bifie | Bildung

*standards

Standard-Setting Mathematik

Technische Dokumentation – BIST-Ü Mathematik, 4. Schulstufe, 2013

Claudia Bazinger
Roman Freunberger
Ursula Itzlinger-Bruneforth

Bundesinstitut bifie

Bundesinstitut für Bildungsforschung, Innovation & Entwicklung
des österreichischen Schulwesens
Alpenstraße 121 / 5020 Salzburg
www.bifie.at

Standard-Setting Mathematik
Technische Dokumentation – BIST-Ü Mathematik, 4. Schulstufe, 2013*
BIFIE | Department Bildungsstandards & Internationale Assessments (BISTA),
Salzburg 2013

* Der ursprüngliche Titel der Publikation lautete: *Standard-Setting Mathematik 4. Schulstufe. Technischer Bericht.*

Inhaltsverzeichnis

Teil I – Das Standard-Setting für Mathematik auf der 4. Schulstufe

Im Rahmen der Bildungsstandardüberprüfung in Österreich wurd beginnend mit 2012 für Mathematik auf der 4. Schulstufe ein Standard-Setting durchgeführt, das insgesamt 3 Phasen umfasste. Phase III beschreibt das eigentliche Standard-Setting und wird im Weiteren näher erläutert. In Phase I und Phase II wurde durch Fachexpertinnen und Fachexperten ein Kompetenzstufenmodell entwickelt, das 4 Stufen (inkl. der Stufe *Bildungsstandards nicht erreicht*, für die keine Beschreibung entwickelt wurde) umfasst und in Abbildung 1 dargestellt ist.

	Inhaltliche Beschreibung der einzelnen Kompetenzstufen
Stufe 3	**Bildungsstandards übertroffen** Du verfügst über grundlegende Kenntnisse und Fertigkeiten in allen Teilbereichen des Lehrplans Mathematik und über erweiterte Wissensstrukturen, welche über die Anforderungen der Stufe 2 hinausgehen, insbesondere über stärker ausgeprägtes analytisches Denken und höhere Kombinationsfähigkeit. Du kannst diese eigenständig in neuartigen Situationen flexibel einsetzen.
Stufe 2	**Bildungsstandards erreicht** Du verfügst über grundlegende Kenntnisse und Fertigkeiten in allen Teilbereichen des Lehrplans Mathematik und kannst diese flexibel nutzen. Du kannst geeignete Lösungsstrategien finden und umsetzen, gewählte Lösungswege beschreiben und begründen. Du kannst relevante Informationen aus unterschiedlich dargestellten Sachverhalten (z. B. Texten, Datenmaterialien, grafischen Darstellungen) entnehmen. Du kannst diese Informationen zueinander in Beziehung setzen sowie mathematische Fragestellungen daraus ableiten und lösen.
Stufe 1	**Bildungsstandards teilweise erreicht** Du verfügst über grundlegende Kenntnisse und Fertigkeiten in allen Teilbereichen des Lehrplans Mathematik und kannst damit reproduktive Anforderungen bewältigen und Routineverfahren durchführen.
unter 1	**Bildungsstandards nicht erreicht**

Abbildung 1: Kopetenzstufenmodell

Ziel der Phase III war es, Schwellenwerte auf der kontinuierlichen Kompetenzskala zu definieren, welche es erlauben, die Schüler und Schülerinnen den einzelnen Stufen zuzuordnen. Hierzu wurde die Methodik des Standard-Settings verwendet, was im weiteren Sinne einen komplexen Entscheidungsprozess beschreibt, der möglichst standardisiert durchgeführt werden sollte, um valide Schwellenwertsetzungen zu ermöglichen.

Der eigentliche Standard-Setting-Prozess mit einer Expertengruppe sollte in der Domäne Mathematik auf der 4. Schulstufe drei Schwellenwerte (Cut-Scores) hervorbringen. Daraus ergibt sich neben den drei definierten Stufen noch die Stufe *Unter Level 1*. Die Cut-Scores wurden unter Anwendung einer modifizierten Item-Descriptor-Matching-Methode (Cizek, 1996; Cizek & Bunch, 2007; Ferrara, Perie & Johnson, 2002) und der Bookmark-Methode (Mitzel, Lewis, Green & Patz, 1999) durch eine Expertengruppe bestimmt. Der Workshop zum Standard-Setting fand von 24. 01. bis 25. 01. 2013 am BIFIE Salzburg statt.

1 Verwendete Methoden

1.1 Bookmark-Methode

Die Bookmark-Methode ist eine der am häufigsten eingesetzten Methoden in Standard-Settings (Karantonis & Sireci, 2006) und wurde von Mitzel et al. (1999) entwickelt. Im Zentrum der Bookmark-Methode steht das sogenannte Ordered-Item-Booklet, das die Items enthält und welches im Folgenden beschrieben wird.

Das Ordered-Item-Booklet. Das gereihte Aufgabenheft *(Ordered-Item-Booklet, OIB)* wurde urspünglich durch die Bookmark-Methode eingeführt (Karantonis & Sireci, 2006; Mitzel et al., 1999). Bei der Bookmark-Methode werden die Items aufsteigend nach den empirisch ermittelten Schwierigkeiten von leicht bis schwierig in einem Aufgabenheft geordnet. Die Itemschwierigkeiten werden durch psychometrische Verfahren der Item-Response-Theorie (IRT) aus den vorhandenen Daten geschätzt (meist durch das Rasch-Modell). Sowohl *Selected-Response-* (SR, z. B. Multiple-Choice-Items) Items als auch *Constructed-Response-* (CR, z. B. offene Antworten mit Punktevergabe) Items werden in einem Ordered-Item-Booklet (OIB) zusammengefasst und den Panel-Teilnehmerinnen und -teilnehmern übergeben. Pro Seite wird ein Item mit der dazugehörigen Schwierigkeit dargestellt. Die Teilnehmer/innen setzen nun unter Berücksichtigung der Schwierigkeiten ein Lesezeichen (Bookmark) an der jeweiligen Stelle, an der sie die Cut-Scores zwischen den unterschiedlichen Niveaustufen vermuten.

Das OIB im Standard-Setting für M4. Für das Standard-Setting in M4 wurden aus dem gesamten Itempool durch ein internes Review 80 Items ausgewählt, die das gesamte Schwierigkeitsspektrum bestmöglich repräsentierten. Aus Zeitgründen können nicht alle verfügbaren Items in den Standard-Setting-Prozess einbezogen werden. Die Items waren im Ordered-Item-Booklet aufsteigend nach Schwierigkeit gereiht, wobei pro Seite nur ein Item gelistet wurde. Jede Seite enthielt den Itemtext (Itemstamm) und dazugehörige Abbildungen, den Itemnamen und die Seitennummer. Zusätzlich zum OIB erhielten die Teilnehmer/innen den Antwortschlüssel zu den einzelnen Items. In einer Online-Kodiersoftware trugen die Rater anschließend die Zuordnung der Items zu den einzelnen Levels ein. Die daraus gewonnenen Daten dienten wiederum als Grundlage für die Diskussionen im Plenum. Insgesamt wurden drei Rating-Runden mit der IDM-Methode (siehe 1.2) für die oberen beiden Cut-Scores und zwei Rating-Runden mit der Bookmark-Methode für den untersten Cut-Score durchgeführt.

