche Situationen wie Elterngespräche. Die Frage, welche Dispositionen – also Kompetenzen – die erfolgreiche Bewältigung dieser professionstypischen Situationen ermöglichen, liegt auf der Hand. Die Verwendung des Kompetenzbegriffs verdeutlicht die Annahme, dass diese Dispositionen nicht angeborene Talente oder Persönlichkeitsstrukturen sind, sondern abgrenzbare Fähigkeiten und Kenntnisse, die ihrerseits im Rahmen geeigneter Lernprozesse – wie der Lehramtsausbildung – vermittelt und vertieft werden können. Die Erforschung der Kompetenzen von Lehrkräften ist somit eine wichtige Aufgabe der Bildungsforschung und kann entscheidende Grundlagen zur Verbesserung von Bildungsprozessen beisteuern.

Die drei Projekte des SPP befassen sich mit unterschiedlichen Anforderungssituationen. Untersucht werden Beratungsgespräche, das Erstellen von Übergangsempfehlungen oder bestimmte Unterrichtssituationen. Ziel ist, die kognitiven Merkmale, die der erfolgreichen Bewältigung dieser Situationen zugrunde liegen, anhand mehrdimensionaler Kompetenzmodelle theoretisch zu beschreiben, empirisch abbildbar zu machen und in den Ausprägungen dieser Kompetenzen interindividuelle Unterschiede zu bestimmen.

1 Die Einschätzung zum Projekt *BITE*, in dem in erster Linie die Bild-Text-Integration und ergänzend auch Lehrerkompetenzen untersucht werden, wird in der Kommentierung der Projekte zu sprachlichen Kompetenzen vorgenommen.

1. Die Projekte im Einzelnen

1.1 Projekt Beratungskompetenz

Beratung ist eine Aufgabe von Lehrkräften, die in verschiedenen Settings stattfindet, etwa in Elterngespräche oder bei Laufbahnberatungen von SchülerInnen. In der Ausbildung für Lehrkräfte werden Strategien zur Führung von Beratungsgesprächen so gut wie nicht vermittelt – ein Großteil der Lehrkräfte meint daher, nicht gut auf Beratungsgespräche vorbereitet zu sein. Beratung ist somit ein wichtiger, aber in Praxis und Forschung bisher vernachlässigter Anforderungsbereich. Es ist sehr zu begrüßen, dass das Projekt die Möglichkeit eröffnet, mehr über die Grundlagen erfolgreicher Beratungen zu erfahren und dadurch Anregungen für eine bessere Vorbereitung der Lehrkräfte auf diese Situationen zu gewinnen.

Die theoretische Grundlage des Projekts ist ein theoretisches Modell mit fünf Merkmalskomplexen, die notwendig erscheinen, um Beratungsgespräche erfolgreich zu führen. Diese Voraussetzungen sind ein angemessener Gesprächsaufbau, hohe Ziel-, Lösungs- und Ressourcenorientierung, angemessene Problemdefinition und Ursachensuche, kooperatives Handeln und die Fähigkeit zum Umgang mit Kritik und schwierigen Beratungssituationen. Ziel der Studie ist, dieses Modell empirisch zu prüfen, Unterschiede in der Beratungskompetenz von Lehrkräften zu beschreiben und diese Unterschiede anhand anderer Lehrermerkmale zu erklären.

Ausgehend vom theoretischen Modell wurde ein Instrumentarium entwickelt, mit dem Lehrkräfte zum einen ihre Kompetenz selbst einschätzen und zum anderen ein Fallbeispiel lösen sollen. Leider ist die Darstellung der Instrumente in der vorliegenden Arbeit sehr knapp und über die Erfassung der Kompetenz speziell mithilfe des Fallszenarios ist wenig bekannt. Inwieweit es sich um einen echten „Test“ handelt, bei dem Können und Wissen anhand des Abgleichs mit normativ richtigen Antworten ermittelt werden, und vor allem, wie sich dieses Können von dem zusätzlich per Test erfassten Beratungswissen abgrenzt, ist nicht erkennbar. Da ein erklärtes Ziel des SPP auch die Entwicklung und Implementierung von Verfahren der Kompetenzmessung ist, wäre es interessant, hier mehr zu erfahren.

