

Frey, Andreas; Seitz, Nicki-Nils

Multidimensionale adaptive Kompetenzdiagnostik: Ergebnisse zur Messeffizienz. Projekt MAT

Klieme, Eckhard [Hrsg.]; Leutner, Detlev [Hrsg.]; Kenk, Martina [Hrsg.]: Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. Weinheim ; Basel : Beltz 2010, S. 40-51. - (Zeitschrift für Pädagogik, Beiheft; 56)



Quellenangabe/ Reference:

Frey, Andreas; Seitz, Nicki-Nils: Multidimensionale adaptive Kompetenzdiagnostik: Ergebnisse zur Messeffizienz. Projekt MAT - In: Klieme, Eckhard [Hrsg.]; Leutner, Detlev [Hrsg.]; Kenk, Martina [Hrsg.]: Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. Weinheim ; Basel : Beltz 2010, S. 40-51 - URN: urn:nbn:de:0111-opus-33446 - DOI: 10.25656/01:3344

<https://nbn-resolving.org/urn:nbn:de:0111-opus-33446>

<https://doi.org/10.25656/01:3344>

in Kooperation mit / in cooperation with:

BELTZ

<http://www.beltz.de>

Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen. Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Zeitschrift für Pädagogik · 56. Beiheft

Kompetenzmodellierung

Zwischenbilanz des DFG- Schwerpunktprogramms und Perspektiven des Forschungsansatzes

Herausgegeben von

Eckhard Klieme, Detlev Leutner und Martina Kenk

BELTZ

Die in der Zeitschrift veröffentlichten Beiträge sind urheberrechtlich geschützt. Alle Rechte, insbesondere das der Übersetzung in fremde Sprachen, vorbehalten. Kein Teil dieser Zeitschrift darf ohne schriftliche Genehmigung des Verlages in irgendeiner Form – durch Fotokopie, Mikrofilm oder ein anderes Verfahren – reproduziert oder in eine von Maschinen, insbesondere Datenverarbeitungsanlagen, verwendbare Sprache übertragen werden. Auch die Rechte der Wiedergabe durch Vortrag, Funk- und Fernsehsendung, im Magnettonverfahren oder auf ähnlichem Wege bleiben vorbehalten. Fotokopien für den persönlichen oder sonstigen eigenen Gebrauch dürfen nur von einzelnen Beiträgen oder Teilen daraus als Einzelkopie hergestellt werden. Jede im Bereich eines gewerblichen Unternehmens hergestellte oder genutzte Kopie dient gewerblichen Zwecken gem. § 5 4(2) UrhG und verpflichtet zur Gebührenzahlung an die VG Wort, Abteilung Wissenschaft, Goethestr. 49, 80336 München, bei der die einzelnen Zahlungsmodalitäten zu erfragen sind.

© 2010 Beltz Verlag · Weinheim und Basel
Herstellung: Lore Amann
Gesamtherstellung: Druckhaus „Thomas Müntzer“, Bad Langensalza
Printed in Germany
ISSN 0514-2717
Bestell-Nr. 41157

Inhaltsverzeichnis

Eckhard Klieme/Detlev Leutner/Martina Kenk
Kompetenzmodellierung. Eine aktuelle Zwischenbilanz des DFG-Schwerpunkt-
programms. Einleitung zum Beiheft 9

Benő Csapó
Goals of Learning and the Organization of Knowledge 12

Mathematische Kompetenzen

Marianne Bayrhuber/Timo Leuders/Regina Bruder/Markus Wirtz
Projekt HEUREKO
Repräsentationswechsel beim Umgang mit Funktionen – Identifikation von
Kompetenzprofilen auf der Basis eines Kompetenzstrukturmodells 28

Andreas Frey/Nicki-Nils Seitz
Projekt MAT
Multidimensionale adaptive Kompetenzdiagnostik: Ergebnisse zur
Messeffizienz 40

*Nina Zeuch/Hanneke Geerlings/Heinz Holling/Wim J. van der Linden/
Jonas P. Bertling*
Projekt Regelgeleitete Itementwicklung
Regelgeleitete Konstruktion von statistischen Textaufgaben: Anwendung von
linear logistischen Testmodellen und Aufgabencloning 52

*Eckhard Klieme/Anika Bürgermeister/Birgit Harks/Werner Blum/Dominik Leiß/
Katrin Rakoczy*
Projekt Co²CA
Leistungsbeurteilung und Kompetenzmodellierung im Mathematikunterricht 64

Olga Kunina-Habenicht/Oliver Wilhelm/Franziska Matthes/André A. Rupp
Projekt Kognitive Diagnosemodelle
Kognitive Diagnosemodelle: Theoretisches Potential und methodische Probleme ... 75