Prozess in der Bookmark-Methode. Wie bereits oben erwähnt, arbeiten die Teilnehmer/innen bei der Bookmark-Methode mit einem Ordered-Item-Booklet, in dem die Items der Schwierigkeit nach aufsteigend gereiht sind. Die Frage, die an die Teilnehmer/innen gestellt wird, lautet (cf. Cizek & Bunch, 2007): „Ist es wahrscheinlich, dass ein/e minimalqualizierte/r Schüler/in bzw. eine Testperson an der Grenze zwischen den Levels X und Y dieses Item richtig beantworten wird?“

Der Term *„wahrscheinlich“* wird meist mit einer 2/3- oder 67 %-Wahrscheinlichkeit, das Item zu lösen, festgelegt (*Response Probability, RP* = .67). Der/die Teilnehmer/in erhält somit die Aufgabe jedes Item zu begutachten und sich die Frage zu stellen, ob ein/e minimalqualizierte/r Schüler/in in 2 von 3 Fällen die Aufgabe richtig beantworten würde. Kommt der/die Teilnehmer/in zu einem Item, bei dem die Wahrscheinlichkeit unter 2/3 fallen würde, setzt er/sie dort eine Marke. Demnach könnten minimalqualizierte Testpersonen alle Items bis zu dieser Bookmark lösen (mit

einer 2/3-Wahrscheinlichkeit)[1]. Hier bleibt zu entscheiden, welche *Response Probability* man festlegt, da diese Auswirkungen auf die Cut-Scores hat (Wyse, 2011). Das Rasch-Modell geht von einer 50%igen Lösungswahrscheinlichkeit aus, was bedeutet, dass, wenn die Personenfähigkeit gleich der Itemschwierigkeit ist, diese Schüler/innen das Item mit einer 50%igen Wahrscheinlichkeit lösen können (Wang, 2003).

Basierend auf der Rasch-Gleichung, in der sich die Lösungswahrscheinlichkeit *p(x = 1)* aus der Itemschwierigkeit β_j und der Personenfähigkeit θ_i zusammensetzt, setzt man *p = 2=3* und löst nach β auf. Man erhält aus der ursprünglichen Gleichung

$$p(x = 1|\theta_i, \beta_j) = \exp(\theta_i - \beta_j)/[1 + \exp(\theta_i - \beta_j)] \quad (1)$$

die Form

$$\theta_i = \beta_j + .708. \quad (2)$$

Um die Personenfähigkeit θ zu ermitteln, die nötig ist, um ein Item mit einer 2/3-Wahrscheinlichkeit zu lösen, muss man der Itemschwierigkeit die Konstante von .708 hinzuaddieren. Im Rasch-Modell würde sich die Reihung der Items nicht ändern, egal, ob man die Schwierigkeit oder die Personenfähigkeit als Wert für die OIB-Generierung verwendet. MacCann und Stanley (2006) verwenden daher anstelle von θ den Begriff der *Bookmark Difficulty Location* (BDL). Im 2PL oder 3PL kann es allerdings durchaus zu einer Änderung der Reihenfolge kommen.

Die Itemschwierigkeiten wurden für M4 durch das Rasch-Modell (Rasch, 1960) ermittelt, wobei die Lösungswahrscheinlichkeit auf 67 % gesetzt wurde. Für das bessere allgemeine Verständnis wurde der Mittelwert der Item- und Personenparameterverteilung für das Standard-Setting auf 500 gesetzt. Dieser ist aufgrund von internationalen Schülerleistungsstudien vertraut.

Nachdem die Teilnehmer/innen ihre Markierungen (Bookmarks) gesetzt haben, wird die jeweilige Seite mit dem dazugehörigen Fähigkeitswert (Theta) notiert. Dieser Theta-Wert stellt nun den Cut-Score dar und kann wieder in einen Rohwert der entsprechenden Test-Skala transformiert werden. Die individuellen Cut-Scores der Teilnehmer/innen können mittels Median oder Mittelwert zu einem Gesamt-Cut-Score zusammengefasst werden.

Vorteile der Bookmark-Methode sind, dass die tatsächlich von den Schülerinnen und Schülern bearbeiteten Items mit den dazugehörigen Test-Scores in den Entscheidungsprozess einfließen und die Methode sehr einfach in der Durchführung ist. Der Nachteil besteht darin, dass oft eine hohe Differenz der Item-Schwierigkeiten zwischen benachbarten Items bestehen kann. Die Cut-Score-Bestimmung ist dann schwierig (diskutiert in Cizek & Bunch, 2007). Eine weitere Kritik besteht in dem Konzept des minimalqualifizierten Schülers, das manche Teilnehmer/innen sehr schwierig finden können, weshalb eine genaue Instruktion und die Diskussion des Konzepts während des Workshops wichtig ist (Cizek & Bunch, 2007).

Karantonis und Sireci (2006) weisen noch auf einige bedeutsame Kritikpunkte hin, die einer genaueren Untersuchung bedürfen:

- Im OIB kann eine *Item-Disordinalität* auftreten, die dem Prozess nicht dienlich ist.
- Es konnte gezeigt werden, dass die Bookmark-Methode im Vergleich zu anderen Methoden und zu simulierten Daten die Cut-Scores meist etwas unterschätzt (negativer Basis).

1 Hinsichtlich der Festlegung der RP herrscht kein klarer Konsens, allerdings scheinen Personen mit dem Term 2 von 3 besser umgehen zu können.

- Generell scheinen Panelisten die Anforderungen in der Bookmark-Methode zu verstehen, das Ausmaß der kognitiven Komplexität und inwiefern die Urteile tatsächlich valide sind, ist allerdings unklar.
- Eine weitere Frage ist, ob der Mittelwert oder Median der individuellen Bookmarks für die Cut-Score-Berechnung verwendet werden sollte. Der Median ist zwar unabhängig von Ausreißern, allerdings könnten solche Ausreißer in Form von Extremmeinungen bezüglich der Position des Bookmarks auch eine wichtige Bedeutung für den Prozess haben.

1.2 Item-Descriptor-Matching-Methode (IDM)

Aufgrund der genannten Nachteile der Bookmark-Methode wurde beim Standard-Setting in M4 daher die Item-Descriptor-Matching-Methode für die beiden oberen Cut-Scores (zwischen Level 1 und Level 2 und zwischen Level 2 und Level 3) verwendet. Diese nutzt ebenfalls ein OIB und das Konzept der Response Probability ist für die Reihung der Items notwendig, geht allerdings nicht in den Entscheidungsprozess ein. Dadurch sollte zumindest der kognitive Aufwand für die Teilnehmer/innen etwas verringert werden.

Die IDM-Methode wurde aus der Motivation heraus entwickelt, eine bessere Verlinkung zwischen den PLDs (*Performance Level Descriptors* = Kompetenzstufenbeschreibungen) und den Cut-Scores zu gewährleisten, was wiederum die Validität der Ergebnisse erhöht (Cizek & Bunch, 2007). Die Methode verwendet ebenfalls ein Ordered-Item-Booklet und die einzelnen Testitems werden den einzelnen PLDs zugeordnet (Ferrara et al., 2002).

