pedocs
Open Access Erziehungswissenschaften

DIPF

Pant, Hans Anand; Tiffin-Richards, Simon P.; Köller, Olaf

Standard-Setting für Kompetenztests im Large-Scale-Assessment. Projekt Standardsetting

Klieme, Eckhard [Hrsg.]; Leutner, Detlev [Hrsg.]; Kenk, Martina [Hrsg.]: Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. Weinheim ; Basel : Beltz 2010, S. 175-188. - (Zeitschrift für Pädagogik, Beiheft; 56)

Quellenangabe/ Reference:
Pant, Hans Anand; Tiffin-Richards, Simon P.; Köller, Olaf: Standard-Setting für Kompetenztests im Large-Scale-Assessment. Projekt Standardsetting - In: Klieme, Eckhard [Hrsg.]; Leutner, Detlev [Hrsg.]; Kenk, Martina [Hrsg.]: Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. Weinheim ; Basel : Beltz 2010, S. 175-188 - URN: urn:nbn:de:0111-opus-34067 - DOI: 10.25656/01:3406

https://nbn-resolving.org/urn:nbn:de:0111-opus-34067
https://doi.org/10.25656/01:3406

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

Mitglied der Leibniz-Gemeinschaft

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 genützte Kopie dient gewerblichen Zwecken gem. § 54 (2) UrhG und verpflichtet zur Gebührenzahlung an die VGWort, AbteilungWissenschaft, 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

Mathematische Kompetenzen

Naturwissenschaftliche Kompetenzen

Sprachliche und Lesekompetenzen

Fächerübergreifende Kompetenzen

Lehrerkompetenzen

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

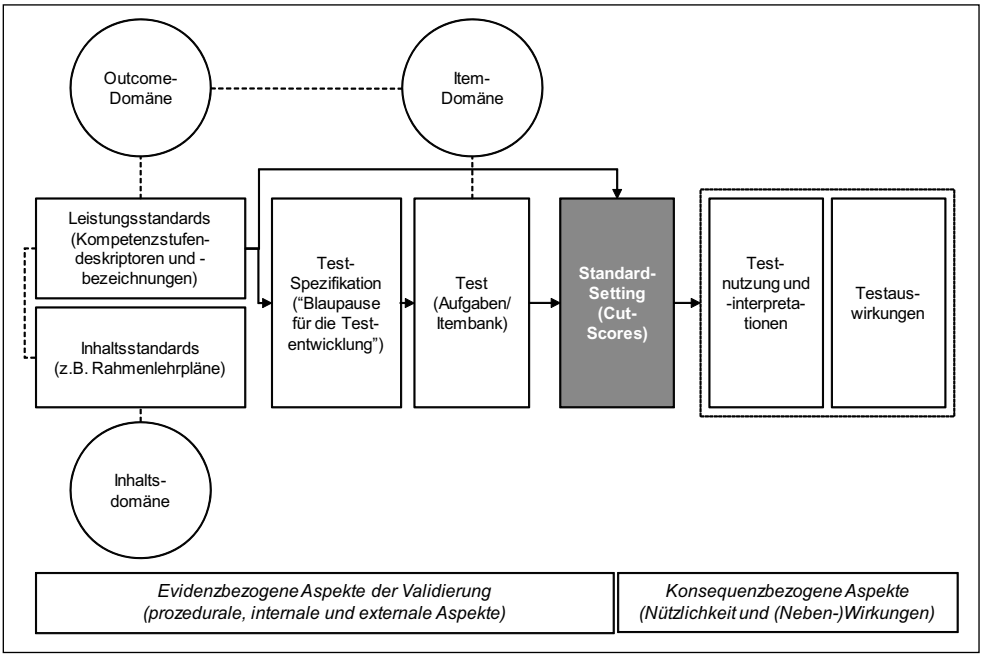

Abb. 1: Schematische Darstellung der Funktion des Standard-Settings im Large-Scale-Assessment

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: $F = 7.1$; $df_1 = 1$, $df_2 = 20$; $p < .05$; $Eta^2_{part.} = .26$;

