Hartig, Johannes

Skalierung und Definition von Kompetenzniveaus

Klieme, Eckhard [Hrsg.]; Beck, Bärbel [Hrsg.]: Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International). Weinheim u.a. : Beltz 2007, S. 83-99

urn:nbn:de:0111-opus-31435

in Kooperation mit:

BELTZ

http://www.beltz.de

Kontakt:

peDOCS

Deutsches Institut für Internationale Pädagogische Forschung (DIPF)

Informationszentrum (IZ) Bildung

Schloßstr. 29, D-60486 Frankfurt am Main

eMail: pedocs@dipf.de

Internet: www.pedocs.de

Bärbel Beck / Eckhard Klieme (Hrsg.)

Sprachliche Kompetenzen

Konzepte und Messung

DESI-Studie
(Deutsch Englisch Schülerleistungen International)

Beltz Verlag · Weinheim und Basel

Dr. *Bärbel Beck* ist Diplompsychologin und Projektkoordinatorin am Deutschen Institut für Internationale Pädagogische Forschung (DIPF) in Frankfurt a.M.

Prof. Dr. *Eckhard Klieme* ist Direktor des Deutschen Instituts für Internationale Pädagogische Forschung (DIPF) in Frankfurt a.M.

Diese Studie wurde im Auftrag der Kultusministerkonferenz erstellt. Für die Richtigkeit des Ergebnisses der Studie trägt das »Deutsche Institut für Internationale Pädagogische Forschung« allein die Verantwortung.

Lektorat: Peter E. Kalb

© 2007 Beltz Verlag · Weinheim und Basel
www.beltz.de
Herstellung: Klaus Kaltenberg
Satz: Deutsches Institut für Internationale Pädagogische Forschung
Druck: Druckhaus »Thomas Müntzer«, Bad Langensalza
Printed in Germany

ISBN 978-3-407-25398-9

Inhaltsverzeichnis

Johannes Hartig

Skalierung und Definition von Kompetenzniveaus

Skalierung der Leistungstests in DESI

Die Erfassung der Sprachkompetenzen in Deutsch und Englisch erfolgt in DESI im Rahmen eines differenzierten multidimensionalen Ansatzes. Sowohl für Deutsch als auch für Englisch werden verschiedene Teilbereiche der jeweiligen Sprachkompetenz untersucht, so z.B. das Verstehen englischer Texte in schriftlicher Form (Leseverstehen) oder gesprochener Sprache (Hörverstehen). Jede dieser Teildimensionen wird in DESI durch einen separaten Test repräsentiert. Die verschiedenen DESI-Tests enthalten, entsprechend der Vielfalt der erfassten Kompetenzen, eine Vielfalt unterschiedlicher schriftlicher und auditiver Teststimuli. Die Testaufgaben haben zum größten Teil ein geschlossenes Antwortformat (multiple-choice), teilweise aber auch offene Antwortformate wie z.B. die zu schließenden Lücken des C-Tests (vgl. Schröder/Harsch in diesem Band). Während bei den geschlossenen Antwortformaten von vornherein klar definiert ist, welche Antworten richtig sind, werden die offenen Antworten anhand entsprechender Kodieranweisungen dahingehend eingeschätzt, für welche Testleistungen wie viele Punkte vergeben werden.

Die Skalierung der DESI-Leistungsdaten erfolgt auf Grundlage der Item-Response-Theorie (IRT, s. z.B. van der Linden/Hambleton 1997; Rost 2004), hierbei werden die Antwortwahrscheinlichkeiten der einzelnen Aufgaben als eine Funktion der zugrundeliegenden Fähigkeit betrachtet. Das in DESI verwendete IRT-Modell ist ein generalisiertes Rasch-Modell, welches in der Analysesoftware ConQuest implementiert ist (Wu/Adams/Wilson 1998). Innerhalb dieses Modells können sowohl dichotome (z.B. falsch/richtig) als auch ordinale Auswertungsformate (partial credit-Modell, z.B. falsch/teilweise richtig/vollständig gelöst) innerhalb derselben Tests modelliert werden. Die DESI-Testaufgaben wurden auf Basis einer Voruntersuchung an 986 Schülern nach Rasch-Homogenität innerhalb der einzelnen Tests selektiert. Hierbei wurde teilweise auch die ursprünglich angenommene Dimensionalität der Tests revidiert.

Die Auswertung auf Basis des Rasch-Modells hat eine Reihe untersuchungs- und auswertungstechnischer Vorteile. Einer davon ist die Möglichkeit der Vorgabe von Aufgaben in einem Matrix-Design, d.h. dass jeder Schüler nur eine Teilmenge aller Aufgaben jedes Tests bearbeitet. Schon aus zeitökonomischen Gründen können nicht alle Schüler der DESI-Stichprobe alle Aufgaben der DESI-Tests beantworten. Noch wichtiger wird die Möglichkeit des Matrix-Designs in DESI im Zusammenhang mit der Messwiederholung zwischen Anfang und Ende des Schuljahres. Aufgrund

möglicher Erinnerungseinflüsse können die Leistungen desselben Schülers zu zwei Zeitpunkten nicht mit denselben Aufgaben gemessen werden. Die Verwendung eines Matrix-Designs erlaubt es, jedem Schüler zu beiden Zeitpunkten unterschiedliche Aufgaben desselben Tests vorzulegen, gleichzeitig werden alle Aufgaben zu beiden Zeitpunkten eingesetzt. Bei einer Testwertbildung im Sinne der klassischen Testtheorie (z.B. Moosbrugger/Hartig 2002) – zum Beispiel als Summe der gelösten Aufgaben – könnten Leistungen zwischen Schülern oder Zeitpunkten, denen die Bearbeitung verschiedener Aufgaben zugrunde liegt, nicht verglichen werden. Im Rahmen der Skalierung der Leistungsdaten auf Basis des Rasch-Modells ist es hingegen möglich, für Schüler, die unterschiedliche Aufgaben bearbeitet haben, Schätzwerte auf einer gemeinsamen Skala zu ermitteln. Die Schätzung der Schülerleistungen erfolgt, wie auch in anderen Large Scale Assessments, mit so genannten *Plausible Values*. Diese ermöglichen durch den Einbezug von erklärenden Hintergrundvariablen wie Schulform, Geschlecht oder sozioökonomischer Hintergrund eine messfehlerbereinigte Schätzung der Zusammenhänge zwischen den erklärenden Variablen und den erfassten Schülerleistungen (vgl. Mislevy/Beaton/Kaplan/Sheehan 1992). Für die Schätzung der Messwerte im selben Test zu zwei Zeitpunkten werden die Plausible Values in einem zweidimensionalen Modell geschätzt, in dem die Aufgabenschwierigkeiten zu beiden Zeitpunkten gleich gesetzt und die Leistungen zu beiden Zeitpunkten als zwei separate Variablen modelliert werden. Dieses Vorgehen erlaubt eine bessere Schätzung der Zusammenhänge der Leistungen zwischen den Zeitpunkten und damit auch eine bessere Schätzung des Leistungszuwachses vom ersten zum zweiten Zeitpunkt (vgl. Hartig/Kühnbach im Druck).