Die Frage, die an das Experten-Panel gestellt wird, ist: „Welcher PLD repräsentiert am besten die Anforderungen des Items?“ Oder genauer: **„Welcher PLD drückt am besten das Wissen, die verlangte Fähigkeit und kognitiven Prozesse aus, die zur Beantwortung des bestimmten Items gefordert sind?“** Die Teilnehmer ordnen danach jedes Item einem bestimmten PLD zu und vermerken dies auf dem Antwortbogen oder in einer entsprechenden Software. Der Schwellenwert, der zwei Kompetenzstufen voneinander trennt, wird dort gesetzt, wo der/die Teilnehmer/in kontinuierlich und systematisch von einem Level ins nächste wechselt. Dies spricht für eine sehr flexible Methode, die nicht von einer strengen Sequenzierung (wie bei der Bookmark-Methode) ausgeht und auch etwas Rauschen zulässt. Da die Schwierigkeiten der Items meist durch Schätzungen basierend auf der IRT erfolgen, kann nicht davon ausgegangen werden, dass die Item-Positionen im Booklet unveränderlich sind, sondern auch einem Schätzfehler unterliegen; eine erlaubte Flexibilität entspricht also einem natürlicheren Matching-Prozess (Cizek & Bunch, 2007; Ferrara et al., 2002). In Regionen alternierender Item-PLD-Matches wird der Threshold-Bereich festgelegt (Ferrara et al., 2002). Da es auch in den PLDs keine absolut festsetzbaren Grenzen gibt, sondern hier die Übergänge eines PLDs zum nächsten fließend sind, wird dieser Bereich als optimal zur Schwellenwertbestimmung angesehen. Mindestens drei aufeinanderfolgende gleiche Klassifizierungen müssen vorliegen, um den Anfang und das Ende eines Grenzbereichs zu definieren. In diesem Bereich wird der Cut-Score ermittelt. Dies kann ähnlich wie bei einer Bookmark-Methode geschehen, indem man die Teilnehmer/innen nochmals entscheiden lässt, wo genau sich in dieser Region der exakte Übergang zwischen den Kompetenzstufen befindet. Genauer kann man es mittels Median oder Mittelwertberechnung erfassen. Im Falle der Mittelwertbestimmung werden nur die Schwierigkeiten der jeweiligen Grenz-Items verwendet (N = 2). Es gibt auch Ansätze, in denen der Schwellenwert mittels logistischer Regression bestimmt wird (Sireci & Clauser, 2001).

Die Identifizierung der Übergangsbereiche, die in der IDM zur Bestimmung der Cut-Scores definiert sind, ist praktisch allerdings oft sehr schwierig umzusetzen. Bei größeren Item-Mengen können auch Ausreißer auftreten, die laut Original-Methode bereits den Beginn oder das Ende eines Grenzbereichs festlegen würden. Für das Standard-Setting in M4 wurde daher die ursprüngliche Methode leicht modifiziert, wie weiter unten (siehe 7) nachzulesen ist.

Die IDM wird grundsätzlich in mehreren Runden durchgeführt, wobei in Runde 1 die Items den PLDs zugeordnet werden: Danach werden die Schwellenwert-Regionen durch die Organisatoren bzw. die Psychometriker/innen des Standard-Settings ermittelt und rückgemeldet. Diese werden dann im Plenum oder in Subgruppen diskutiert. In Runde 2 wird derselbe Prozess nochmals durchgeführt, Änderungen können vorgenommen werden und ein erster Cut-Score wird berechnet. In Runde 3 werden die Werte diskutiert und es werden den Teilnehmerinnen und Teilnehmern zusätzlich Informationen über die Konsequenzen, Mittelwerte, Verteilungen usw. vermittelt. Der endgültige Cut-Score wird danach festgelegt und nochmals zur Begutachtung präsentiert. Zusätzlich könnte die IDM noch durch eine Item-Map ergänzt werden, da eine solche auch Item-Untergruppen besser darstellt (Schulz, Kolen & Nicewander, 1999; Schulz, Lee & Mullen, 2005).

Die sogenannten *Threshold Regions* (TR) sind Bereiche, in denen der Match zwischen Item-Anforderung (Wissen, Fähigkeit etc.) und die Anforderungen des Descriptors (PLDs) nicht klar sind. Dies kann mehrere Gründe haben und die Teilnehmer/innen müssen darauf sensibilisiert und trainiert werden. Gründe können sein:

- Item Ordering Effects (inkl. methodische Aspekte der OIB-Generierung)
- Unklarheit in Beschreibung der PLDs
- Unsicherheit der Teilnehmer/innen bzgl. Zuordnung

Der wesentliche Vorteil der Methode liegt darin, dass der kognitive Anspruch an die Teilnehmer/ innen gering gehalten wird (Ferrara et al., 2002). Die Items müssen lediglich den PLDs zugeordnet werden, es bedarf keiner zusätzlichen Instruktion, wie z. B. sich eine bestimmte Schülergruppe vorzustellen, die einer gewissen Mindestanforderung entspricht. Da Personen generell Probleme haben, Urteile auf Grund von Wahrscheinlichkeitsangaben zu machen (Impara & Plake, 1998; Plous, 1993), bietet diese Methode auch den Vorteil, dass Antwortwahrscheinlichkeiten zwar in die Generierung des OIB miteinfließen, für den Entscheidungsprozess allerdings irrelevant sind (im Gegensatz zur Bookmark-Methode).

2 Die Expertengruppe

Die insgesamt 14 Teilnehmer/innen setzten sich aus unterschiedlichen Teilgruppen (siehe Abb. 2) zusammen, die ein bestimmtes Spektrum repräsentierten. Die direkte Auswahl geschah durch das BIFIE in Zusammenarbeit mit den verschiedenen Institutionen und Behörden. Unter den Teilnehmerinnen und Teilnehmern befanden sich Vertreter/innen der Fachdidaktik, des Bundesministeriums für Unterricht, Kunst und Kultur (BMUKK), praktizierende Lehrer/innen für M4 und M8 sowie Personen aus Forschungseinrichtungen (BIFIE, Universitäten etc.).

Laut Einführungsfragebogen waren zum Zeitpunkt des Standard-Settings 81 % der Teilnehmer/innen[2] mit dem Prozess des Setzens von Standards und den Kompetenzstufen der Bildungsstandards für Mathematik auf der 4. Schulstufe vertraut. Alle Teilnehmer/innen stimmten zu, dass die Gruppenzusammensetzung für das Standard-Setting passend war.

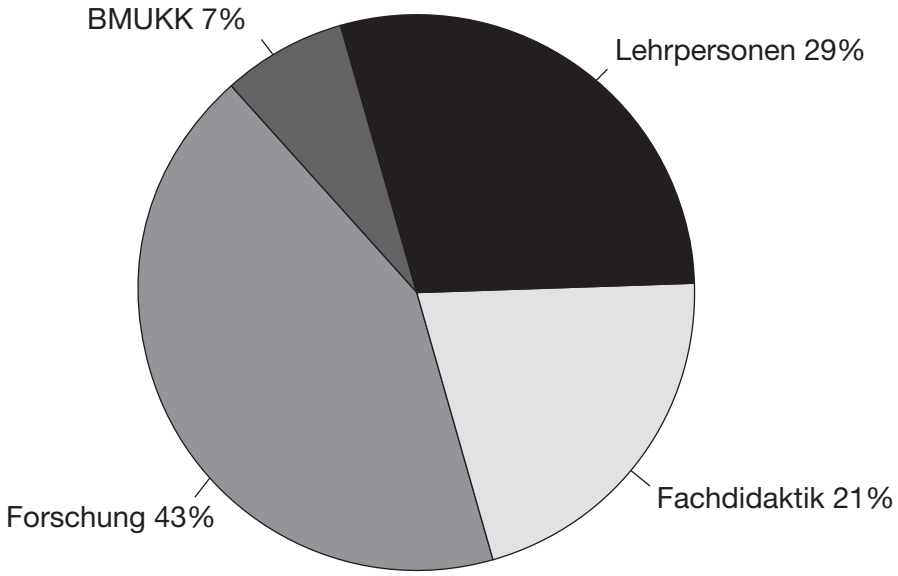

Abbildung 2: Zusammensetzung der Teilnehmer/innen

2 Rücklaufquote der Fragebögen 78 %

3 Training und Vorbereitung auf den Beurteilungsprozess

Die Teilnehmer/innen müssen eine umfassende Schulung erhalten, damit sie mit dem Material, der Methode und dem Ablauf vertraut sind. Es ist von enormer Wichtigkeit, dass die Aufgaben verstanden werden. Cizek und Bunch (2007) geben einen kleinen Leitfaden, an dem der Ablauf des Standard-Settings für M4 ausgerichtet wurde.