Aiso Heinze

Review

Mathematische Kompetenz modellieren und diagnostizieren: Eine Diskussion der Forschungsprojekte des DFG-Schwerpunktprogramms „Kompetenzmodelle“ aus mathematikdidaktischer Sicht 86

Naturwissenschaftliche Kompetenzen

Tobias Viering/Hans E. Fischer/Knut Neumann

Projekt Physikalische Kompetenz

Die Entwicklung physikalischer Kompetenz in der Sekundarstufe I 92

Renate Soellner/Stefan Huber/Norbert Lenartz/Georg Rudinger

Projekt Gesundheitskompetenz

Facetten der Gesundheitskompetenz – eine Expertenbefragung 104

Ilonca Hardy/Thilo Kleickmann/Susanne Koerber/Daniela Mayer/

Kornelia Möller/Judith Pollmeier/Knut Schwippert/Beate Sodian

Projekt Science – P

Die Modellierung naturwissenschaftlicher Kompetenz im Grundschulalter 115

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

Projekt Umweltkompetenz

Umweltkompetenz – Modellierung, Entwicklung und Förderung 126

Ilka Parchmann

Review

Kompetenzmodellierung in den Naturwissenschaften – Vielfalt ist wertvoll, aber nicht ohne ein gemeinsames Fundament 135

Sprachliche und Lesekompetenzen

Wolfgang Schnotz/Nele McElvany/Holger Horz/Sascha Schroeder/Mark Ullrich/

Jürgen Baumert/Axinja Hachfeld/Tobias Richter

Projekt BITE

Das BITE-Projekt: Integrative Verarbeitung von Bildern und Texten in der Sekundarstufe I 143

Tobias Dörfler/Stefanie Golke/Cordula Artelt

Projekt Dynamisches Testen

Dynamisches Testen der Lesekompetenz: Theoretische Grundlagen, Konzeption und Testentwicklung 154

<i>Thorsten Roick/Petra Stanat/Oliver Dickhäuser/Volker Frederking/ Christel Meier/Lydia Steinhauer</i>	
Projekt Literarästhetische Urteilskompetenz	
Strukturelle und kriteriale Validität der literarästhetischen Urteilskompetenz	165

<i>Hans Anand Pant/Simon P. Tiffin-Richards/Olaf Köller</i>	
Projekt Standard-Setting	
Standard-Setting für Kompetenztests im Large-Scale-Assessment	175

<i>Johannes Hartig/Jana Höhler</i>	
Projekt MIRT	
Modellierung von Kompetenzen mit mehrdimensionalen IRT-Modellen	189

<i>Albert Bremerich-Vos</i>	
Review	
Modellierung von Aspekten sprachlich-kultureller Kompetenz. Anmerkungen zu den Projektberichten	199

Fächerübergreifende Kompetenzen

<i>Ellen Gausmann/Sabina Eggert/Marcus Hasselhorn/Rainer Watermann/ Susanne Bögeholz</i>	
Projekt Bewertungskompetenz	
Wie verarbeiten Schüler/-innen Sachinformationen in Problem- und Entscheidungssituationen Nachhaltiger Entwicklung – Ein Beitrag zur Bewertungskompetenz	204

<i>Samuel Greiff/Joachim Funke</i>	
Projekt Dynamisches Problemlösen	
Systematische Erforschung komplexer Problemlösefähigkeit anhand minimal komplexer Systeme	216

<i>Klaus Lingel/Nora Neuenhaus/Cordula Artelt/Wolfgang Schneider</i>	
Projekt EWIKO	
Metakognitives Wissen in der Sekundarstufe: Konstruktion und Evaluation domänenspezifischer Messverfahren	228

<i>Jens Fleischer/Joachim Wirth/Stefan Rumann/Detlev Leutner</i>	
Projekt Problemlösen	
Strukturen fächerübergreifender und fachlicher Problemlösekompetenz – Analyse von Aufgabenprofilen	239

Melanie Schütte/Joachim Wirth/Detlev Leutner

Projekt Selbstregulationskompetenz

Selbstregulationskompetenz beim Lernen aus Sachtexten – Entwicklung und
Evaluation eines Kompetenzstrukturmodells 249

Tobias Gschwendtner/Bernd Geißel/Reinhold Nickolaus

Projekt Berufspädagogik

Modellierung beruflicher Fachkompetenz in der gewerblich-technischen
Grundbildung 258

Franziska Perels

Review

Modellierung und Messung fächerübergreifender Kompetenzen und ihre
Bedeutung für die Bildungsforschung. Kritische Reflexion der Projektbeiträge ... 270