Der für die Skalierung der in DESI verwendeten Leistungstests entscheidende Vorteil der Rasch-Skalierung ist die Möglichkeit, Aufgabenschwierigkeiten und Schülerleistungen auf einer gemeinsamen Skala abzubilden (z.B. Wilson 2003). Bei einer klassischen Testwertbildung lässt sich zwischen der Kompetenz einer Schülerin – z.B. wie viel Prozent aller Aufgaben sie gelöst hat – kein Bezug zur Schwierigkeit einer Aufgabe – z.B. von wie vielen Prozent der Schülerinnen und Schüler wurde sie gelöst – herstellen. Die Raschskalierung bildet für beide Größen eine gemeinsame Skala, der Bezug zwischen der Kompetenz der Personen und der Schwierigkeit der Aufgaben wird in Form von Lösungswahrscheinlichkeiten hergestellt. Abbildung 1 veranschaulicht diesen Zusammenhang am Beispiel des Rasch-Modells für dichotome Antwortalternativen (zu Grundlagen und Erweiterungen des Rasch-Modells s. z.B. Fischer/Molenaar 1991). Die Personenfähigkeiten werden mit θ, die Aufgabenschwierigkeiten mit σ bezeichnet. Die Schwierigkeit σ einer Aufgabe ist im Rasch-Modell definiert als der Punkt auf der Kompetenzskala, an dem die

Wahrscheinlichkeit von Personen[1] mit einer Fähigkeit von $\theta = \sigma$ 50% beträgt, die Aufgabe zu lösen.

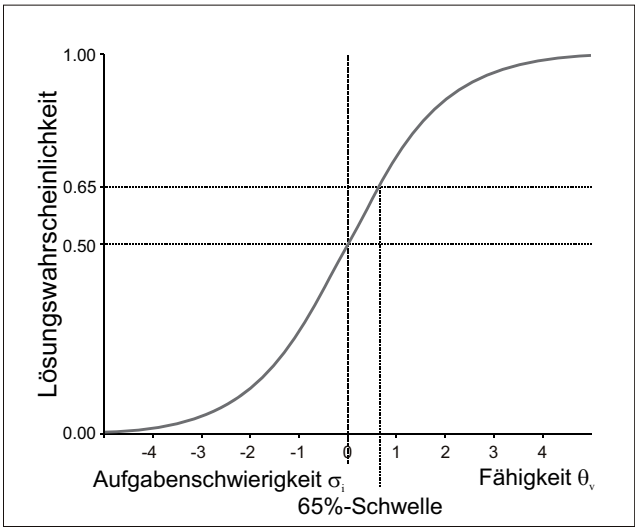

Abbildung 1: Verortung von Aufgabenschwierigkeit σ_i und Personenfähigkeit θ_v auf einer gemeinsamen Kompetenzskala am Beispiel des dichotomen Rasch-Modells.

Anhand des im Rasch-Modell angenommenen Zusammenhanges zwischen Personenfähigkeit und Lösungswahrscheinlichkeit lassen sich auch Punkte auf der Kompetenzskala bestimmen, an denen die Lösungswahrscheinlichkeit für eine spezifische Aufgabe einen beliebigen anderen Wert als 50% annimmt. Bei der Beschreibung von Kompetenzen interessiert, ob eine Population die Anforderungen bestimmter Aufgaben mit einer hinreichenden Sicherheit bewältigen kann – eine Lösungswahrscheinlichkeit von 50% erscheint hierfür relativ niedrig. Daher werden oft höhere Werte gewählt, um einzelne Aufgaben auf der Kompetenzskala zu verorten. In Abbildung 1 ist dies am Beispiel der „65%-Schwelle“, wie sie z.B. auch in der *Third International Mathematics and Science Study* (TIMSS) verwendet wurde (vgl. Klieme/Baumert/Köller/Bos 2000), veranschaulicht. Auch in DESI werden die Skalen der Leistungstests auf die 65%-Schwelle der Testaufgaben bezogen charakterisiert.

1 Es ist ein nahe liegender aber voreiliger Schluss, diese Lösungswahrscheinlichkeiten auf die Ebene der untersuchten Individuen zu übertragen. Das Rasch-Modell nimmt lediglich an, dass eine Teilpopulation von Individuen mit einer Fähigkeit von $\theta = \sigma$ die Aufgabe zu 50% lösen sollte, *nicht*, dass die „Lösungschance“ eines einzelnen Individuums mit einer Fähigkeit von $\theta = \sigma$ für eine einzelne Aufgabe 50% betragen würde (vgl. Borsboom/Mellenbergh/Heerden 2003).

Ziele und mögliche Methoden bei der Definition von Kompetenzniveaus

Aus den DESI-Leistungstests resultieren quantitative Messungen auf kontinuierlichen Skalen. Diese Zahlenwerte sind gut geeignet, um Zusammenhänge der erfassten Kompetenzen mit anderen Variablen zu untersuchen (z.B. dem sozioökonomischen Status der Eltern) oder die Kompetenzen verschiedener Gruppen zu vergleichen (z.B. von Schülern aus verschiedenen Schulformen). Neben derartigen quantitativen Zusammenhangsanalysen ist es aber gerade in DESI von herausragendem Interesse, über welche *spezifischen* Kompetenzen Schüler auf einem bestimmten Niveau verfügen bzw. welche fachbezogenen Leistungsanforderungen sie bewältigen können. Es besteht also der Bedarf an einer *kriteriumsorientierten Interpretation* der quantitativen Leistungswerte. Die numerischen Werte auf der Kompetenzskala sollen zu konkreten, fachbezogenen Kompetenzen in Bezug gesetzt werden. Genau dieses Ziel soll mit der Bildung so genannter *Kompetenzniveaus* erreicht werden. Es wäre in der Praxis nicht realisierbar, für jeden einzelnen Punkt einer quantitativen Skala eine Beschreibung der jeweiligen Kompetenz vorzunehmen (Beaton/Allen 1992). Daher wird eine Unterteilung der Skala in Abschnitte vorgenommen, welche als Kompetenzniveaus bezeichnet werden. Für diese Skalenabschnitte wird dann eine kriteriumsorientierte Beschreibung der Schülerkompetenzen vorgenommen.

Eine in anderen Studien geläufige Bezeichnung für derartige zur Beschreibung kontinuierlicher Skalen gebildeter Abschnitte ist der Begriff *Kompetenzstufe* (z.B. in der PISA-Studie, z.B. OECD 2004; PISA-Konsortium Deutschland 2004). Die Übersetzung des englischen „proficiency level“ mit „Stufe“ ist jedoch nicht ideal (vgl. auch Helmke/Hosenfeld 2004). Stufen werden innerhalb der Erziehungswissenschaften und Psychologie oft mit *qualitativen* Unterschieden – wie z.B. im Piagetschen Modell der kognitiven Entwicklung – assoziiert. Den in Schulleistungsstudien verwendeten „Kompetenzstufen“ liegen in aller Regel jedoch keine echten Stufenmodelle zugrunde, sie dienen lediglich einer einfacheren Kommunikation und Veranschaulichung der erfassten quantitativen Leistungsdimensionen: “Dividing (…) these continua into levels, though useful for communication about students’ development, is essentially arbitrary.” (Adams/Wu 2002, S. 197). Angesichts der möglicherweise irreführenden Konnotationen des Stufenbegriffs wurde für DESI der Begriff *Kompetenzniveau* vorgezogen. Es soll an dieser Stelle jedoch hervorgehoben werden, dass dieser Begriff dasselbe bezeichnet wie *Kompetenzstufen* im Kontext anderer Studien: Abschnitte auf kontinuierlichen Kompetenzskalen, die mit dem Ziel einer kriteriumsorientierten Beschreibung der erfassten Kompetenzen gebildet werden.