Am ersten Tag des Workshops wurden die Teilnehmer/innen eingegeschult. Nach einer ausführlichen Einführung in die Bildungsstandardüberprüfung sowie zum bisherigen Verlauf des Standard-Setting-Prozesses (Phase I und II) bekamen die Experten einen Übungstest mit 10 Items vorgelegt. Dadurch sollte ihnen die Testsituation vermittelt werden und ihnen zeitlicher Druck, der in die tatsächliche Bearbeitung der Items miteinfließt, bewusst gemacht werden. Danach folgte eine Einführung in die Standard-Setting-Methode und den Ratingprozess. Nach genauerer Erläuterung der Kompetenzstufenbeschreibungen folgte eine kurze Diskussion in Kleingruppen, in denen die Teilnehmer/innen auf Unterschiede zwischen den Stufenbeschreibungen achten und Unklarheiten bezüglich Begrifflichkeiten klären konnten. Anschließend wurde im Plenum nochmals über kritische Punkte diskutiert und erste Ratings anhand von einigen Items in der Gesamtgruppe vorgenommen. Darauf folgten die weiteren Runden.

4 Runde 1

4.1 Aufgabe und Instruktion

In Runde 1 wurden die Experten/innen aufgefordert, die Items den Kompetenzstufenbeschreibungen (= PLDs) zuzuordnen. Die genaue Instruktion lautete: Beantworten Sie folgende Fragen: **Welche Kompetenzanforderung stellt das Item an die Schüler/innen? Welche Kompetenzstufenbeschreibung drückt das am besten aus?** Die Teilnehmer/innen wurden aufgefordert, das OIB individuell durchzuarbeiten und in die Kodier-Software einzutragen.

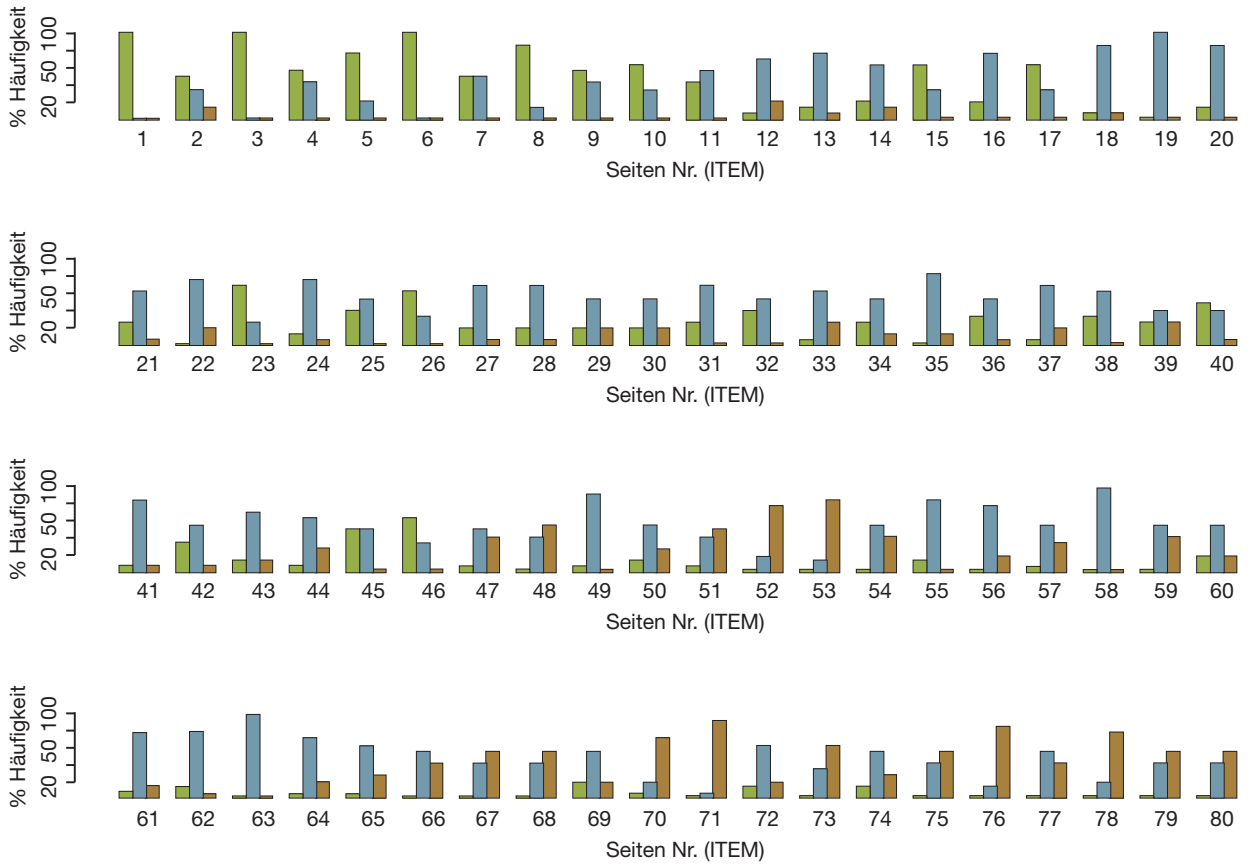

Abbildung 3: Rating-Daten aus Runde 1, die an die Teilnehmer/innen rückgemeldet wurden. Pro Item wird die prozentuelle Häufigkeit der Zuordnung zu einem bestimmten Level dargestellt. Die Teilnehmer/innen können dadurch Items mit niedriger oder hoher Übereinstimmung erkennnen und über diese Items diskutieren (GRÜN = Level 1, BLAU = Level 2, ORANGE = Level 3)

4.1.1 Auswertung der Ratingdaten

Aus der Software erhält man eine Datenmatrix mit Panelisten x Items mit den Werten 1, 2 und 3 (Level-Ratings 1–3). Zur Auswertung wurde für jedes Item separat die prozentuelle Häufigkeit der einzelnen Kategorien ermittelt und grafisch aufbereitet (siehe Abb. 3). Dieses Datenblatt diente als Diskussionsgrundlage. Diskussionspunkte waren Items mit hoher Konvergenz bzw. Divergenz, augenscheinliche Übergänge zwischen Levels sowie Abschnitte, die sich bereits als einzelne Levels herauskristallisierten. Zusätzlich erhielt jede/r Teilnehmer/in eine Auflistung seiner/ihrer individuellen Ratings.

5 Runde 2

5.1 Aufgabe und Instruktion

Die Diskussion in Runde 1 diente dazu, die Expertengruppe in bestimmten Diskussionspunkten, die für den Entscheidungsprozess wichtig sind, zu *homogenisieren*. In Runde 2 arbeiteten die Experten/innen erneut individuell das OIB durch und adjustierten ihre Item-PLD-Zuordnungen, und sie vermerkten ihre Urteile wiederum in der entsprechenden Software. Als Feedback zur Runde 2 wurden den Expertinnen und Experten die Ratingdaten vorgelegt, die in ähnlicher Weise wie in Runde 1 diskutiert wurden. Zusätzlich wurden hier auch die Cut-Scores ermittelt und den Teilnehmerinnen und Teilnehmern rückgemeldet.

