bifie | Bildung

*standards

Die Itemerstellung für Bildungsstandardüberprüfungen

Technische Dokumentation – BIST-Ü

Ursula Itzlinger-Bruneforth
Jörg-Tobias Kuhn

Bundesinstitut bifie

Bundesinstitut
bifie

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

Die Itemerstellung für Bildungsstandardüberprüfungen
Technische Dokumentation – BIST-Ü
BIFIE | Department Bildungsstandards & Internationale Assessments (BISTA),
Salzburg 2015

Inhaltsverzeichnis

Dieses Kapitel beschreibt allgemein die Itemerstellung für die Bildungsstandardüberprüfungen (BIST-Ü), wobei mit „Item“ oder „Testitem“ alle Arten von „Testaufgaben“ gemeint sind, die den Schülerinnen und Schülern bei den Bildungsstandardüberprüfungen in den verschiedenen Fächern vorgelegt werden. Der Ausdruck „Item“ ist „broad enough to allow for a variety of item formats and item classifying categories, yet sufficiently precise to be useful for technical discussion“ (Osterlind, 2001). Fach- und kompetenzspezifische Beschreibungen können unter der jeweiligen Standardüberprüfung gefunden werden. Zunächst wird das Ziel der Itemerstellung vorgestellt, danach kurz auf die getesteten Inhalte eingegangen. Zuletzt werden die Prozessschritte aufgelistet, die bei der Erstellung von Items durchlaufen werden.

Ziel der Itemerstellung

Das Ziel der Itemerstellung für die Bildungsstandardüberprüfungen ist es, einen möglichst großen Pool qualitativ hochwertiger Items zu erstellen, der den jeweiligen Fachbereich bzw. die gesuchte Kompetenz valide abbildet, in einer flächendeckenden Überprüfung einsetzbar ist und reliable und objektive Messergebnisse produziert. Die Verwendung der Testergebnisse muss dabei schon bei der Itemerstellung beachtet werden (Kane, 2006). Die Definition von Validität, auf die sich Kane bezieht und die auch vom BIFIE verwendet wird, lautet: „Validität bezieht sich auf den Grad, in dem Evidenz und Theorie die Interpretationen der Testergebnisse stützen, welche sich aus der vorgesehenen Testverwendung ergeben“ (AERA et al., 2008). Da die Rückmeldung der Testergebnisse in Österreich hauptsächlich darauf abzielt, die Unterrichtsentwicklung zu fördern (BIFIE, 2012), ist eine möglichst breite inhaltliche Abdeckung der Fach- und Kompetenzbereiche unerlässlich.

Die Kosten mangelhaft erstellter Items drücken sich nicht nur in kleineren Itempools aus, sondern auch in möglicher einseitiger Verzerrung gegenüber verschiedenen Gruppen von Testpersonen und in größerer Fehlervarianz (Tarrant & Ware, 2008), was die Konstruktvalidität schmälert (Downing, 2002, 2005). Die Itemwriter produzieren das Material, welches beobachtbare Schulleistungen mit dem zu messenden zugrunde liegenden Konstrukt in Verbindung bringt. Ihre Wichtigkeit kann kaum überbewertet werden, denn ein mangelhafter oder mittelmäßiger Itempool kann auch durch die besten statistischen Testdesigns oder Skalierungsmethoden nicht mehr korrigiert werden (Downing & Haladyna, 1997). Eine fachliche Passung der Items ist Grundvoraussetzung für einen validen Test (Webb, 2006). Neben der fachlichen Passung der Items müssen diese auch für die Schüler/innen der Zielgruppe passen; daher werden als Itemwriter hauptsächlich Lehrpersonen mit mehrjähriger Unterrichtserfahrung gesucht. In wenigen Fällen werden Studierende herangezogen; diese müssen jedenfalls ein fachspezifisches Studium absolvieren und werden beim Reviewprozess von erfahrenen Itemwritern gecoacht.