Die Befunde belegen, dass das fünfdimensionale theoretische Modell empirisch abbildbar ist und dass mittelhohe Zusammenhänge zwischen selbst eingeschätzter und im Fallbeispiel ermittelter Beratungskompetenz bestehen. Gleichzeitig werden interindividuelle Unterschiede sichtbar, die zum Teil durch professionsbiografische Merkmale erklärbar sind. So zeigen sich unterschiedlich hohe Korrelationen zwischen Kompetenz und der Teilnahme an Fortbildungen zum Thema, was darauf hinweist, dass es sich bei den erfassten Dispositionen tatsächlich um eine „Kompetenz“ handelt, die in geeigneten Lerngelegenheiten vermittelt werden kann. Entsprechend weist der erwartungswidrige Befund eines negativen Zusammenhangs zwischen Berufserfahrung und Beratungskompetenz darauf hin, dass sich diese nicht zwangsläufig durch bloßes Ausüben des Berufs verbessert, sondern dass sich hier bereits bessere Ausbildungsbedingungen niederschlagen könnten. Explizite Fördermaßnahmen scheinen somit geboten, und es wäre

interessant, in späteren Projektphasen die Vermittelbarkeit dieser Kompetenz näher zu erforschen.

1.2 Projekt: Diagnostische Kompetenz bei der Erstellung von Übergangsempfehlungen

Auch die diagnostische Kompetenz von Lehrkräften gilt als zentrale Komponente ihrer Professionalität. Das Projekt setzt an einer Situation an, in der mangelnde diagnostische Kompetenz von Lehrkräften gravierende Folgen für die Biografien von SchülerInnen haben kann, nämlich die Übergangsempfehlungen am Ende der Grundschulzeit, und untersucht Entscheidungsprozesse, die dem Lehrerurteil zugrunde liegen. Die theoretischen Grundlagen sind sozialpsychologische Arbeiten, die zwischen zwei Strategien der Urteils- und Entscheidungsfindung unterscheiden. Da Übergangsentscheidungen komplexe Urteile mit hoher Relevanz darstellen, wird angenommen, dass eine sorgfältige, informationsintegrierende Entscheidung einem heuristisch gefällten Urteil überlegen ist.

Das Programm umfasst mehrere Studien, u.a. explorativ-qualitative Vorstudien und experimentelle Designs, bei denen die Qualität der Entscheidungsprozesse von Lehrkräften anhand neu konstruierter situativ variierender Fallbeispiele untersucht wird. Die Ergebnisse der experimentellen Studien liegen noch nicht vollständig vor. Bisher zeigt sich, dass Lehrkräfte Informationen über SchülerInnen unterschiedlich gewichten und dass sich relevante von weniger relevanten Merkmalen trennen lassen. Ein wichtiges und absolut stimmiges Element ist die geplante Validierungsstudie, in der die im Labor generierten Fallbeispiele mit realen Entscheidungen von Lehrkräften zusammengebracht werden sollen, um die prognostische Validität der entwickelten Verfahren zu prüfen.

Das Projekt überzeugt durch die Anwendung sozialpsychologischer Grundlagenforschung auf eine praktisch hochrelevante Situation des Lehrerhandelns. Das aufeinander aufbauende Forschungsprogramm kombiniert unterschiedliche Designs und stellt damit sicher, dass belastbare Ergebnisse produziert werden, die hohe praktische Relevanz haben können. Eine noch offene theoretische Frage ist hingegen, inwieweit das vorliegende Projekt tatsächlich „Kompetenz“ im Sinne eines individual-differentiellen Konstrukts erfasst, das veränderbares Wissen und Können beschreibt. Die Designs zielen darauf ab, Urteilsprozesse aus allgemein-psychologischer Sicht abzubilden. Interindividuelle Variabilität, möglicherweise auch in Abhängigkeit von Lernerfahrungen, scheint nicht Thema zu sein. Hier wäre es wichtig, in späteren Projektphasen – nach erfolgreicher Etablierung des Erfassungsparadigmas – den Effekt von Fördermaßnahmen zu untersuchen.