Lehrerkompetenzen

Simone Bruder/Julia Klug/Silke Hertel/Bernhard Schmitz

Projekt Beratungskompetenz

Modellierung der Beratungskompetenz von Lehrkräften 274

Cornelia Gräsel/Sabine Krolak-Schwerdt/Ines Nölle/Thomas Hörstermann

Projekt Diagnostische Kompetenz

Diagnostische Kompetenz von Grundschullehrkräften bei der Erstellung der
Übergangsempfehlung: eine Analyse aus der Perspektive der sozialen
Urteilsbildung 286

Tina Seidel/Geraldine Blomberg/Kathleen Stürmer

Projekt OBSERVE

„OBSERVER“ – Validierung eines videobasierten Instruments zur Erfassung
der professionellen Wahrnehmung von Unterricht 296

Mareike Kunter

Review

Modellierung von Lehrerkompetenzen. Kommentierung der
Projektdarstellungen 307

Multidimensionale adaptive Kompetenzdiagnostik: Ergebnisse zur Messeffizienz

Projekt MAT¹

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 dreiparametrischen 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 $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)$ und drei Itemparametern konzeptualisiert:

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

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 $\hat{\theta}$ 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/Schaffer 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-priori-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:

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

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-

Dimensionsanzahl	Korrelation	Testalgorithmus					
		FIT		U-CAT		MAT	
		<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
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