Durch Schulungen und Materialien für Itemwriter soll gewährleistet werden, dass für alle Bereiche gute Items zur Verfügung stehen. Ein Ziel dabei ist auch, den in der Unterrichtspraxis verankerten Itemwritern den Unterschied zwischen Schulaufgabe und Testitem zu erläutern. Vor allem der Fokus auf eine große, in ihrer Heterogenität unbekannte Schülergruppe – im Gegensatz zur bekannten Klasse – ist für neue Itemwriter erfahrungsgemäß oft schwierig.

Testinhalte

Das BIFIE führt Bildungsstandardüberprüfungen in den Fächern Deutsch und Mathematik auf der 4. und 8. Schulstufe sowie in Englisch auf der 8. Schulstufe durch. Informationen zu den jeweiligen Kompetenzen, Teilkompetenzen und Deskriptoren können im fachspezifischen Teil gefunden werden.

Die österreichischen Standards definieren „grundlegende Kompetenzen, über die die Schüler/innen bis zum Ende der Primar- bzw. Sekundarstufe I in der Regel verfügen sollen“ (BIFIE, 2012). Sie beziehen sich entweder auf ein Kompetenzmodell, das als solches eine Grundlage der Itemerstellung darstellt (Mathematik, 8. Schulstufe: https://www.bifie.at/node/1347; Mathematik, 4. Schulstufe: https://www.bifie.at/node/1346), oder auf mehrere Kompetenzmodelle für Teilkompetenzen eines Fachs (Deutsch, 8. Schulstufe: https://www.bifie.at/node/325; Deutsch, 4. Schulstufe: https://www.bifie.at/node/1345;

Englisch, 8. Schulstufe: https://www.bifie.at/node/1348). Um die Messung der jeweiligen Kompetenzen bzw. Teilkompetenzen zu präzisieren, werden die fachspezifischen Konstrukte in Bereiche bzw. Kompetenzfelder, Deskriptoren oder Knoten „zerlegt". Diese Operationalisierungen, entweder als zielgenaue Knoten eines Kompetenzmodells (Mathematik) oder in Form differenzierter Can-do-Statements (Deutsch bzw. Englisch), dienen als grundlegendes Raster für die Itementwicklung. Eine Orientierung am Lehrplan ist jedoch ebenfalls unerlässlich, da dieser wesentlich detailreicher ist.

Bevor mit der Itementwicklung begonnen werden kann, müssen die zugrundeliegenden Standards hinsichtlich Testbarkeit überprüft werden sowie die administrativen oder technischen Voraussetzungen in Erwägung gezogen werden: Beschränkungen der Testbarkeit ergeben sich sowohl durch Testökonomie und Ausstattung der Schulen, etwa mit Nachschlagewerken und Computern, sowie durch den Inhalt mancher Verordnungen. Obwohl Lehrpläne, Kompetenzmodelle und Bildungsstandards sehr umfassende Dokumentationen darstellen, was den Inhalt der Lehre auf den jeweiligen Schulstufen und Fächern betrifft, muss doch in Betracht gezogen werden, dass diese ursprünglich nicht mit dem Fokus auf eine Überprüfung erstellt wurden, sondern für die Lehre entwickelt wurden. Ein Beispiel für einen Inhalt, der sich einer flächendeckenden Testung widersetzt, wäre z. B. in Deutsch zu finden. Deskriptor 22 (Lesen) der Standards für die 8. Schulstufe besagt „Schüler/innen können gezielt Informationen in unterschiedlichen Medien aufsuchen und beherrschen insbesondere die Internetrecherche und Benützung von Nachschlagewerken" (https://www.bifie.at/node/325). Es ist nicht davon auszugehen, dass allen Schülerinnen und Schülern der 8. Schulstufe Nachschlagewerke während der Bildungsstandardüberprüfung zur Verfügung stehen. Während die Benützung von Nachschlagewerken beim Test jedoch noch simuliert werden kann, verbietet sich die Internetrecherche während eines Tests von selbst – nicht zuletzt, weil aufseiten der Schulen nicht garantiert werden kann, dass genügend funktionierende Geräte zur Verfügung gestellt werden können, sodass alle Schüler/innen dieselben Chancen hätten, diese Testaufgaben zu lösen. Diese Beschränkungen müssen von Fall zu Fall geprüft werden und diese Limitationen in einem Testkonzept dokumentiert werden. Diese Prüfung findet durch das Team Fachdidaktik am BIFIE statt, das bei diesem Prozess durch Kooperation mit Universitäten und Pädagogischen Hochschulen unterstützt wird.