5.2 Bestimmung der Cut-Scores

Die Bestimmung der Cut-Scores erfolgt in mehreren Analyseschritten. Wie bereits erwähnt, wurde im Standard-Setting für M4 eine alternative Strategie zur Auswertung des Ratingverhaltens und der damit verbundenen Cut-Score-Bestimmung verwendet. Ziel der Methode ist es, Übergänge zwischen den einzelnen Levels zu detektieren, was in drei Schritten vorgenommen wurde:

1. Als erster Schritt wird jede individuelle Ratingserie durch einen symmetrischen Moving Average geglättet (*order* = 1, Filterfenster ergibt sich aus 2 * order + 1, ungewichtet). Um in den Randbereichen keinen Datenverlust durch die Filterung zu erleiden, wurden die mittleren Ratingwerte dem Beginn und Ende der Serie angefügt. Abbildung 4 zeigt die Rating-Serie (*series*, obere Graphik) einer Person und die gefilterte Funktion dieser Serie darunter. Die individuelle Ratingserie besteht aus 80 Werten (pro Item ein Wert). Die Itemnummer entspricht exakt der Seitenzahl im OIB, die Items sind nach Schwierigkeit geordnet.
2. Die geglättete Funktion jedes Panelisten steigt mit zunehmender Kategorienzahl an. Es wurden zwei Schwellen definiert, die jeweils erstenWerte, die diese Schwellen überschreiten, liefern den Seiten-Index für den jeweiligen Cut-Score. Die dazugehörige Schwierigkeit des Items auf der jeweiligen Schwelle definiert des Weiteren den Cut-Score auf der Theta-Metrik. Die Schwellenwerte wurden auf 1.7 für den ersten Cut und auf 2.4 für den zweiten Cut gesetzt. DieseWerte ergaben sich aus zusätzlich in einem Probelauf erhaltenen Daten[3].
3. Nach anschließender manueller Kontrolle erhält man pro Teilnehmer/in Index-Werte mit Angabe der Seitenzahl des *Cut-Score-Items* sowie die dazugehörigen Theta-Werte. Um einen Gruppen-Wert für die jeweiligen Cut-Scores zu erhalten, wurde der Mittelwert über alle individuellen Cut-Scores berechnet.

Die Methode erbrachte bei allen Teilnehmerinnen und Teilnehmern reliable Werte der Übergange zwischen den Levels.

Zur Rückmeldung an die Teilnehmer/innen wurde eine Tabelle präsentiert, in der die Cut-Scores mit dazugehöriger OIB-Seitennummer dargestellt wurde (Abb. 5). So konnten sich die Teilnehmer/innen ein erstes Bild von den Cut-Scores machen.

3 Überschreitet wie in Abbildung 4 die geglättete Funktion (*filtered*, mittlere Graphik) den ersten Schwellenwert von 1.7, definiert dieser Punkt den Index für die Seite im OIB. In diesem Fall Seite 20. Dieses Item mit der entsprechenden Schwierigkeit (auf Theta-Metrik) liefert den ersten Cut-Score für diese/n Experten/in. Analog verfährt man mit dem zweiten Cut-Score.

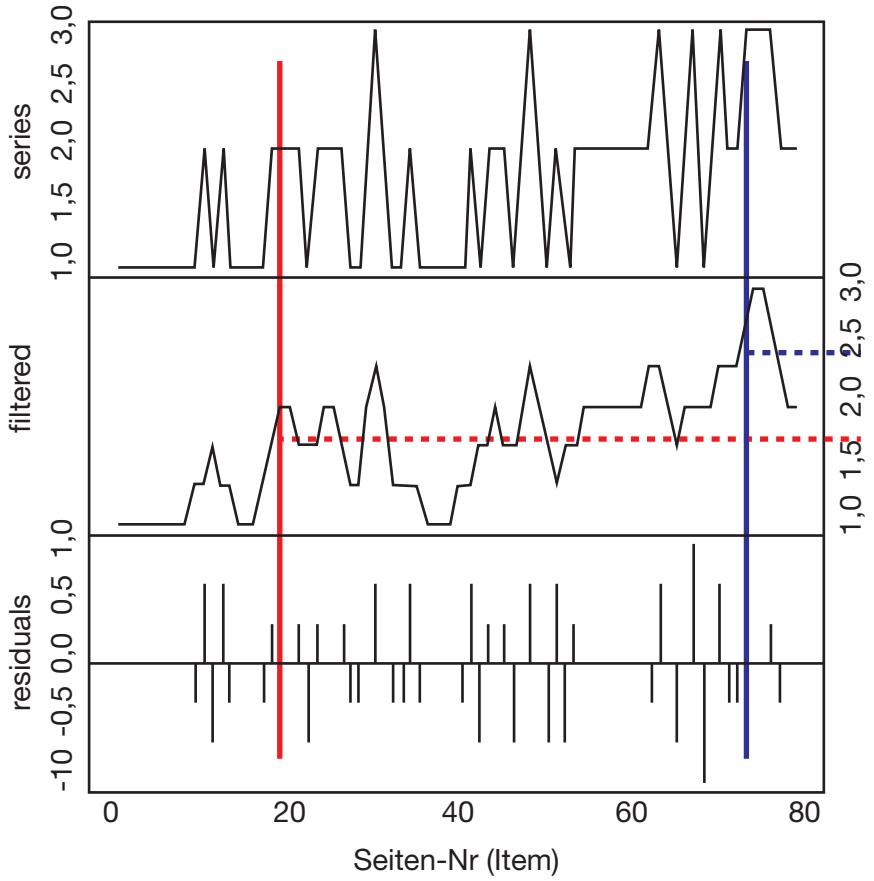

Abbildung 4: Methodik der Cut-Score-Bestimmung. Die oberste Grafik zeigt eine einzelne beispielhafte Ratingserie einer Person. Darunter ist die geglättete Funktion dieser Serie (unten: Filter-Residuen). Gestrichelte horizontale Linien zeigen die beiden Schwellenwerte bei 1.7 und 2.4. Vertikale Linien stellen die Schnittpunkte der geglätteten Funktion mit den Schwellenwerten dar. Aus diesen Punkten kann man auf der X-Achse die Seitennummer des Items ablesen, das den Cut-Score repräsentiert.

	Seite-Cut1	Seite-Cut2	Diff-Cut1	Diff-Cut2
Mean	11	46,29	444,2	618,82
SD	3,88	9,91	36,54	47,62
SE	1,04	2,65	9,77	12,73
	[1,444)		[444,619)	[619,900)
Anz. Items pro Level	10		36	34
MW ItemDiff pro Level	384,53		539,76	702,34
SD ItemDiff pro Level	47,47		48,4	52,06

Abbildung 5: Feedback in Runde 2: Deskriptive Statistiken zu den Cut-Scores sowie Anzahl der Items pro Level.

6 Runde 3

Nach der Diskussion zu Runde 2 wurden die Teilnehmer/innen gebeten, das OIB ein letztes Mal durchzuarbeiten, die Zuordnungen zu adjustieren und sich auf endgültige Urteile festzulegen. Dann wurden Rückmelde- und Konsequenzdaten präsentiert, danach folgte eine abschließende Diskussion über die Setzung der Cut-Scores.