Um für die Kompetenzskala eines existierenden Tests Niveaus zu definieren, stehen verschiedene Methoden zur Verfügung. Entscheidend für die Definition der Niveaus ist bei praktisch jeder Vorgehensweise die Definition der *Schwellen zwischen den Niveaus*. Innerhalb eines Skalenabschnittes, der als ein Kompetenzniveau betrachtet wird, wird keine weitere inhaltliche Differenzierung der Schülerkompetenzen

vorgenommen. Abbildung 2 veranschaulicht die Unterteilung einer kontinuierlichen Kompetenzskala in Kompetenzniveaus.

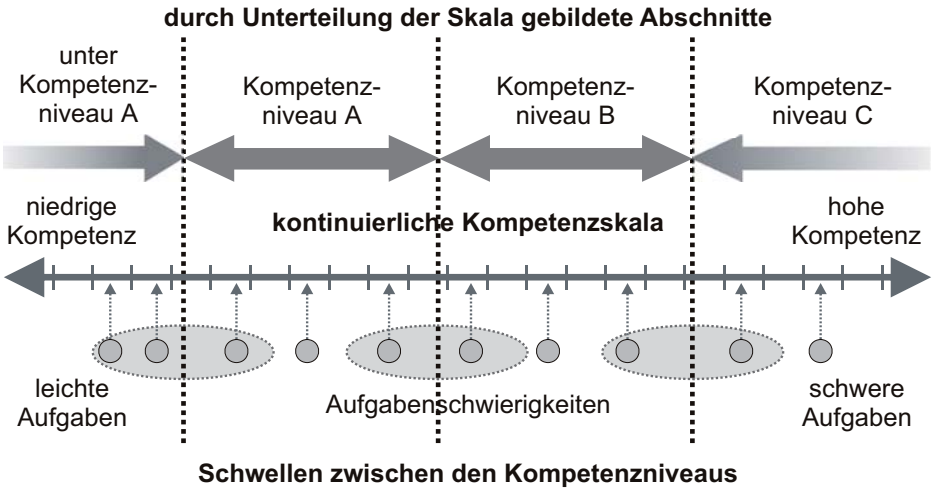

Abbildung 2: Veranschaulichung der Unterteilung einer kontinuierlichen Kompetenzskala mit darauf verorteten Aufgabenschwierigkeiten in Kompetenzniveaus. Die Aufgabenschwierigkeiten in Nachbarschaft der Schwellen zwischen den Niveaus sind unterlegt.

Als inhaltliche Grundlage für die Bildung und die Beschreibung von Skalenabschnitten werden die Schwierigkeiten der Aufgaben eines Tests und deren fachbezogene Anforderungen herangezogen. Bei der Interpretation der Kompetenzniveaus in Bezug auf die Aufgabenschwierigkeiten ist zu beachten, dass zur inhaltlichen Charakterisierung der Niveaus diejenigen Aufgaben herangezogen werden, welche *in Nachbarschaft* der Schwellen am Beginn des jeweiligen Niveaus liegen (vgl. auch Beaton/Allen 1992). In Abbildung 2 sind die Aufgaben in Nachbarschaft der Schwellen zwischen den Niveaus grau unterlegt. Anhand dieser Aufgaben können die Kompetenzen von Schülern beschrieben werden, deren Leistung innerhalb des Niveaus liegt. Es ist hingegen *nicht* angezeigt, die Aufgabenschwierigkeiten *innerhalb* eines Niveaus (d.h. zwischen der unteren und oberen Grenze) zur Beschreibung der Niveaus heranzuziehen – die Aufgaben am „oberen Ende" eines Niveaus sind eher charakteristisch für die Leistungen von Schülern auf dem nächsthöheren Niveau.

Die in verschiedenen vorliegenden Studien eingesetzten Methoden zur Definition und inhaltlichen Beschreibung der Niveaus bzw. der Schwellen zwischen den Niveaus unterscheiden sich vor allem hinsichtlich zweier Merkmale, nämlich

- wie differenziert und mit welchem Abstraktionsgrad die inhaltlichen Anforderungen der Testaufgaben systematisiert werden und
- inwieweit inhaltliche Hypothesen über spezifische Aufgabenanforderungen und über mögliche Kompetenzniveaus schon a priori – vor der Erhebung empirischer Leistungsdaten – formuliert werden.

Die Kompetenzstufenmodelle der in den letzten Jahren durchgeführten internationalen Large Scale Assessments basieren größtenteils auf unterschiedlich systematischen Post-Hoc-Analysen der Aufgabeninhalte und Aufgabenanforderungen (vgl. z.B. Klieme et al. 2000; Artelt/Stanat/Schneider/Schiefele 2001; Bos et al. 2003). Zur Darstellung eines häufig zitierten systematischen Vorgehens siehe Beaton und Allen (1992); eine kurze aktuelle Übersicht über das Vorgehen in jüngeren Studien findet sich bei Helmke und Hosenfeld (2004).

Definition von Kompetenzniveaus in DESI

Formulierung von Aufgabenmerkmalen

In vielen Studien fand eine genauere Betrachtung der Zusammenhänge zwischen Aufgabeninhalten und -schwierigkeiten erst nach der Verankerung auf der Kompetenzskala statt (z.B. Beaton/Allen 1992; Watermann/Klieme 2002). Im Unterschied dazu wurden in DESI für viele der einzelnen Leistungstests bereits vorab Beschreibungen der Testaufgaben hinsichtlich möglicher schwierigkeitsbestimmender Charakteristika vorgenommen. Diese Beschreibungen wurden aus Modellen der jeweils zu erfassenden Kompetenzen abgeleitet und beinhalten Annahmen darüber, welche *spezifischen Anforderungen* zur Schwierigkeit einer Aufgabe beitragen sollten. Aus diesen anforderungsrelevanten *Aufgabenmerkmalen* lässt sich also im vornherein ableiten, welche Aufgaben leichter oder schwerer sein sollten und worauf diese Schwierigkeitsunterschiede zurückzuführen sind. Die Aufgabenmerkmale liegen im einfachsten Fall als globale Einschätzungen ganzer Aufgaben vor, für die meisten Tests jedoch in Form mehrerer differenzierter Merkmale für jede Aufgabe.

Die Aufgabenmerkmale erleichtern es, die jeweilige zu erfassende Kompetenz auf einer verallgemeinerten Ebene, unabhängig von den konkreten Testaufgaben, zu beschreiben. Ein wichtiger Ausgangspunkt für die Identifikation und Beschreibung relevanter Aufgabenmerkmale sind Hypothesen über die Anforderungen, die beim Bearbeiten und Lösen der Aufgaben bewältigt werden müssen. Aufgabenmerkmale können sich auf unterschiedliche theoretisch angenommene Prozesse beim Lösen, aber auch auf technische Oberflächencharakteristika der Aufgaben beziehen. Bereiche, in denen Merkmale kodiert werden können, sind z.B.

- zum Lösen der Aufgabe auszuführende kognitive Operationen (z.B. Suche von Informationen beim Lesen eines Textes);
- die Schwierigkeit hinsichtlich spezifischer inhaltlicher Kriterien (z.B. Wortschatz eines Lesetextes);
- spezifische Phänomene im jeweiligen Leistungsbereich (z.B. bilden von Konjunktiv-Formen in einem Grammatiktest);
- Aufgabenformate (z.B. geschlossene vs. offene Antworten).