Ein so ausgearbeitetes Testkonzept stellt die Grundlage für die Erstellung eines Itempools dar, mit dem die Kompetenzen valide überprüft werden können. Insgesamt ergibt das in jedem Fach einen Itempool, der mehrere hundert Items pro Kompetenz bzw. Teilkompetenz umfasst. Dieser kann in seiner Gesamtheit nicht den einzelnen Schülerinnen und Schülern vorgelegt werden; auf eine ausgewogene Auswahl pro Schüler/in wird beim Testdesign (bspw. Kiefer und Fellinger, 2015) geachtet.

Der Prozess der Itementwicklung

Der Prozess der Itementwicklung involviert in jedem Fach die Mitarbeit von Personen aus der Unterrichtspraxis, Fachdidaktik und Psychometrie. In einem mehrstufigen Verfahren werden Items erstellt, gereviewt, getestet und analysiert, bevor sie in der BIST-Ü eingesetzt werden können. Dabei werden sowohl fachliche als auch lernpsychologische und psychometrische Eigenschaften eines Items mehrfach untersucht, bevor es als „testreif" bezeichnet werden darf.

Bedarfsanalyse

Vor dem Start einer Runde der Itementwicklung steht eine Bedarfsanalyse. Ausgehend von der geplanten Rückmeldung wird analysiert, wie viele Items mit welchen Eigenschaften benötigt werden. Dabei spielen vor allem inhaltliche Überlegungen eine Rolle, aber auch das zu verwendende Itemformat und die angepeilte Itemschwierigkeit. Ergebnis der Bedarfsanalyse ist ein sogenannter „test blueprint", ein Raster, in dem für jeden Punkt des Anforderungsprofils die Anzahl zu erstellender Items aufgelistet sind. Diese werden dann auf die einzelnen Itemwriter aufgeteilt; die Aufteilung ist für die Itemwriter bindend, um eine valide Abdeckung der Inhalte gewährleisten zu können.

Itemschwierigkeit	Leicht			Mittel			Schwer		
Format / Handlungs- und Inhaltsbereiche	**MC**	**Halb-offen**	**offen**	**MC**	**Halb-offen**	**offen**	**MC**	**Halb-offen**	**offen**
H1: „Darstellen, Modellbilden“ I1: „Zahlen und Maße“	1	2	2	3	2	1	1	2	3
H1: „Darstellen, Modellbilden“ I2: „Variable, funkt. Abhängigkeiten“	3	4	1	2	2	1	4	2	1
H1: „Darstellen, Modellbilden“ I3: „Geometrische Figuren u. Körper“	3	4	0	1	2	2	5	1	1
H1: „Darstellen, Modellbilden“ I4: „Statist. Darstellungen u. Kenngrößen“	8	2	1	6	1	1	2	4	0

Abbildung 1: Beispiel (Ausschnitt aus M8) eines „test blueprints“ (fiktive Zahlen)

In Tabelle in Abb. 1 ist ein stark vereinfachtes Beispiel eines „test blueprints“ für Mathematik, 8. Schulstufe, zu finden. Dargestellt wird hier nur H1 (Handlungsbereich „Darstellen, Modellbilden“). Aus Lesbarkeitsgründen ist die Tabelle gekürzt und vereinfacht. Die angeführten Itemformate (nicht vollständig) sind:

- „MC“ = Multiple-Choice-Format;
- Halboffen: eine Kurzantwort soll gegeben werden (z. B. eine Zahl, wenige Wörter, eine Formel);
- Offen: eine ausführliche Antwort soll gegeben werden (z. B. eine Begründung, ein Rechengang, eine Konstruktion).