1.3 Projekt: Professionelle Wahrnehmung von Unterricht

Das Projekt beschäftigt sich mit dem Kerngeschäft von Lehrkräften, dem Unterricht. Ziel ist die Entwicklung eines Instruments, das die professionelle Wahrnehmung von

Unterrichtssituationen bei (angehenden) Lehrkräften erfasst und testet, ob sie Unterrichtsqualität angemessen beurteilen können. Der theoretische Hintergrund ist die Annahme, dass die angemessene Wahrnehmung pädagogischer Situationen das eigene Handeln bedingt. Dies ist vor allem für die Lehramtsausbildung relevant: Aufgabe der universitären Lehramtsausbildung sollte nicht sein, konkrete Handlungspläne für alle potenziellen Unterrichtssituationen zu liefern, sondern angehenden Lehrkräften einen konzeptuellen Rahmen zu vermitteln, mit dem sie neue, unbekannte Situationen theoretisch einordnen und praktisch bewältigen können.

Die Arbeit stellt die Entwicklung eines Instruments dar, in dem Lehrkräfte unterschiedliche Unterrichtssituationen beschreiben, erklären und hinsichtlich ihres möglichen weiteren Verlaufs beurteilen. Nach Vorstudien wurden Videoszenarien mit Unterrichtssituationen zusammengestellt, die sich – wie anhand mehrerer Expertenratings belegt – in den postulierten Dimensionen der Unterrichtsqualität (Zielorientierung, Lernbegleitung und Lernatmosphäre) unterscheiden. Das Instrument hat Testcharakter, da aufgrund des Abgleichs von Lehrerantworten und Expertenurteilen normativ richtige oder falsche Antworten kodiert werden können. Erste Ergebnisse zeigen, dass das video- und computerbasierte Instrument ökonomisch einsetzbar ist und von Lehramtsstudierenden als praktisch hoch relevant beurteilt wird.

Die noch anstehende Hauptstudie muss nun die Dimensionalität des theoretisch postulierten Kompetenzmodells empirisch prüfen. Inwieweit sich die angenommenen Prozesse der Beschreibung, Erklärung und Vorhersage empirisch trennen und orthogonal zu den drei Bereichen der Unterrichtsqualität abbilden lassen, ist eine offene Frage. Auch das für die Hauptuntersuchung angekündigte Vorhaben, interindividuelle Unterschiede in der Wahrnehmungsqualität zu beschreiben und durch studienbezogene Merkmale zu erklären, verspricht relevante Erkenntnisse. In einer weiteren Projektphase wäre die ökologische Validität des Instruments zu prüfen. Hier ist vor allem an den Zusammenhang zwischen Wahrnehmungsqualität und tatsächlichem Unterrichtshandeln zu denken. Ließe sich mit diesem Instrument nachweisen, dass die angemessene Beurteilung fremder Unterrichtssituationen ein kausaler Faktor für das eigene Unterrichtshandeln ist, wäre es ein interessantes Werkzeug, das zu Evaluationszwecken, aber möglicherweise auch in der Lehreraus- und -weiterbildung einsetzbar ist.

2. Zusammenschau

Die vorliegenden Projekte haben im Kontext des SPP Kompetenzmodelle eine Sonderstellung, da nicht SchülerInnen, sondern Lehrkräfte im Mittelpunkt stehen. Sie untersuchen relevante Situationen des beruflichen Alltags von Lehrkräften und haben auf Basis fundierter theoretischer Überlegungen Kompetenzmodelle entwickelt, die mithilfe neu konstruierter, innovativer Vorgehensweisen empirisch überprüfbar sind.

In der bisherigen Projektphase standen die theoretische Konzeption von Kompetenzmodellen und die Entwicklung von Messverfahren im Vordergrund. Der gemeinsame theoretische Nenner der drei Projekte ist ein Verständnis von Kompetenz als eine kogni-