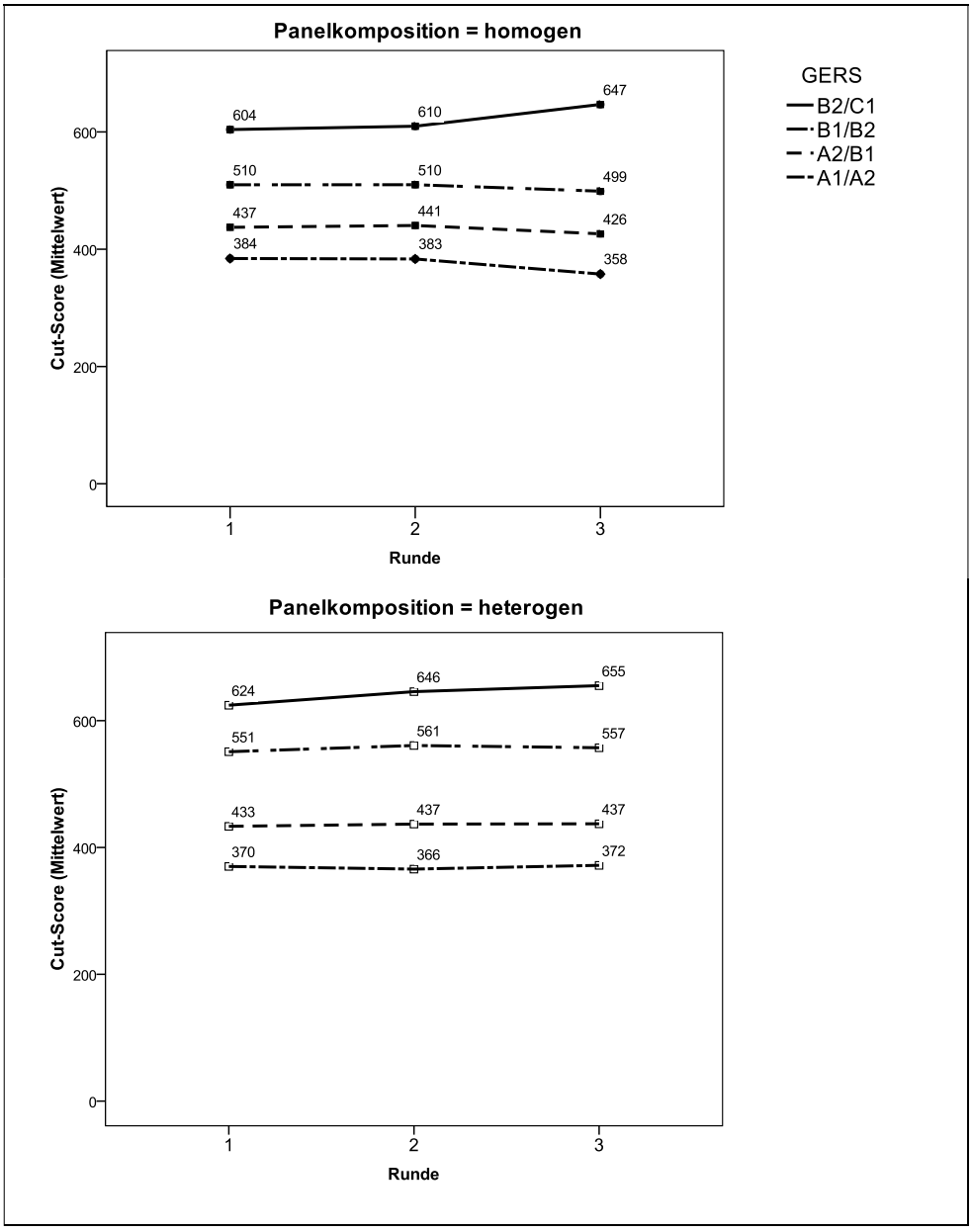

Abb. 2: Gemittelte Cut-Scores im Leseverstehen nach Durchgangsrunde, GERS-Niveau und Panelkomposition unter Verwendung der klassischen Bookmark-Methode