Entwicklung

Kim et al. (2010) konnten zeigen, dass zielgruppengerechte Testspezifikationen den Erstellungsprozess deutlich verbessern können und so zu insgesamt höherwertigen Items beitragen. Eine Schulung der Itemwriter, die – neben inhaltlichen Aspekten – auf die Testspezifikationen ebenso eingeht wie auf Aspekte der Itemkonstruktion, ist deshalb von größter Wichtigkeit. Es wurde gezeigt, dass die Kosten pro Item in traditionellen Testkommissionen (Gruppen geschulter Itemersteller/innen) niedriger liegen als bei Adhoc-Gruppen (Case, Holtzman & Ripkey, 2001). Jeder Itemerstellung geht eine mindestens eintägige Schulung voraus, in der die Grundlagen der Itementwicklung mit den Itemwritern besprochen werden. Zusätzlich erhalten die Itemwriter schriftliche Unterlagen, die die fachlichen und psychometrischen Aspekte des jeweiligen Fachbereichs und der Schulstufe abdecken:

1. Zugrundeliegende Verordnung;
2. Lehrplan;
3. Zusätzliche inhaltliche Materialien, kompetenzspezifisch;
4. Itemwriter-Richtlinien mit „Checkliste“ (BIFIE, 2014);
5. „test blueprint“ (Abbildung 1), personenspezifisch.

Die Itemerstellung wird von Fachdidaktikern und Psychometrikern des BIFIE sowie von externen Experten im Prozess begleitet. Die Prozesse laufen nicht in jedem Fach genau gleich ab, eine prototypische Darstellung des Workflows ist in Abbildung 2 dargestellt.

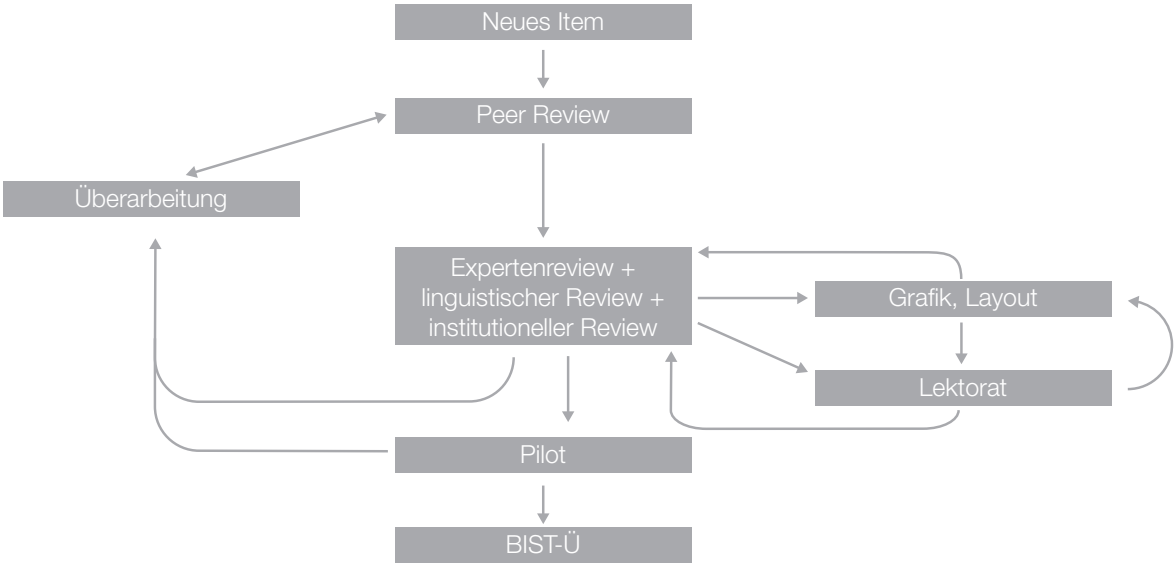

Abbildung 2: Prototypischer Workflow

Nach der Erstellung durch den Itemwriter überprüft dieser anhand der Checkliste, ob das Item allen Anforderungen genügt. Zusammengefasst umfasst diese Prüfung folgende Punkte:

1. Das Item ist valide, d. h. es misst den beabsichtigten Inhalt auf Ebene der Knoten, Operatoren oder Deskriptoren[1];
2. das vereinbarte Format wurde technisch korrekt umgesetzt;
3. das Item liegt im vereinbarten Schwierigkeitsbereich, wobei schwierigkeitsbestimmende Merkmale hauptsächlich inhaltlicher Natur sind;
4. Anforderungen an die Items bezüglich Stil, Inhalt und Antworten wurden umgesetzt;
5. Inhalt und Thema sind dem Alter/der Schulstufe angemessen;
6. die verwendete Terminologie ist korrekt und entspricht der Schulstufe;
7. zur Lösung des Items sind v. a. fachliche Kompetenzen nötig, weniger reines Erinnern oder Allgemeinwissen;
8. das Item bevorzugt keine Schülergruppe;
9. die Items sind voneinander unabhängig, d. h. die Lösung von Item A hilft nicht bei der Beantwortung des Items B;
10. das Item ist meinungsunabhängig, verständlich und hat eine eindeutige Lösung;
11. das Item ist sachrichtig;
12. Grammatik und Rechtschreibung sind korrekt;
13. der Leseumfang ist auf das notwendige Minimum reduziert;
14. die Frage oder Handlungsanweisung ist eindeutig und klar;
15. Antwortoptionen, soweit vorhanden, sind inhaltlich plausibel;
16. der Lösungsschlüssel bzw. die möglichen richtigen Schülerantworten liegen vor;
17. alle Urheberrechte an dem Item liegen beim Itemwriter oder sind klar angeführt, wobei Schulbücher und ähnliche Materialien als Quellen ausgeschlossen sind.

1 Dies wird kritisiert, siehe z. B. Stern (2010). Es ist jedoch eine Anforderung, die sich aus der Rückmeldung ergibt, die zwangsläufig stark standardisiert und klar am Kompetenzmodell ausgerichtet erfolgen muss.

Review & Überarbeitung

Nach der Erstellung durchläuft das Item mehrere Review- und Feedbackrunden. Im Peer Review sind es Mitglieder der Itementwicklungsgruppe, die die Items der anderen Mitglieder kritisch bewerten und ggf. Veränderungsvorschläge machen. Diese Gruppenphase sorgt für eine hohe Homogenität in der Itemerstellung, sowohl in inhaltlicher als auch in qualitativer Hinsicht. Nach dieser Reviewphase kommen viele Items in eine Phase der Überarbeitung, bevor sie in den nächsten Review gesandt werden. Nach dem erfolgreichen Peer Review wird das Item in den Expertenreview bzw. BIFIE-Review geschickt; dieser umfasst sowohl die inhaltlichen Experten sowie die Psychometrie als auch, wo dies zutrifft, linguistische Experten bzw. Native Speakers. Hierbei werden die gleichen Richtlinien angewandt, wie sie für die Itemwritern gelten. Auch in dieser Reviewschleife besteht die Möglichkeit, das Item zur Überarbeitung zurück an den Itemwriter zu senden; erst nachdem diese Reviewschleife positiv durchlaufen ist, wird ein Item vom BIFIE abgenommen. In diesem Prozess werden die Items oft mehrfach überarbeitet, bevor sie abgenommen werden. Aus der Erfahrung der bisherigen Itementwicklung lässt sich sagen, dass die Quote der Items, die nach der Pilotierung verworfen oder zurück in eine Überarbeitungsschleife gesandt werden, bei derart begleiteten Itementwicklungsprozessen wesentlich geringer ist als bei Items, die unbegleitet entwickelt werden.