tive Leistungsdisposition. Die untersuchten Zielkonstrukte werden vielfältig definiert und umfassen kognitive Merkmale wie Wahrnehmungstendenzen (Projekt *Diagnostische Kompetenz*, Projekt *Observe*), deklaratives Wissen (Projekt *Observe*, Projekt *Beratungskompetenz*), Entscheidungs- und Handlungsstrategien (Projekt *Diagnostische Kompetenz*, Projekt *Beratungskompetenz*). Alle Konstruktdefinitionen konzipieren die jeweils untersuchten kognitiven Dispositionen als Voraussetzungen für angemessenes bzw. erfolgreiches Handeln in komplexen Situationen (z.B. „Lösung" des Beratungsfalls oder die „richtige" Übergangsempfehlung). Der Spezifitätsgrad der Situationen reicht von der sehr konkreten, eng umschriebenen Situation der Übergangsempfehlung bis zum Unterricht im Allgemeinen. Für weitere Projektphasen wäre eine genauere Analyse des theoretischen Wirkungsfelds bzw. der Kontextspezifität der jeweiligen Kompetenz interessant, etwa die Frage, ob Lehrkräfte, die kompetente Übergangsempfehlungen geben, auch während des Unterrichts gut diagnostizieren können oder ob Lehramtskandidat/innen die von ihnen studierten Unterrichtsfächer professionell unterschiedlich wahrnehmen.

Ein weiterer wichtiger Aspekt für künftige Arbeit ist die Veränderbarkeit der untersuchten Kompetenzen. Eine Kernannahme des Kompetenzbegriffs ist, dass es sich um erlernbare und entwicklungsfähige Merkmale handelt (vgl. z.B. Sternberg/Grigorenko 2003). Während es unmittelbar plausibel erscheint, dass erlerntes theoretisch-didaktisches Wissen die angemessene Beurteilung einer Unterrichtssequenz erleichtert oder dass bestimmte Gesprächsführungsstrategien trainierbar sind, könnte die Einwirkung auf Urteilsprozesse deutlich schwieriger sein. Zu Recht betonen daher die vorliegenden Arbeiten die Wichtigkeit von Untersuchungen interindividueller Unterschiede etwa im Experten-Novizen-Vergleich, um Anhaltspunkte zu gewinnen, ob die untersuchten Merkmale tatsächlich durch geeignete Lerngelegenheiten formbar sind. Hier wären hochinteressante Befunde und damit substanzielle Beiträge zur praktischen Gestaltung von Lernumgebungen für Lehrkräfte zu erwarten.

Da in der Lehrerforschung die systematische Entwicklung von Kompetenzmodellen und entsprechenden Erfassungsinstrumenten mehr oder weniger Neuland ist, haben sich die hier vorgestellten Arbeiten zu Recht zunächst auf die Entwicklung von Verfahren zum Einsatz in der Grundlagenforschung konzentriert. Die empirische Befundlage im Bereich der Lehrerkompetenzen ist bisher noch äußerst defizitär. Die SPP-Projekte zeigen eindrucksvoll, wie durch theoretisch fundierte und klar abgeleitete Kompetenztheorien und die kreative Entwicklung von Erhebungsmethoden theoretisch wie praktisch bedeutsame Fortschritte erzielt werden können, die das Feld mit Sicherheit voranbringen werden.

Literatur

Baumert, J./Kunter, M. (2006): Stichwort: Professionelle Kompetenz von Lehrkräften. In: Zeitschrift für Erziehungswissenschaft 9, H. 4, S. 469–520.

Klieme, E./Hartig, J./Rauch, D. (2008): The concept of competence in educational contexts. In: Hartig, D./Klieme, E./Leutner, D. (Hrsg.): Assessment of Competencies in Educational Contexts. Göttingen: Hogrefe & Huber, S. 3–22.

Klieme, E./Leutner, D. (2006): Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. In: Zeitschrift für Pädagogik 52, S. 876–903.
Lipowsky, F. (2006): Auf den Lehrer kommt es an. Empirische Evidenzen für Zusammenhänge zwischen Lehrerkompetenzen, Lehrerhandeln und dem Lernen der Schüler. 51. Beiheft der Zeitschrift für Pädagogik, S. 47–70.
Sternberg, R.J./Grigorenko, E.L. (Hrsg.) (2003): The psychology of abilities, competencies, and expertise. New York: Cambridge University Press.

Anschrift der Autorin

Prof. Dr. Mareike Kunter, Johann Wolfgang Goethe-Universität Frankfurt am Main,
Institut für Psychologie, Arbeitsbereich Pädagogische Psychologie, Senckenberganlage 15,
D-60325 Frankfurt a. M.
E-Mail: kunter@paed.psych.uni-frankfurt.de