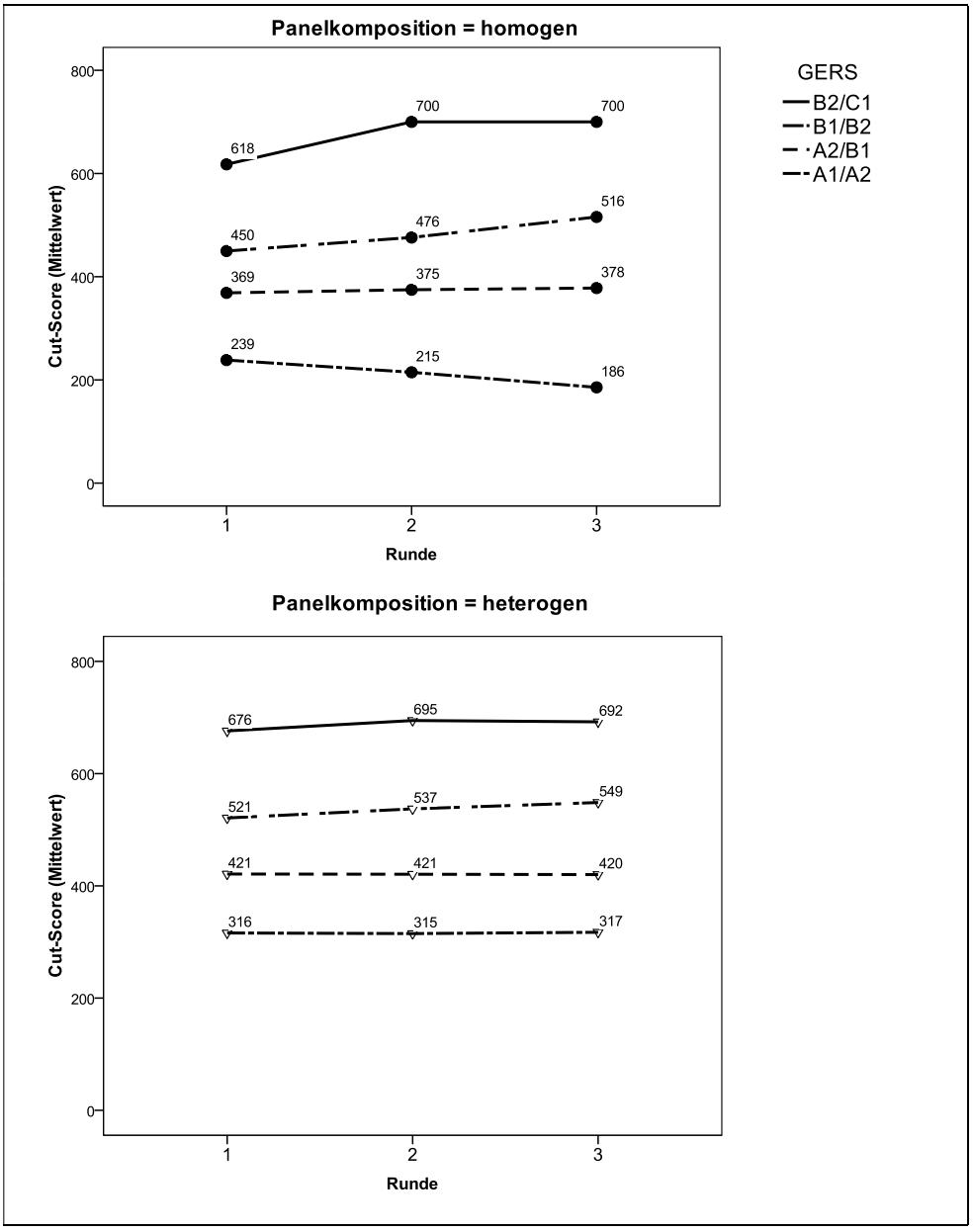

Abb. 3: Gemittelte Cut-Scores im Hörverstehen nach Durchgangsrunde, GERS-Niveau und Panelkomposition unter Verwendung der klassischen Bookmark-Methode

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.