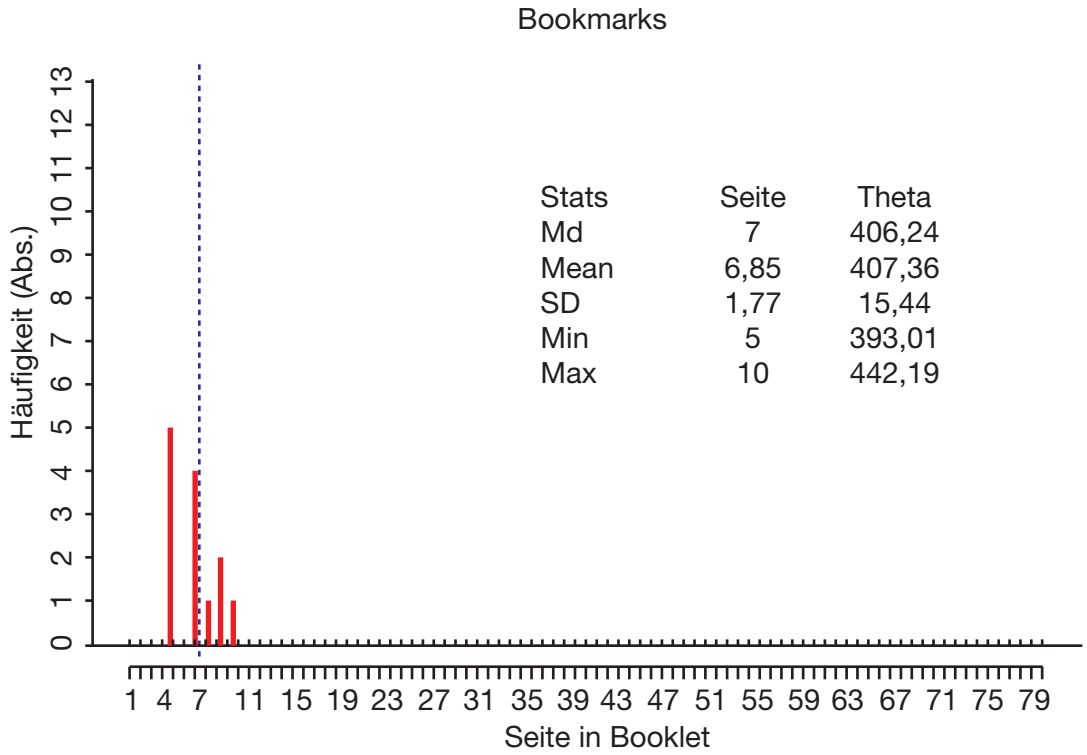

Abbildung 6: Feedback in Runde 4 und 5: Deskriptive Statistiken zum Bookmark-Cut-Score sowie Häugkeit gewählter Bookmarks und Median (blaue gestrichelte Linie).

7 Setzung der Schwelle zu Unter Level 1 – Runden 4 und 5

Nach einer endgültigen Entscheidung über die Cut-Scores zu Level 1–2 und Level 2–3 wurde abschließend noch die Grenze zu *Unter Level 1* bestimmt. Dazu wurde die Bookmark-Methode (siehe Abschnitt 1.1) verwendet. Die Teilnehmer/innen mussten sich, beginnend beim ersten Item des OIBs folgende Frage stellen:

Könnte ein/e minimalqualizierte/r Schüler/in bzw. eine Testperson an der Grenze zwischen dem untersten Level und Level 1 das jeweilige Item in 2 von 3 Fällen beantworten? Falls die Frage mit Ja beantwortet wurde, gingen die Teilnehmer/innen zum nächsten Item über, war die Antwort Nein, wurde hier ein *Bookmark* (Lesezeichen) gesetzt, welches den Cut-Score zwischen den Levels repräsentiert.[4]

Nachdem die Teilnehmer/innen ihre Markierungen gesetzt haben, wird die jeweilige Seite mit dem dazugehörigen Fähigkeitswert (Theta) notiert. Dieser Theta-Wert ist nun der Cut-Score und kann wieder in einen Rohwert der entsprechenden Test-Skala transformiert werden. Die individuellen Cut-Scores der Teilnehmer/innen können nun mittels Mittelwert oder Median zu einem Gesamt-Score zusammengefasst werden.

Rückmeldung Bookmark-Methode. Den Teilnehmerinnen und Teilnehmern wurden deskriptive Statistiken zum Cut-Score präsentiert sowie in einer Grafik die Häufigkeiten, mit denen bestimmte Seiten als Bookmark gewählt wurden und der Median (siehe Abb. 6). Aufgrund dieser Informationen konnte über die Items auf den gewählten Seiten diskutiert werden.

Finale Runde 5. Nach der Diskussion des Feedbacks (siehe Abb. 6) setzten die Teilnehmer/innen ihre finalen Bookmarks. Abschließend wurde ihnen erneut das Feedback für eine abschließende Diskussion über das Setzen des unteren Cut-Scores präsentiert.

4 Dabei wurde den Teilnehmerinnen und Teilnehmern erklärt, darauf zu achten, den Bookmark nicht an einem Ausreißer-Item festzusetzen, sondern stattdessen auch die nächsten folgenden Items mit in die Entscheidung einzubeziehen.

Teil II – Validität und Post-Standard-Setting

8 Prozessevaluation und Evaluation der Cut-Score-Urteile

Es ist von großer Bedeutung, am Ende wichtiger Entscheidungsrunden interne Evaluationen durchzuführen (Hambleton, 2001). Mit diesen soll geklärt werden, ob die Teilnehmer/innen alles verstanden haben, ob es Verbesserungsvorschläge für die Vorgehensweise gibt und wie einig man sich bei den Ergebnissen ist (Raymond & Reid, 2001).

Für Cizek, Bunch und Koons (2004) besteht die Evaluation aus mehreren Teilen: Nach einer ersten Orientierung wird der Grad des Bereitseins der Experten erhoben (Training, Aufgabenverständnis, Überzeugung gegenüber der Methode). Danach folgt eine Evaluation über das Ergebnis des Standard-Settings (Pitoniak, 2003). Für das M4-Standard-Setting wurden ein Eingangsfragebogen und ein Abschlussfragebogen verwendet sowie ein Fragebogen nach jeder Runde.

Aus der Evaluation durch die Experten/innen konnte ebenfalls ein positives Bild des Standard-Setting-Prozesses hinsichtlich Methodik, Durchführung und Organisation gezeichnet werden. Alle Teilnehmer/innen[5] gaben an, dass sie von ihren Empfehlungen zur Schwellenwertsetzung überzeugt wären und sie die ermittelten Cut-Scores als verlässlich einstufen würden. Die Teilnehmer/innen gaben großteils an (ca. 92 %), dass die Cut-Scores von Politik, Lehrerinnen und Lehrern, der Bevölkerung und Abnehmerinnen und Abnehmern aus der Wirtschaft als verlässlich akzeptiert werden würden. Viele Personen waren ebenfalls der Meinung, dass die Verteilung, die sich aus den Konsequenzdaten ergab, ein sehr gutes Abbild aus der praktischen Erfahrung widerspiegelt.

5 Rücklaufquote der Fragebögen 85 %

9 Rating-Verhalten

Um Aufschluss über das Rating-Verhalten zu bekommen, wurde in Runde 3 für jedes Item der Modalwert[6] berechnet. Jede individuelle Ratingserie eines Raters wurde anschließend mit der Reihe an Modalwerten korreliert. Wie Abbildung 7 zeigt, sind die Korrelationen generell hoch. Allerdings sind bei zwei Ratern (R16 und R19) die Korrelationen niedriger als bei den anderen – diese Rater zeigten auch bei anderen Maßen der Übereinstimmung Auffälligkeiten (siehe 9.1) und wurden daher von der Analyse zur Berechnung der Cut-Scores ausgeschlossen.