Werden Annahmen über die schwierigkeitsbestimmenden Merkmale einer Aufgabe vor der Erhebung empirischer Daten formuliert, gewinnen diese den Status empirisch prüfbarer Hypothesen über Charakteristika der zu erfassenden Kompetenz. Selbstverständlich ist ein Large Scale Assessment mit standardisierten Tests nicht geeignet, kognitive Prozesse beim Bearbeiten der Testaufgaben zu untersuchen. Dennoch kann eine differenzierte Analyse schwierigkeitsbestimmender Aufgabenmerkmale und eines daraus abgeleiteten Messmodells für die zu erfassende Kompetenz mehr über die Validität (Gültigkeit) eines Kompetenztests aussagen als korrelative Zusammenhänge des Testwertes mit weiteren Leistungsmaßen oder sonstigen Variablen (vgl. Borsboom/Mellenbergh/Heerden 2004).

Die meisten Aufgabenmerkmale der DESI-Tests sind derart definiert, dass jede Testaufgabe sinnvoll hinsichtlich jedes Merkmals einzustufen ist. Dies bedeutet, dass auch die einfachste Ausprägung der Merkmale hinsichtlich der Aufgabenanforderungen inhaltlich definiert ist. So enthalten z.B. Aufgaben im Englisch Lese- und Hörverstehen, die im Merkmal *sprachliche Anforderungen* auf der einfachsten Stufe eingeordnet sind, hochfrequente Wörter und einfache grammatische Strukturen (vgl. Nold/Rossa in diesem Band). Die Sprachkompetenz eines Schülers, der eine solche einfache Aufgabe löst, lässt sich also u.a. durch das Beherrschen hochfrequenten englischen Vokabulars charakterisieren. Die höchste Ausprägung des Merkmals *sprachliche Anforderungen* bedeutet, dass Aufgaben einen erweiterten Wortschatz und komplexe grammatische Strukturen beinhalten. Die Kompetenz von Schülern, die derartige Aufgaben lösen, kann entsprechend durch das Beherrschen dieser Anforderungen beschrieben werden. In einigen Fällen sind die Aufgabenmerkmale hingegen stärker qualitativer Natur, z.B. „Dativ/Akkusativ unterscheiden“ in Sprachbewusstheit Deutsch (vgl. Eichler in diesem Band). Die Kompetenz von Schülern, die eine solche Aufgabe lösen, kann durch das Beherrschen des in diesem Aufgabenmerkmal beschriebenen grammatischen Phänomens charakterisiert werden. Es ist in diesem Fall jedoch nicht inhaltlich definiert, was das Nicht-Vorliegen eines solchen Merkmals hinsichtlich der Aufgabenanforderungen bedeutet.

Nach der Benennung und Beschreibung relevanter Aufgabenmerkmale muss durch Fachexperten eine Einschätzung jeder Aufgabe eines Tests hinsichtlich dieser Merkmale vorgenommen werden. Hierbei ist eine qualitative und/oder quantitative Einschätzung, ob bzw. in welcher Ausprägung ein Merkmal bei einer Aufgabe vorliegt, möglich. Diese Einschätzungen sollten im Idealfall gut dokumentiert sein, so dass eine Nachvollziehbarkeit für dritte möglich ist. Für DESI werden die Kriterien zur Einschätzungen der Aufgabenmerkmale auch dahingehend überprüft, inwieweit unabhängige, geschulte Rater hinsichtlich ihrer Einschätzungen übereinstimmen. Eine ausführliche Dokumentation der Kriterien zur Aufgabeneinschätzung ermöglicht neben der Verwendung bei der Definition von Kompetenzniveaus auch die Übertragung auf andere Tests oder noch zukünftig zu konstruierende Aufgaben.

Modellierung der Effekte der Aufgabenmerkmale und erwartete Schwierigkeiten

Kodierung der Aufgabenmerkmale

Zur empirischen Analyse der Zusammenhänge zwischen Aufgabenmerkmalen und -schwierigkeiten werden die Abstufungen der Merkmale in Zahlen ausgedrückt. Hierbei werden die Aufgabenmerkmale i.d.R. so kodiert, dass sie einen positiven Zusammenhang mit der Aufgabenschwierigkeit aufweisen. Ist ein Merkmal zweifach gestuft, wird z.B. die Ausprägung, welche als schwerer angenommen wird, mit 1, die leichtere Ausprägung mit 0 kodiert. Bei Merkmalen mit mehr als zwei Stufen (z.B. „leicht", „mittel", „schwer") werden die einzelnen Stufen in so genannte *Dummy-Variablen* für die einzelnen Merkmalsstufen übersetzt, die wiederum nur Werte von 0 und 1 annehmen. Ein dreistufiges Merkmal wird z.B. in zwei Dummy-Variablen überführt, eine für das Vorliegen der mittleren und eine für das Vorliegen der schwierigsten Merkmalsausprägung. Bei der Untersuchung der Effekte der Dummy-Variablen auf die Aufgabenschwierigkeit wird die angenommene Schwierigkeitsabfolge der Merkmalsstufen einer empirischen Prüfung unterzogen, da erst gezeigt werden muss, dass eine als mittel angenommene Stufe tatsächlich schwieriger ist als die zugehörige als leicht eingestufte Ausprägung desselben Merkmals.

Auswertung mittels linearer Regressionsanalyse

Nach der Definition und Kodierung der schwierigkeitsbestimmenden Merkmale können diese zu den empirisch ermittelten Aufgabenschwierigkeiten in Beziehung gesetzt werden. In DESI werden als Aufgabenschwierigkeiten die oben genannten, aus der Rasch-Skalierung resultierenden 65%-Schwellen verwendet, d.h. diejenigen Werte auf der Kompetenzskala, mit denen Schüler die jeweilige Aufgabe mit einer Wahrscheinlichkeit von 65% lösen sollten. Als Modell für den Zusammenhang zwischen Aufgabenmerkmalen und Aufgabenschwierigkeiten wird für einen großen Teil der DESI-Tests ein additives, lineares Modell gewählt. Die Schwierigkeit jeder Aufgabe ergibt sich in diesem Modell aus der Summe ihrer anforderungsrelevanten Merkmale. Die empirische Analyse dieses Modells für die einzelnen Tests erfolgt mit linearen Regressionsanalysen (z.B. Moosbrugger 2002). Hierbei wird die Aufgabenschwierigkeit als eine gewichtete Summe ihrer einzelnen Merkmale modelliert:

(1) $$\sigma_i = \beta_0 + \beta_1 \cdot q_{i1} + \cdots + \beta_m \cdot q_{im} + \cdots + \beta_M \cdot q_{iM} + \varepsilon_i$$

σ_i = Schwierigkeit von Aufgabe i;
β_m = Regressionsgewicht für Merkmal m;
M = Anzahl der Aufgabenmerkmale;
q_{im} = Kodierung des Merkmals m für Aufgabe i (0=liegt vor, 1=liegt nicht vor);
ε_i = Verbleibende Abweichung zwischen im Modell erwarteter und tatsächlicher Aufgabenschwierigkeit (Residuum).

Die Regressionsgewichte β_m drücken hierbei den Einfluss eines Aufgabenmerkmals auf die Aufgabenschwierigkeit aus. Die Dummy-Kodierungen von mehreren Abstufungen desselben Merkmals werden technisch als separate Merkmale behandelt, für die jeweils eigene Regressionsgewichte geschätzt werden. Die Werte für β_m sind so zu interpretieren, dass eine Aufgabe, bei der das Merkmal m gegeben ist, auf der Kompetenzskala um β_m schwerer ist, als eine Aufgabe, bei der das Merkmal m nicht gegeben ist. Die Werte für die Einflussgewichte β_m werden in der Regressionsanalyse so geschätzt, dass die tatsächlichen Aufgabenschwierigkeiten möglichst gut wiedergegeben werden, d.h. dass die verbleibenden Abweichungen (ε_i in Formel 1) möglichst gering sind.