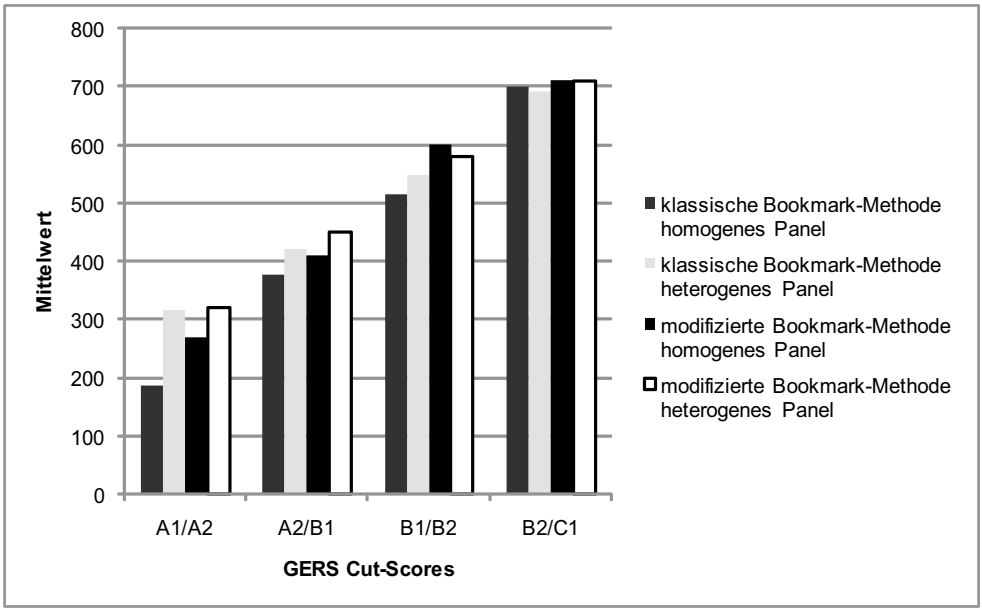

Abb. 4: Finale Cut-Scores aller vier Experimentalgruppen im Leseverstehen nach GERS-Niveau

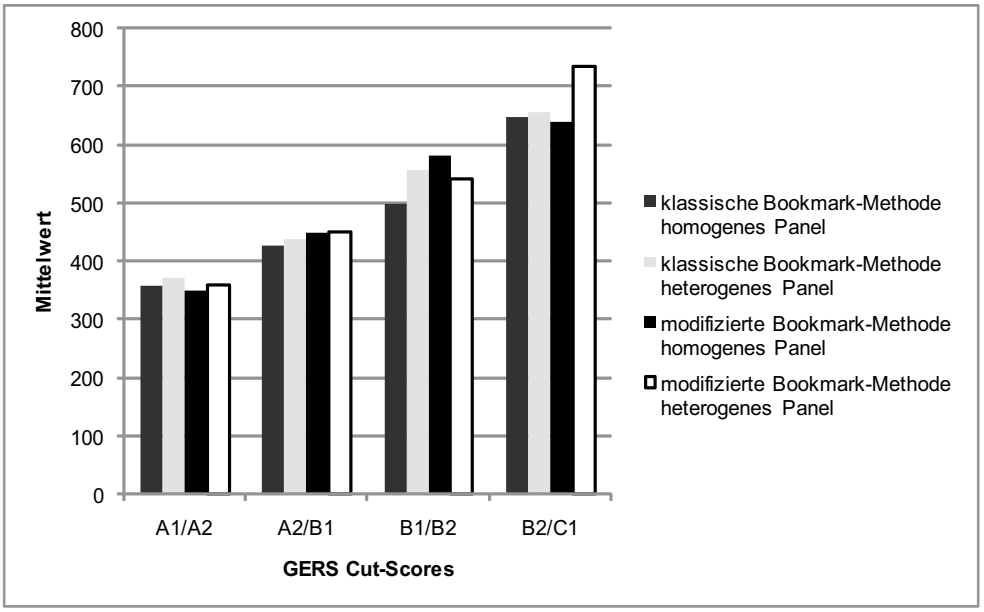

Abb. 5: Finale Cut-Scores aller vier Experimentalgruppen im Hörverstehen nach GERS-Niveau

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