9.1 Interrater-Reliabilität

Für jeden Rater wurde ein mittleres Kappa, also die mittlere Übereinstimmung mit allen anderen Ratern, sowie die dazugehörige Standardabweichung berechnet (Abb. 8)[7]. Ein niedriger Mittelwert zeigt hier eine geringe Übereinstimmung des Raters mit allen anderen Ratern an. Ein niedriger Mittelwert und eine niedrige Standardabweichung würde ein konsistent abweichendes Rating-Verhalten bedeuten, d. h., der oder die Teilnehmer/in würde konsistent von der Gruppen-Meinung abweichen. Für die Beurteilung des Verhaltens der Rater folgte die Orientierung an den Richtlinien von Landis und Koch (1977), die zwischen $0.41 < \kappa < 0.60$ von einer moderaten Übereinstimmung sprechen, die in diesem Standard-Setting angestrebt wurde. Wenn der mittlere Kappa-Koeffizient also unter 0.41 lag, wurde der entsprechende Rater von der Analyse ausgeschlossen. Dies betraf zwei Rater, die auch bereits in einem anderen Maß des Rating-Verhaltens (Abb. 7) Auffälligkeiten zeigten. Für die Berechnung der Cut-Scores (sowohl bei der IDM- als auch bei der Bookmark-Methode) wurden diese Rater ausgeschlossen.

Als weitere Analyse zur Übereinstimmung der Raterurteile wurde der von Fleiss vorgeschlagene Kappa-Koeffizient für die dritte, finale IDM-Runde berechnet (Fleiss, 1971). Fleiss‘ Kappa ist eine Erweiterung zu Cohen‘s Kappa (Cohen, 1960) bei mehr als 2 Raterurteilen, wobei $\kappa = 1$ perfekte Übereinstimmung bedeutet. Für Runde 3 ergab sich für 14 Teilnehmer/innen und 80 Items $\kappa = 0.46$, für 12 Teilnehmer/innen (unter Ausschluss der beiden Rater mit auffälligem Rating-Verhalten) $\kappa = 0.51$. Interpretiert man die Werte nach Landis und Koch (1977), so liegt hier eine *moderate* Übereinstimmung ($0.41 < \kappa < 0.60$) der Expertenurteile vor.

Die Intraklassen-Korrelation *(intraclass correlation coefficient, ICC)* kann auf Basis von verschiedenen Varianzanteilen sowohl zur Bestimmung von Konsens als auch für Konsistenz (siehe Abb. 9)[8] eingesetzt werden. Der ICC beschreibt das Verhältnis der Varianz einer abhängigen Variable (z. B. Ratings) zur Gesamtvarianz. In einem idealen Fall wäre die Varianz in den Ratings ausschließlich auf die Items und nicht auf die unterschiedlichen Rater zurückzuführen, dann würde der ICC einen Wert von 1 (Bartko, 1966; McGraw & Wong, 1996) erreichen.

6 Der am häufigsten vorkommende Wert.
7 Für die Berechnung der Rater-Analysen wurden nur Daten aus der IDM herangezogen, da bei der Bookmark-Methode zu wenige Daten pro Rater vorliegen
8 Rater-Übereinstimmung (oder Rater-Konsens) beschreibt hier die exakte Übereinstimmung der einzelnen Ratings zwischen den Ratern. Rater-Konsistenz hingegen gibt an, inwieweit die Rater bestimmte Objekte (Personen, Items etc.) in eine ähnliche Reihung bringen.

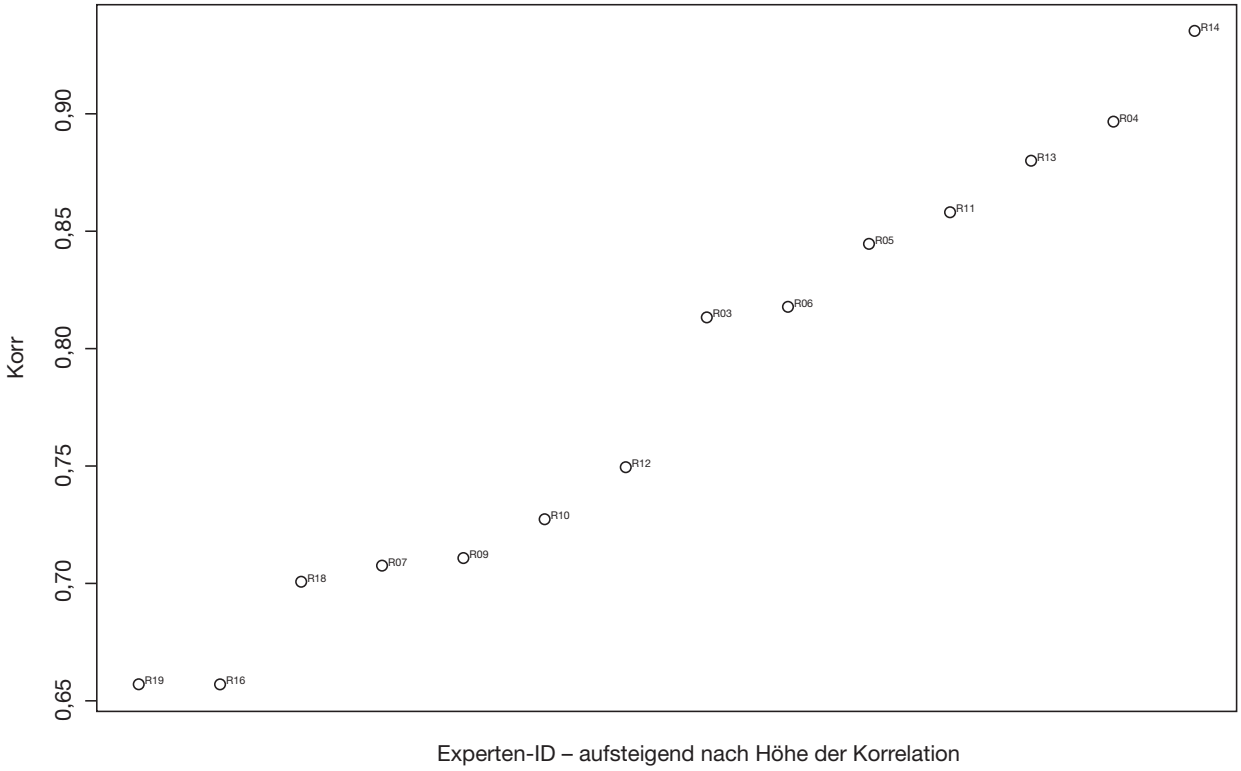

Abbildung 7: Korrelation zwischen Modalwerten der Items und individuellen Ratings der Teilnehmer/innen.