Prüfung der Vorhersagekraft der Merkmale

Für die Prüfung der Übereinstimmung zwischen tatsächlichen und im Modell erwarteten Aufgabenschwierigkeiten erlaubt die Regressionsanalyse zunächst eine Einschätzung, inwieweit die Unterschiede zwischen Aufgabenschwierigkeiten durch die Aufgabenmerkmale erklärt werden können. Als Maß hierfür wird der so genannte Determinationskoeffizient R^2 herangezogen. Dieser gibt an, welcher Anteil der Unterschiedlichkeit der Aufgabenschwierigkeiten durch die Aufgabenmerkmale erklärt werden kann. Ein Wert von $R^2 = 1.0$ würde bedeuten, dass das Modell eine perfekte Vorhersage der Aufgabenschwierigkeiten erlaubt (alle $\varepsilon_i = 0$), bei einem Wert von $R^2 = 0.0$ wären die Aufgabenmerkmale vollkommen irrelevant für die Schwierigkeiten der Testaufgaben. Ein Wert von $R^2 = 0.5$ bedeutet, dass die Hälfte der Unterschiede zwischen den Aufgabenschwierigkeiten durch die Aufgabenmerkmale erklärt werden kann. Das Ausmaß erklärter Unterschiede ist somit ein Indikator dafür, inwieweit die Annahmen über die schwierigkeitsbestimmenden Aufgabenmerkmale sich durch die tatsächlichen Aufgabenschwierigkeiten stützen lassen. Über die generelle Erklärungskraft der Gesamtheit der Aufgabenmerkmale hinaus kann auf Basis der Regressionsanalyse auch beurteilt werden, welche spezifischen Merkmale mehr oder weniger bedeutend für die Schwierigkeit der Aufgaben sind. Merkmale mit einem hohen Regressionsgewicht β_m sind besonders bedeutsam für die Schwierigkeiten der Aufgaben, Merkmale mit einem Gewicht nahe null haben unter Berücksichtigung der übrigen Merkmale kaum Erklärungskraft hinsichtlich der Aufgabenschwierigkeiten.

Ermittlung erwarteter Aufgabenschwierigkeiten zur Schwellendefinition

Neben der Prüfung der Annahmen zu den Einflüssen von Aufgabenmerkmalen auf Aufgabenschwierigkeiten ermöglicht es die Regressionsanalyse, die *erwarteten Schwierigkeiten* von Aufgaben mit spezifischen Merkmalskombinationen zu bestimmen. Die erwartete Schwierigkeit $\hat{\sigma}_i$ einer Aufgabe i lässt sich als Summe der Regressionsgewichte ermitteln, die für die jeweilige Aufgabe relevant sind. Mit anderen Worten ergibt sich die erwartete Aufgabenschwierigkeit aus Formel 1 ohne das Regressionsresiduum ε_i:

(2) $$\hat{\sigma}_i = \beta_0 + \beta_1 \cdot q_{i1} + \cdots + \beta_m \cdot q_{im} + \cdots + \beta_M \cdot q_{iM}$$

$\hat{\sigma}_i$ = erwartete Schwierigkeit von Aufgabe i;
β_m = Regressionsgewicht für Merkmal m;
M = Anzahl der Aufgabenmerkmale;
q_{im} = Kodierung des Merkmals m für Aufgabe i (0=liegt vor, 1=liegt nicht vor).

Die so ermittelten erwarteten Aufgabenschwierigkeiten werden in DESI herangezogen, um Schwellen zwischen Kompetenzniveaus zu definieren. Unter den verschiedenen Kombinationen von Aufgabenmerkmalen für jeden Test werden solche herausgesucht, die geeignet zur Charakterisierung von Übergängen auf ein neues Kompetenzniveau sind.

Dieses Vorgehen wird im Folgenden an einem hypothetischen Beispiel mit vier Aufgabenmerkmalen schematisch dargestellt; die Anwendung auf einen konkreten Test wird z.B. im Kapitel zum C-Test gut anschaulich (Harsch/Schröder in diesem Band). In Tabelle 1 sind Regressionsgewichte für vier Merkmale aufgelistet; als Werte sind Zahlen in plausiblen Größenordnungen auf der Logit-Skala gewählt, auf der die Aufgabenschwierigkeiten und Personenfähigkeiten im Raschmodell dargestellt werden (vgl. Abbildung 1). Die Regressionskonstante β_0 entspricht der erwarteten Schwierigkeit einer maximal einfachen Aufgabe, die hinsichtlich aller schwierigkeitsbeeinflussenden Merkmale auf den jeweils einfachsten Stufen eingeordnet ist (alle q_{im} in Formel 2 sind null).

Tabelle 1: Hypothetische Regressionsgewichte bei der Vorhersage von Aufgabenschwierigkeiten mit vier Aufgabenmerkmalen.

Merkmale und Regressionsgewichte		**Werte für β_m**
Regressionskonstante	(β_0)	-2.0
Merkmal 1	(β_1)	1.0
Merkmal 2	(β_2)	0.5
Merkmal 3	(β_3)	1.0
Merkmal 4	(β_4)	0.5

Die erwartete Schwierigkeit einer Aufgabe, bei der nur Merkmal 1 vorliegt und alle übrigen Merkmale nicht, wäre in diesem Beispiel 1.0 Logits höher als β_0, nämlich $\beta_0 + \beta_1 = -2.0 + 1.0 = -1.0$. In Tabelle 2 sind mögliche Schwellen zwischen Kompetenzniveaus dargestellt, die sich aus den erwarteten Schwierigkeiten für ausgewählte Merkmalskombinationen ergeben. Der Beginn eines Kompetenzniveaus wird durch Merkmalskombinationen und die daraus resultierenden erwarteten Schwierigkeiten definiert. Schüler auf Niveau A sollten also die einfachsten Aufgaben beherrschen, die keines der schwierigkeitsbestimmenden Merkmale aufweisen. Im hier dargestell-

ten hypothetischen Beispiel würden Schüler auf Niveau B die Aufgaben beherrschen, die das Merkmal 1 aufweisen. Auf Niveau C werden auch die Anforderungen der Merkmale 2 und 3 beherrscht, und auf Niveau D beherrschen Schüler auch Aufgaben, bei denen alle vier Aufgabenmerkmale vorliegen.

Tabelle 2: Definition von Schwellen zwischen vier Kompetenzniveaus durch die aus bestimmten Aufgabenmerkmals-Kombinationen resultierenden erwarteten Schwierigkeiten $\hat{\sigma}_i$.

		Kodierungen q_{im} der Aufgabenmerkmale*			
Niveau	$\hat{\sigma}_i$	**M1**	**M2**	**M3**	**M4**
A	-2.0	0	0	0	0
B	-1.0	1	0	0	0
C	0.5	1	1	1	0
D	1.0	1	1	1	1

**0 = Merkmal liegt nicht vor; 1 = Merkmal liegt vor. Die schattierten Felder zeigen an, auf welchem Niveau ein Merkmal erstmalig auftritt.*

Natürlich sind die zur Schwellendefinition verwendeten Merkmalskombinationen nicht die einzigen, die im Test tatsächlich vorkommen. So können Aufgaben im hypothetischen Beispiel auch die Merkmalskombination 1 und 2 (ohne 3 und 4) aufweisen, hier würde sich eine erwartete Schwierigkeit von $\beta_0 + \beta_1 + \beta_2 = -2.0 + 1.0 + 0.5 = -0.5$ ergeben. In Abbildung 3 sind die erwarteten Schwierigkeiten aus Tabelle 2 sowie die dadurch gebildeten Kompetenzniveaus grafisch auf der fiktiven Kompetenzskala abgetragen.