Nach dieser Abnahme muss das Item ein Lektorat durchlaufen, erst danach gilt ein Text als freigegeben zur Pilotierung. Manchen Items sind Grafiken, Tabellen oder Schaubilder zugeordnet, die grafisch nachbearbeitet werden müssen. Im Layout wird schließlich jedes Item in eine einheitliche Schriftart überführt, wobei auch Schriftgröße, Absätze etc. standardisiert sind, um die Lesearbeit möglichst zu erleichtern.

Itemanalysen nach der Pilotierung

Jedes Item, das den Prozess der Überprüfung erfolgreich durchlaufen hat, wird vor dem Einsatz in der Bildungsstandardüberprüfung in einer Pilotierung getestet. In der Regel wird dabei jedes Item von ca. 200 Schülerinnen und Schülern bearbeitet.

Die Ergebnisse der Pilotierung werden Rasch-skaliert und auf folgende Eigenschaften untersucht:

1. Itempoolebene:
   a. Entspricht die Schwierigkeit des eingesetzten Itempools den Fähigkeiten der getesteten Population?
   b. Verteilen sich die Itemschwierigkeiten gleich auf die Dimensionen (Deskriptoren, Bereiche, Teilkompetenzen)?
   c. Entspricht die eingeschätzte (theoretische) Itemschwierigkeit etwa der empirischen Schwierigkeit?

2. Testheftebene:
   a. Ist die Anzahl der Items der Testzeit angemessen oder gab es zeitliche Probleme?

3. Itemebene:
   a. Ist das Item im akzeptierten Schwierigkeitsbereich? Zu leichte / zu schwierige Items werden ausgeschlossen. Als Ausschlussgrund gilt > 95 % Lösungswahrscheinlichkeit bzw. < 5 % Lösungshäufigkeit.
   b. Ist das Item trennscharf, das bedeutet, unterscheidet es zwischen Schülerinnen und Schülern mit hoher und solchen mit niedriger Kompetenz? Eine Trennschärfe von < .1 ist ein Ausschlussgrund, eine Trennschärfe zwischen 0.1 und 0.2 führt zu einem Markieren des Items – ein solches Item wird bei möglichem Ersatz durch ein anderes Item aus derselben Zelle des „blueprints“ nicht eingesetzt.
   c. Weist einer der Distraktoren eine positive Korrelation mit dem Testergebnis auf? Wenn diese 0.05 übersteigt und der Distraktor von mindestens 10 Schülerinnen und Schülern gewählt wurde, wird das Item nicht eingesetzt.

d. Wird jeder der Distraktoren von den Schülerinnen und Schülern angenommen? Wird ein Distraktor von sehr wenigen Schülerinnen und Schülern gewählt, ist er nicht plausibel und das Item wird ausgeschlossen. < 5 % gilt als Ausschlussgrund, die Regel wird jedoch für sehr leichte Items angepasst (ungefähre Gleichverteilung auf Distraktoren).
e. Für offene Formate: Ist das Item kodierbar? Bei der Kodierung der gegebenen Schülerantworten wird auf die Anzahl der „Problemfälle" sowie auf die Übereinstimmung der Kodierer bei der Bewertung der Aufgaben (Double Coding; bspw. Kiefer und Fellinger, 2015) geachtet. Gestaltet sich die Kodierung als zu aufwändig oder ist die Übereinstimmung zu gering, wird das Item ausgeschlossen.
f. Gibt es viele Schülerantworten, die fehlend oder ungültig sind? Je nach Format kommen verschiedene Entscheidungsregeln zum Tragen, bei offenen Formaten werden höhere Fehlanteile toleriert als bei geschlossenen Formaten. Items mit zu hohen Fehleranteilen oder zu hohen Anteilen invalider Antworten werden ausgeschlossen.
g. Bevorzugt das Item bestimmte Gruppen? Differential-Item-Functioning (DIF; bspw. Kiefer und Fellinger, 2015)-Analysen werden für verschiedene Kriterien durchgeführt: Geschlecht, Erstsprache, Stadt/Land, Schultyp. Ist ein Item in einem Kriterium auffällig und kann ausgeschlossen werden, dass es sich um eine konstruktinhärente Auffälligkeit handelt, wird das Item nicht eingesetzt.
h. Ein Item, das zwar kein Ausschlusskriterium gänzlich erfüllt, aber bei mehreren Kriterien in der Nähe des Ausschlusswerts liegt, wird „geflaggt" und wenn möglich durch ein besseres Item aus derselben Zelle des „blueprints" ersetzt.