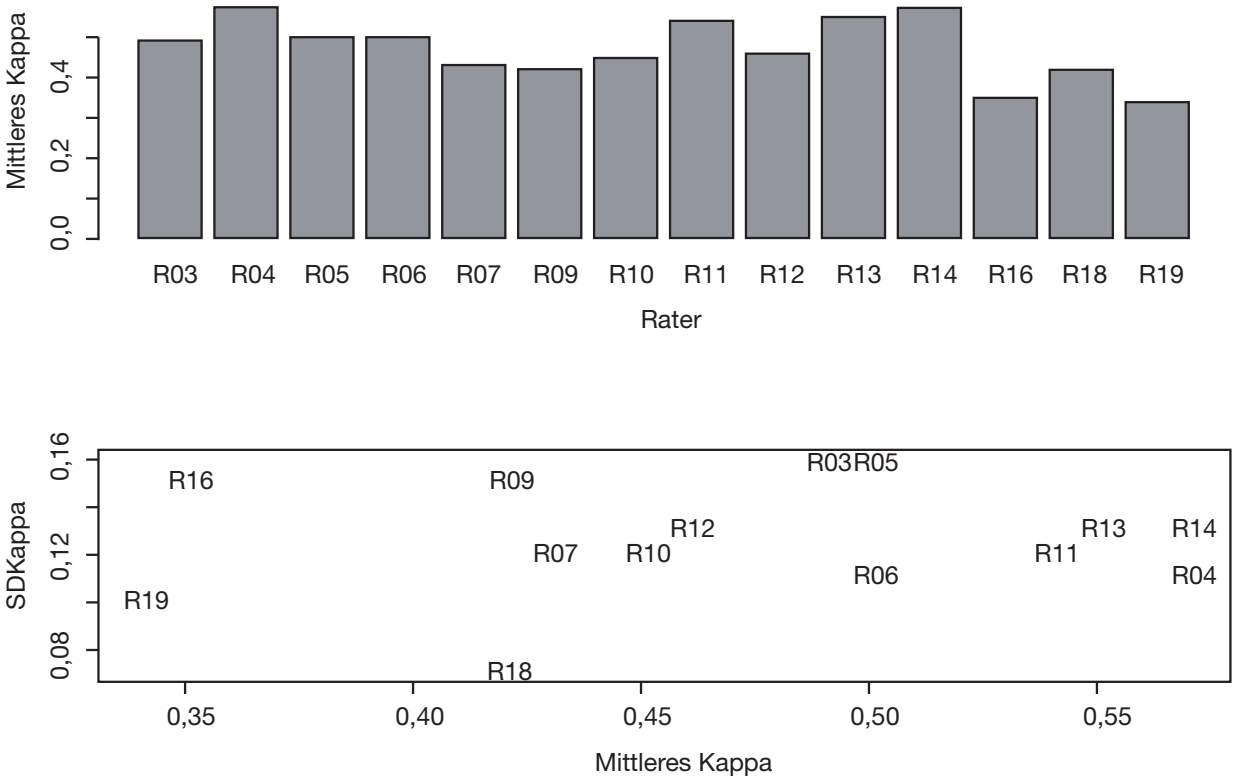

Abbildung 8: Mittleres Kappa und Standardabweichung pro Rater.

		ICC	
	N	Agreement	Consistency
Runde 3	14	0.62	0.67
Runde 3	12	0.65	0.68

Abbildung 9: Analyse zur Übereinstimmung und Konsistenz der Ratings. ICC = Intraclass Correlation Coefficient.

10 Endgültige M4-Cut-Score-Werte aus der IDM- und Bookmark-Methode

Die finale Runde der Cut-Scores mit der IDM-Methode war die Runde 3, dies wurde auch den Teilnehmerinnen und Teilnehmern deutlich signalisiert. Die Cut-Scores wurden daher aus dieser Runde ermittelt, und über ein Jackknife-Verfahren wurde der Standardfehler berechnet. Wie bereits in den Abschnitten 9 und 9.1 beschrieben, wurden für die finale Berechnung der Cut-Scores zwei Teilnehmer/innen ausgeschlossen.

Für die Bestimmung der Cut-Scores wurde der Mittelwert herangezogen. Hier fließen alle Meinungen der Teilnehmer/innen gleichermaßen ein, das heißt, auch extremere Meinungen (die durchaus wichtig sind) gehen in die Cut-Score-Bestimmung mit ein. Durch den Median würde man solche Meinungen verlieren, da dieser ein robustes Maß gegenüber Ausreißern darstellt. Aus diesen Gründen sowie durch Erfahrungen aus früheren Standard-Settings für M8 wurden daher alle Cut-Scores durch eine Mittelwertberechnung bestimmt und der jeweilige Standardfehler errechnet.

Der Cut-Score zwischen Level 1 *Bildungsstandards teilweise erreicht* und Level 2 *Bildungsstandards erreicht* beträgt 458.31 (SEJack = 4.02). Der Cut-Score zwischen Level 2 *Bildungsstandards erreicht* und Level 3 *Bildungsstandards übertroffen* beträgt 651.86 (SEJack = 13.04).

Für den Cut-Score zwischen dem untersten Level *Bildungsstandards nicht erreicht* und *Bildungsstandards teilweise erreicht* wurde die Bookmark-Methode herangezogen. Wie bereits erwähnt, kann für die Berechnung der Cut-Scores entweder der Median oder der Mittelwert herangezogen werden, die beide gleichermaßen angebracht sind. Üblicherweise wird bei der Bookmark-Methode aber der Mittelwert gewählt (Cizek & Bunch, 2007), der Cut-Score für den untersten Level beträgt daher 409.66 (SEJack = 4.25).

Literatur

Bartko, J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19 , 3–11.

Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice, 15* (1), 13–21. doi: 10.1111/j.1745-3992.1996.tb00802.x

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Sage.

Cizek, G. J., Bunch, M. B. & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23* , 31–50.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20* , 37–46.

Ferrara, S., Perie, M. & Johnson, E. (2002, April). *Matching the judgemental task with standard setting panelist expertise: The item-descriptor (id) matching procedure.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76(5)*, 378–382.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Hrsg.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (S. 89–116). New York: Routledge.

Impara, J. & Plake, B. S. (1998). Teacher‘s ability to estimate item difficulty: A test of the assumption in the modified angoff standard setting method. *Journal of Educational Measurement, 35(1)*, 69–81.

Karantonis, A. & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25* (1), 4–12. doi: 10.1111/j.1745-3992.2006.00047.x

Landis, J. & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33* , 159–174.

MacCann, R. G. & Stanley, G. (2006). The use of rasch modeling to improve standard setting. *Practical Assessment, Research and Evaluation, 11(2)*, available online.

McGraw, K. & Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1(1)*, 30–46.

Mitzel, H. C., Lewis, D. M., Green, D. R. & Patz, R. J. (1999). *The bookmark standard setting procedure.* Monterey, CA: McGraw-Hill.

Pitoniak, M. (2003). *Standard setting methods for complex licensure examinations.* Unveröffentlichte Dissertation, University of Massachusetts, Amherst.

Plous, S. (1993). *The psychology of judgement and decision making.* New York: McGraw-Hill.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche.

Raymond, M. R. & Reid, J. B. (2001). Who made thee a judge? selecting and training participants for standard-setting. In G. J. Cizek (Hrsg.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (S. 119–158). New York: Routledge.

Schulz, E. M., Kolen, M. J. & Nicewander, W. A. (1999). A rationale for defining achievement levels using irt-estimated domain scores. *Applied Psychological Measurement, 23* (4), 347–362. doi: 10.1177/01466219922031464

Schulz, E. M., Lee, W.-C. & Mullen, K. (2005). A domain-level approach to describing growth in achievement. *Journal of Educational Measurement, 42* (1), 1–26. doi: 10.1111/j.0022-0655.2005.00002.x

Sireci, S. G. & Clauser, B. E. (2001). Practial issues in setting standards on computerized adaptive tests. In G. J. Cizek (Hrsg.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (S. 355–369). New York: Routledge.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement, 40* , 231–253.

Wyse, A. E. (2011). The similarity of bookmark cut scores with different response probability values. *Educational and Psychological Measurement, 71*, 963–985. doi: 10.1177/0013164410395577