Die Auswahl der Merkmalskombinationen, welche tatsächlich zur Schwellendefinition herangezogen werden, muss nach theoretischen und empirischen Kriterien erfolgen. Bei der Skalierung der DESI-Tests werden die Aufgaben hierzu nach ihren erwarteten Schwierigkeiten sortiert und dann auf folgende Kriterien geachtet:

- Die Merkmalskombinationen, die als Schwellen verwendet wurden, sollen inhaltlich gut geeignet sein, Unterschiede zwischen Schülern mit niedrigerer und höherer Kompetenz zu charakterisieren.
- Der Test soll mehrere Aufgaben enthalten, bei denen diese Merkmalskombinationen tatsächlich realisiert sind.
- Für diese Merkmalskombinationen sollte es Aufgaben geben, deren tatsächliche Schwierigkeit σ_i nicht zu stark von der erwarteten Schwierigkeit $\hat{\sigma}_i$ abweicht.
- Die Schwellen auf der Kompetenzskala sollten nicht zu dicht beieinander liegen, um eine klare Trennung der Niveaus zu gewährleisten.

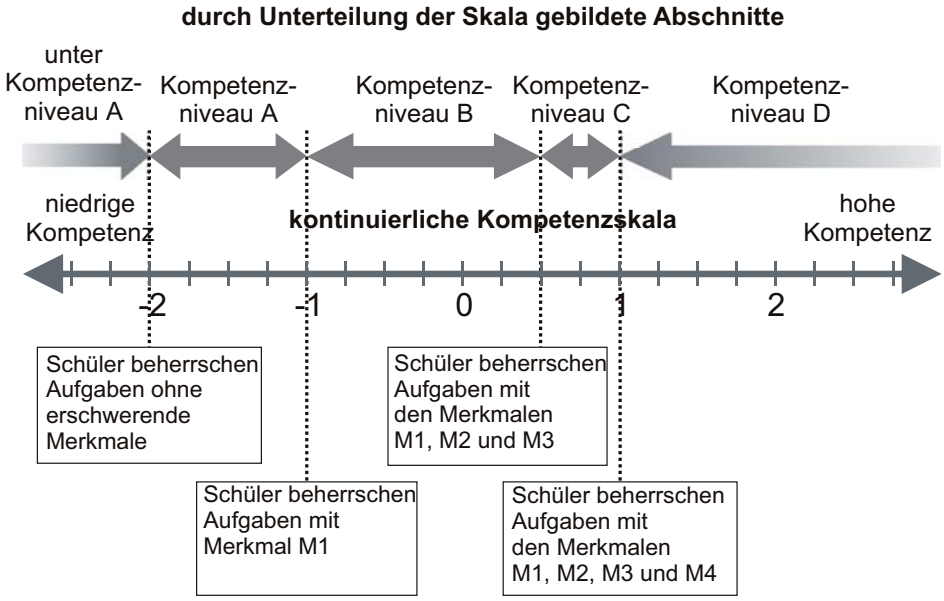

Abbildung 3: Durch Kombinationen von Aufgabenmerkmalen gebildete Kompetenzniveaus auf einer hypothetischen Kompetenzskala.

Merkmalsauswahl und Beschreibung der Kompetenzniveaus

Wie am Anfang dieses Abschnitts erläutert, erlauben die Ergebnisse der Regressionsanalyse sowohl eine Einschätzung der gemeinsamen Vorhersagekraft aller Aufgabenmerkmale als auch eine Einschätzung der Einflüsse der einzelnen Merkmale. Bei der empirischen Analyse der Aufgabenmerkmale und -schwierigkeiten werden für jeden Test die Merkmale mit der höchsten Erklärungskraft für die Aufgabenmerkmale identifiziert. Bei keinem der DESI-Tests werden letztlich alle ursprünglich definierten Merkmale im Regressionsmodell behalten. Dies ergibt sich nicht nur dadurch, dass einzelne Merkmale wider Erwarten keinen deutlichen Einfluss auf die Aufgabenschwierigkeiten zeigten, sondern auch durch Abhängigkeiten (Kollinearitäten) zwischen den Merkmalen. Bei der Aufgabenkonstruktion werden schwierigkeitsbestimmende Merkmale häufig in Kombination miteinander realisiert, so dass Aufgaben oft gleichzeitig hinsichtlich mehrerer Merkmale als leicht oder schwer einzuschätzen sind. Die hieraus resultierenden Zusammenhänge zwischen den Merkmalen über Aufgaben hinweg führen dazu, dass einige Merkmale keine zusätzliche Erklärungskraft mehr haben, wenn andere Merkmale bereits berücksichtigt sind.

Die Schwellen zwischen den Kompetenzniveaus der DESI-Tests werden daher als die erwarteten Schwierigkeiten auf Basis derjenigen Aufgabenmerkmale gebildet, welche die größte Erklärungskraft für Unterschiede in den Aufgabenschwierigkeiten haben. Die inhaltliche Definition der Schülerkompetenzen auf den so gebildeten Kompetenzniveaus wird entsprechend zuerst durch diese vorhersagestärksten

Merkmale vorgenommen. Dennoch werden zusätzlich auch Merkmale, welche aufgrund von Zusammenhängen innerhalb der Merkmale nicht ins Regressionsmodell aufgenommen wurden, zur inhaltlichen Beschreibung der erfassten Sprachkompetenzen herangezogen.

Es ist hier zu betonen, dass als Kriterium bei der Auswahl der Merkmale *nicht* die Signifikanz der einzelnen Regressionsgewichte oder des inkrementellen Anteils zusätzlich erklärter Varianz herangezogen wird. Bei den zur Niveaudefinition vorgenommenen Regressionsanalysen geht es nicht darum, allgemeingültige Modelle zu prüfen, welche für Populationen von Items Gültigkeit haben sollen. Ziel ist es vielmehr vor allem, Modelle zur *Beschreibung* der spezifischen in DESI verwendeten Leistungstests und der darin enthaltenen Aufgaben zu entwickeln. Als Kriterium für die Aufnahme von Aufgabenmerkmalen in diese beschreibenden Modelle wurde die absolute Größe der Regressionsgewichte sowie inhaltlich sinnvolle Vorzeichen herangezogen. So wurden z.B. Merkmale aus den Modellen ausgeschlossen, wenn sie im Gesamtmodell wider Erwarten ein negatives Vorzeichen aufwiesen. Der Verzicht auf eine Generalisierbarkeit der Modelle macht es umso wichtiger, die Passung des Modells auf Ebene der einzelnen Aufgaben zu inspizieren. Dies wurde insbesondere bei der Auswahl der Schwellen zwischen den Kompetenzniveaus beachtet (s.o., Punkte 2 und 3).