Durch diese Auswahl, die sich an internationalen Standards orientiert (vgl. OECD, 2012) und diese durch Erfahrungswerte der durchgeführten Pilotierungen ergänzt, ergibt sich ein Itempool, der sowohl fachdidaktisch als auch testtheoretisch hohen Ansprüchen genügt. Neue Erkenntnisse, die bei Pilotierungen gewonnen werden, fließen in neue Versionen der Materialien für Itemwriter und in die Überarbeitung der Arbeitsprozesse ein, sodass die Qualität der erstellten Items kontinuierlich verbessert werden kann.

Literatur

American Educational Research Association, American Psychological Association & National Council of Measurement in Education (AERA) (Eds.). (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

BIFIE (Hrsg.). (2012). *Bildungsstandards in Österreich: Überprüfung und Rückmeldung.* 4. aktualisierte Auflage. Salzburg: BIFIE. [Online] www.bifie.at/node/560 [29.7.2014.].

BIFIE (Hrsg.). (2014). *Itemwriter-Guidelines.* Unveröffentlichtes Manuskript. Salzburg: BIFIE.

Case, S. M., Holtzman, K., & Ripkey, D. R. (2001). *Developing an item pool for DBT: A practical comparison of three models of item writing.* Academic Medicine, 76, 111–113.

Downing, S. M. (2002). *Construct-irrelevant variance and flawed test questions: Do multiplechoice itemwriting principles make any difference?* Academic Medicine, 77, 103–104.

Downing, S. M. (2005). *The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education.* Advances in Health Sciences Education, 10, 133–143.

Downing, S. M., & Haladyna, T. M. (1997). *Test item development: Validity evidence from quality assurance procedures.* Applied Measurement in Education, 10, 61–82.

Kane, M. (2006). Content-related validity evidence in test developement. In Downing, S. & Haladyna, T. (2006). *Handbook of Test Development (S. 131–153).* New Jersey: Lawrence Erlbaum.

Kiefer, T. & Fellinger, R. (2015). *Pilotierung und Testdesign. Technische Dokumentation – BIST-Ü Mathematik, 8. Schulstufe, 2012.* Salzburg: BIFIE. In Vorbereitung.

Kim, J., Chi, Y., Huensch, A., Jun, H., Li, H. & Roullion, V. (2010). *A case study on an item writing process: Use of test specifications, nature of group dynamics, and individual item writers' characteristics.* Language Assessment Quarterly, 7, 160–174.
OECD (2012). *PISA 2009 Technical Report.* OECD Publishing. [Online] http://www.oecd.org/pisa/pisaproducts/50036771.pdf [12.09.2014]

Österreichisches Zentrum für Persönlichkeitsbildung und soziales Lernen (ÖZEPS) (Hrsg.). (2010). *Förderliche Leistungsbewertung.* ÖZEPS: Wien.

Osterlind, S. J. (2001). *Constructing Test Items: Multiple Choice, Constructed-Response, Performance, and Other Formats* (2. Aufl.). Boston/Dordrecht/London: Kluwer Academic.

Tarrant, M. & Ware, J. (2008). *Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments.* Medical Education, 42, 198– 206.

Webb, N. L. (2006). Identifying content for student achievement tests. In Downing, S. & Haladyna, T. (2006). *Handbook of Test Development (S. 299–309).* New Jersey: Lawrence Erlbaum.