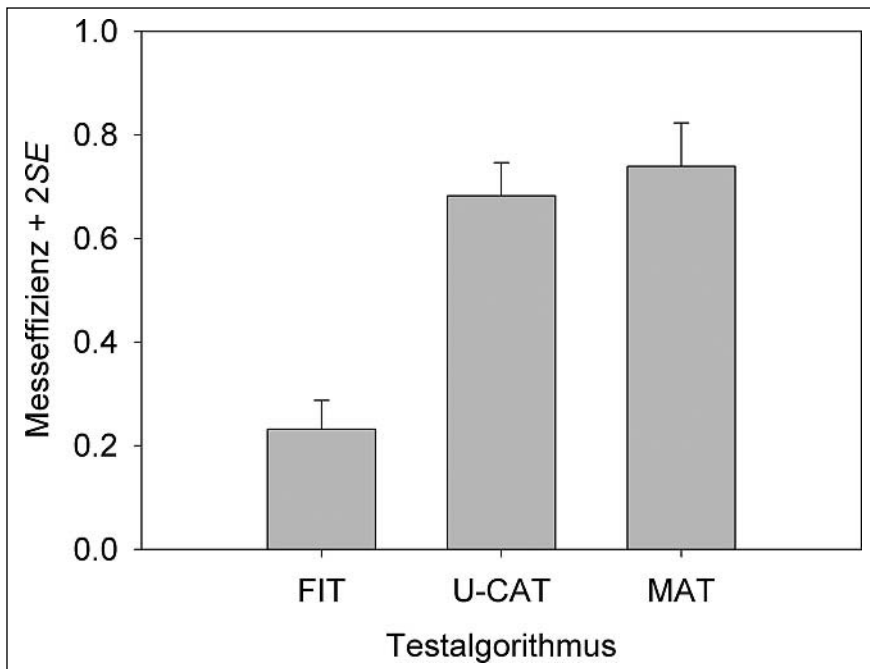


Abb. 1: Messeffizienz in Abhängigkeit des Testalgorithmus

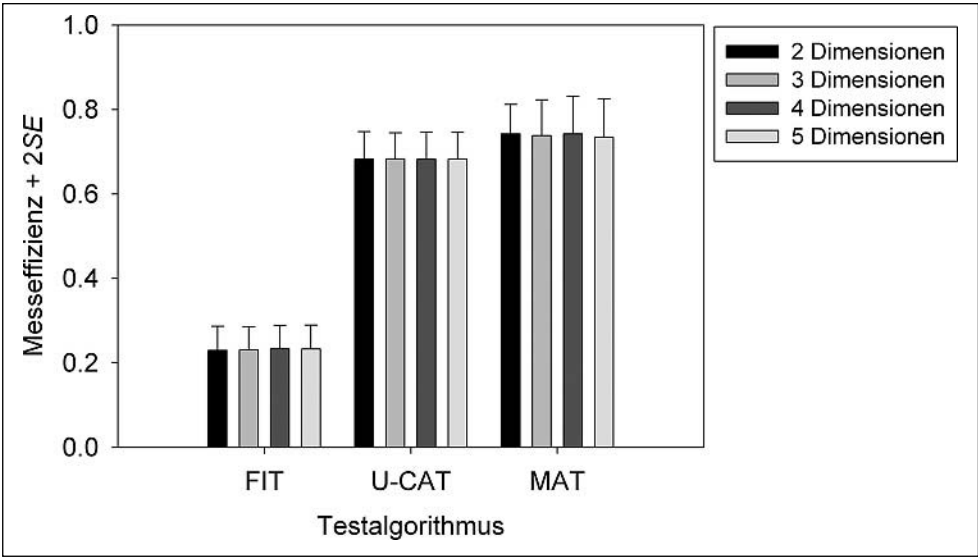


Abb. 2: Messeffizienz in Abhängigkeit von Testalgorithmus und Dimensionsanzahl

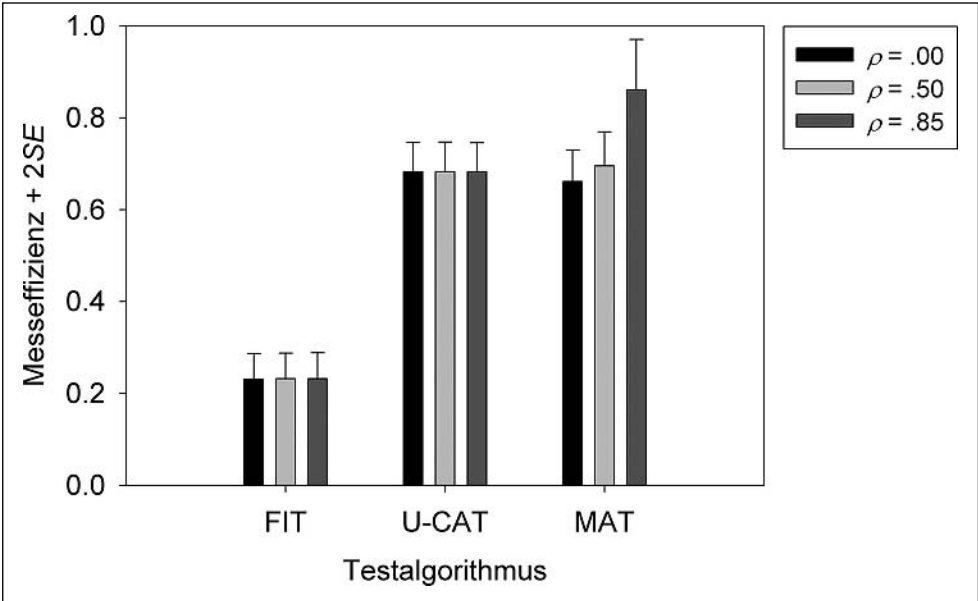


Abb. 3: Messeffizienz in Abhängigkeit von Testalgorithmus und Korrelation zwischen den Dimensionen

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.

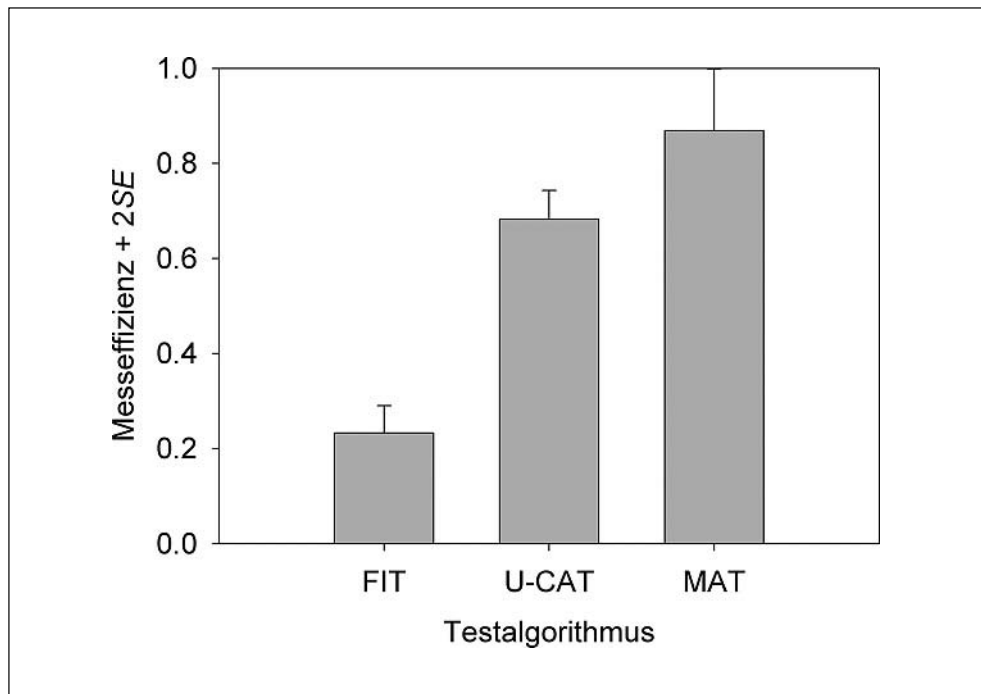


Abb. 4: Messeffizienz bei für Erhebungen zu den Bildungsstandards typischen Bedingungen nach Testalgorithmus

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 Ergebnismeldung 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