Vorgehen bei globalen Aufgabeneinstufungen

Nicht bei allen DESI-Tests ist es möglich, eine differenzierte Beschreibung der Aufgaben hinsichtlich mehrerer Merkmale vorzunehmen. Für einige Tests (Deutsch Lesekompetenz, Deutsch Wortschatz und Deutsch Argumentation) liegen die Aufgabenbeschreibungen in Form globaler Einstufungen des Anspruchsniveaus vor, d.h. jede Aufgabe wird einem angenommenen Schwierigkeitsgrad zugeordnet. Für diese Tests wird ein anderes Vorgehen zur Definition der Kompetenzniveaus gewählt, da die einfachere Aufgabenbeschreibung eine differenziertere Betrachtung der genauen Verteilungen der Aufgabenschwierigkeiten erlaubt. Zur Definition der Kompetenzniveaus wurden die Aufgaben jedes angenommenen Niveaus nach ihrer Schwierigkeit sortiert und dann anhand der tatsächlichen Schwierigkeiten jener Punkt auf der Kompetenzskala ermittelt, ab dem 50% der Aufgaben eines Niveaus beherrscht werden. Bei ungerader Aufgabenanzahl innerhalb eines Niveaus wird die nächste Schwierigkeit unterhalb von 50% gewählt (z.B. die fünfte von elf Aufgaben), bei einer geraden Aufgabenanzahl die mittlere Aufgabenschwierigkeit (d.h. der Median). Dieses Vorgehen führt zu Schwellen zwischen Niveaus, bei welchen die Kompetenz der Schüler dadurch charakterisiert ist, dass sie ca. die Hälfte der Aufgaben eines Niveaus mit hinreichender Sicherheit (65%-Schwelle) beherrschen. Die inhaltliche Beschreibung der Niveaus ergibt sich unmittelbar aus der Beschreibung der vorab angenommenen Niveaus.

Diskussion und Ausblick

Generell gilt, dass die Unterteilung einer kontinuierlichen Skala in ordinale Niveaus mit einem Informationsverlust verbunden ist, da die Unterschiede zwischen Schülern innerhalb eines Niveaus nicht mehr berücksichtigt werden. Dennoch wird der Mehrwert, der durch die kriteriumsorientierte Skaleninterpretation gewonnen wird, als hinreichender Ausgleich für diese Informationsreduktion betrachtet. Es darf jedoch nicht in Vergessenheit geraten, dass sich die Leistungen von Schülern, die kurz unter- und oberhalb einer Schwelle zwischen zwei Niveaus liegen, ähnlicher sind, als die Leistungen von Schülern, die im unteren und oberen Bereich desselben liegen. Die durch die Einteilung in Niveaus vorgenommene Informationsreduktion ist jedoch insofern nicht gravierend, als diese Niveaus im Wesentlichen zur Veranschaulichung deskriptiver Ergebnisse verwendet werden. Für differenzierte Analysen, z.B. zur Vorhersage von Leistungsunterschieden, werden i.d.R. weiterhin die ursprünglichen quantitativen Skalenwerte verwendet.

Grenzen des Vorhersagemodells

Zur Beurteilung der hier geschilderten Vorgehensweise ist anzumerken, dass das linear-additive Modell bei der Vorhersage der Aufgabenschwierigkeiten ein relativ einfaches Modell darstellt. Andere Modelle zur Vorhersage der Schwierigkeiten sind ohne weiteres denkbar. So sind z.B. Wechselwirkungen zwischen Merkmalen dahingehend möglich, dass eine Kombination von zwei Merkmalen die Testaufgaben schwieriger macht, als aufgrund einer einfachen Addition der Schwierigkeiten der einzelnen Merkmale zu erwarten wäre. In DESI wurde routinemäßig eine Prüfung vorgenommen, ob derartige Wechselwirkungen zu beobachten sind. Die Ergebnisse sprechen nicht dagegen, das hier geschilderte einfache Modell ohne Wechselwirkungen beizubehalten. Zudem nimmt die Interpretierbarkeit der Kompetenzniveaus mit zunehmender Komplexität des Vorhersagemodells ab. Zu beachten ist auch, dass die Anzahl der Aufgaben für einzelne Tests relativ gering ist – die Schätzungen der Einflüsse der Aufgabenmerkmale sollte insbesondere in diesen Fällen ohne zusätzliche Untersuchungen an neuen Aufgaben nicht über das hier verwendete Testmaterial hinaus generalisiert werden. Wichtig ist in jedem Fall eine Prüfung der Abweichungen ε_i zwischen beobachteten und vorhergesagten Schwierigkeiten, diese können auch zur Suche nach Fehlspezifikationen des Modells herangezogen werden.

Alternative Schätzverfahren

Das für die Definition der Kompetenzniveaus gewählte Vorgehen beinhaltet im Wesentlichen zwei separate Analyseschritte: Die Rasch-Skalierung der Testdaten mit Schätzung der Aufgabenschwierigkeiten und die daran anschließende Vorhersage der Schwierigkeiten durch angenommene Aufgabenmerkmale. Der Vorteil dieses Vorgehens in zwei Schritten ist, dass die Analysen auf weit verbreite-

ten Standardmethoden basieren, womit sie für ein breites Publikum nachvollziehbar und anwendbar sind. Dass angenommene linear-additive Modell zur Erklärung der Aufgabenschwierigkeiten könnte jedoch auch in einer gemeinsamen Analyse mit der Skalierung geschätzt werden. Eine Möglichkeit hierzu ist die Skalierung mit einem Linear-Logistischen Testmodell (LLTM, z.B. Fischer 1996). Hierbei wird schon bei der Rasch-Skalierung eine Dekomposition der Aufgabenschwierigkeiten in eine gewichtete Kombination von so genannten Basisparametern vorgenommen, wobei jeder dieser Basisparameter ein Aufgabenmerkmal repräsentiert. Diese Basisparameter können analog den Regressionsgewichten bei der für DESI gewählten Methode zur Definition von Kompetenzniveaus herangezogen werden. Ein wesentlicher Unterschied der Auswertung mittels LLTM ist, dass die Aufgabenschwierigkeiten durch die Aufgabenmerkmale vollständig erklärt werden, d.h. es wird kein Residuum zugelassen. Zudem ist es im Rahmen des LLTM schwierig, die Güte hinsichtlich der Aufgabenmerkmale gemachten Annahmen einzuschätzen (Hartig 2004a). Analysen mit einzelnen DESI-Tests ergeben allerdings, dass die Verwendung von Regressionsanalysen und LLTM zu inhaltlich sehr ähnlichen Ergebnissen führen (Hartig 2004b). Neuere Item-Response-Modelle ermöglichen wie das LLTM die Schätzung von Merkmalseffekten auf die Aufgabenschwierigkeiten innerhalb der Skalierung, lassen jedoch Residuen zu (Janssen/Tuerlinckx/Meulders/De Boeck 2000; Janssen/Schepers/Peres 2004). Diese Modelle führen mit einem Analyseschritt zu inhaltlich fast identischen Ergebnissen wie das in DESI gewählte zweischrittige Vorgehen (Hartig/Frey 2005). Die simultane Modellierung beider Schritte im Rahmen komplexerer Analysemodelle erscheint der Datenlage und Fragestellung zwar methodisch angemessener, ist jedoch mit einem deutlichen Mehraufwand verbunden. Da die hierzu durchgeführten Analysen keine nennenswerten inhaltlichen Unterschiede zwischen den Ergebnissen aus den verschiedenen Modellierungsverfahren ergeben, wird in DESI auch aus Gründen der Nachvollziehbarkeit der Ergebnisse das einfachere Verfahren mit separater Skalierung und Regressionsanalyse gewählt.

Vorzüge von a priori definierten Merkmalen

In der empirischen Praxis sind die Grenzen zwischen einem Post-Hoc-Vorgehen und einem streng hypothesengeleiteten und modellbasierten Vorgehen fließend. Auch im Nachhinein kann z.B. eine Vorhersage der Aufgabenschwierigkeiten mittels Aufgabenmerkmalen vorgenommen werden (z.B. Prenzel/Häußler/Rost/Senkbeil 2002). Mit einer A-Priori-Beschreibung der Aufgabenmerkmale gehen jedoch verschiedene Vorteile einher. Vor allem lassen sich empirisch gestützte Aussagen, welche auf Basis vorab formulierter Kompetenzmodelle und daraus abgeleiteter Hypothesen erzielt wurden, besser über das eingesetzte Aufgabenmaterial hinaus verallgemeinern. So können Aufgabenmerkmale und die darauf basierenden Kompetenzniveaus z.B. auf andere Tests oder neu konstruierte Aufgaben übertragen werden. Die Aufgabenmerkmale in DESI wurden für alle Tests vorab definiert, teilweise jedoch auch noch nach Vorliegen der Skalierungsergebnisse revidiert. Insofern

sind die resultierenden Kompetenzniveaus auch in DESI zunächst an das verwendete Testmaterial gebunden. Es erscheint daher als eine interessante und vielversprechende Forschungsfrage, inwieweit die in DESI entwickelten Kompetenzmodelle sich mit neuen Aufgaben und an anderen Stichproben bestätigen lassen. Mit den vorliegenden differenzierten Merkmalsdefinitionen sind die Voraussetzungen für derartige Untersuchungen ideal.

Abschließend ist hervorzuheben, dass die Möglichkeiten eines hypothesengeleiteten Vorgehens, völlig unabhängig von der eingesetzten Analysemethodik, an die Qualität der zugrunde liegenden inhaltlich-theoretischen Kompetenzmodelle gebunden sind: Je differenzierter die Vorstellungen über die Natur der zu erfassenden Kompetenz und der Prozesse beim Bewältigen spezifischer Aufgabenanforderungen, desto differenzierter können die Anforderungsmerkmale konkreten Testmaterials beschrieben werden und desto eher führt die Verankerung der Aufgabenanforderungen auf der Kompetenzskala zu praktisch nutzbaren Kompetenzstufen.

Literatur

Adams, R./Wu, M. (Eds.) (2002): PISA 2000 technical report. Paris: OECD.

Artelt, C./Stanat, P./Schneider, W./Schiefele, U. (2001): Lesekompetenz: Testkonzeption und Ergebnisse. In: Deutsches PISA-Konsortium (Hrsg.): PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske + Budrich.

Beaton, E./Allen, N. (1992): Interpreting scales through scale anchoring. In: Journal of Educational Statistics 17, S. 191-204.

Borsboom, D./Mellenbergh, G.J./van Heerden, J. (2003): The Theoretical Status of Latent Variables. In: Psychological Review 110, S. 203–219.

Borsboom, D./Mellenbergh, G.J./van Heerden, J. (2004): The Concept of Validity. In: Psychological Review 111, S. 1061–1071.

Bos, W./Lankes, E.-M./Schwippert, K./Valtin, R./Voss, A./Badel, I./Plaßmeier, N. (2003): Lesekompetenzen deutscher Grundschülerinnen und Grundschüler am Ende der vierten Jahrgangsstufe im internationalen Vergleich. In: Bos, W./Lankes, E.-M./Prenzel, M./Schwippert, K./Walther G./Valtin, R. (Hrsg.): Erste Ergebnisse aus IGLU. Münster, New York: Waxmann, S. 69-142.

Fischer, G.H. (1996): Unidimensional linear logistic rasch models. In: Linden, W.J./van der/Hambleton, R.K. (Eds.): Handbook of modern item response theory. New York, Berlin: Springer. S. 225-243.

Fischer, G.H./Molenaar, I.W. (Eds.) (1991): Rasch models. Foundations, recent developments, and applications. New York, Berlin: Springer.

Hartig, J. (2004a): Assessing the appropriateness of specifications in LLTM weight matrices. Paper presented at the 24th Biennial Conference of the Society for Multivariate Analysis in the Behavioral Sciences in Jena, July 18th to July 21st.

Hartig, J. (2004b): Methoden zur Bildung von Kompetenzstufenmodellen. In: Moosbrugger, H./Rauch, W./Frank, D. (Hrsg.): Qualitätssicherung im Bildungswesen. Frankfurt a.M.: Arbeiten aus dem Institut der J.W. Goethe-Universität, Heft 2004/03.

Hartig, J./Frey, A. (2005): Application of different explanatory item response models for model based proficiency scaling. Paper presented at the 70th Annual Meeting of the Psychometric Society in Tilburg, July 5–8.

Hartig, J./Kühnbach, O. (im Druck). Schätzung von Veränderung mit Plausible Values in mehrdimensionalen Rasch-Modellen. In: Ittel, A./Merkens, H. (Hrsg.): Veränderungsmessung und Längsschnittstudien in der empirischen Erziehungswissenschaft. Wiesbaden: VS Verlag für Sozialwissenschaften.

Helmke, A./Hosenfeld, I. (2004): Vergleichsarbeiten – Standards – Kompetenzstufen: Begriffliche Klärungen und Perspektiven. In: Jäger, R.S./Frey, A. (Hrsg.): Lernprozesse, Lernumgebung und Lerndiagnostik. Wissenschaftliche Beiträge zum Lernen im 21. Jahrhundert. Landau: Verlag Empirische Pädagogik.

Janssen, R./Schepers, J./Peres, D. (2004): Models with item and item group predictors. In De Boeck, P./Wilson, M. (Eds.): Explanatory item response models: A generalized linear and nonlinear approach. New York, Berlin: Springer. S. 189-212.

Janssen, R./Tuerlinckx, F./Meulders, M./De Boeck, P. (2000): A hierarchical IRT model for criterion-referenced measurement. Journal of Educational and Behavioral Statistics 25, S. 285-306.

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.H. (Hrsg.): TIMSS/III. Dritte internationale Mathematik- und Naturwissenschaftsstudie. Band 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit. Opladen: Leske + Buderich.

Linden, W.J. van der /Hambleton, R.K. (Eds.) (1997): Handbook of modern item response theory. New York, Berlin: Springer.

Mislevy, R.J./Beaton, A.E./Kaplan, B./Sheehan, K.M. (1992): Estimating population characteristics from sparse matrix samples of responses. In: Journal of Educational Measurement 29, S. 133-161.

Moosbrugger, H. (^{3}2002): Lineare Modelle. Regressions- und Varianzanalysen. Göttingen, Bern: Huber.

Moosbrugger, H./Hartig, J. (2003): Klassische Testtheorie. In Kubinger, K./Jäger, R. (Hrsg.): Schlüsselbegriffe der Psychologischen Diagnostik. Weinheim: Psychologie Verlags Union. S. 408-415.

OECD (2004): Lernen für die Welt von morgen. Erste Ergebnisse von PISA 2003. Paris: OECD.

PISA-Konsortium Deutschland (Hrsg.) (2004): PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs. Münster: Waxmann.

Prenzel, M./Häußler, P./Rost, J./Senkbeil, M. (2002): Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? In: Unterrichtswissenschaft 30, S. 120-135.

Rost, J. (^{2}2004): Lehrbuch Testtheorie – Testkonstruktion. Bern, Göttingen: Huber.

Watermann, R./Klieme, E. (2002): Reporting Results of Large-Scale Assessment in Psychologically and Educationally Meaningful Terms – Construct Validation and Proficiency Scaling in TIMSS. European Journal of Psychological Assessment 18, S. 190-203.

Wilson, M.R. (2003): On choosing a model for measuring. Methods of Psychological Research Online 8, S. 1-22.

Wu, M.L./Adams, R.J./Wilson, M.R. (1998): ConQuest: Generalized item response modelling software. Melbourne: Australian Council for Educational Research.