bifie | Bildung

*standards

Skalierung der Leistungsdaten und Linking zur Baseline-Erhebung

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

Matthias Trendtel

Bundesinstitut bifie

Bundesinstitut
bifie

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

Skalierung der Leistungsdaten und Linking zur Baseline-Erhebung
Technische Dokumentation – BIST-Ü Mathematik, 4. Schulstufe, 2013
BIFIE | Department Bildungsstandards & Internationale Assessments (BISTA),
Salzburg 2015

Inhaltsverzeichnis

1 Einführung

Bei der Überprüfung der Bildungsstandards in Mathematik auf der 4. Schulstufe im Jahr 2013 (BIST-Ü-M4) wurden 254 Items in einem Testdesign mit 30 Testheften an 73.655 Schülerinnen und Schülern administriert. Dabei bearbeiteten Schülerinnen und Schüler jeweils ein Testheft mit 68 bzw. 72 Items. Die Leistungen der Schülerinnen und Schüler wurden zum einen vergleichbar mit den Leistungen der Baseline-Testung (BL) im Jahr 2010 gemacht und zum anderen auf Kompetenzstufen anhand von aus Standard-Settings gewonnenen Cut-Scores (siehe Bazinger, Freunberger & Itzlinger-Bruneforth, 2013) verortet. D. h. um (1) Unterschiede in den mittleren Schwierigkeiten der einzelnen Testhefte innerhalb der BIST-Ü-M4 berücksichtigen, (2) die Schülerleistungen mit denen der BL vergleichen und (3) die Schülerleistungen Kompetenzstufen zuordnen zu können, ist eine Skalierung der Schülerergebnisse erforderlich.

Die BIST-Ü-M4 wurde in zwei Stichproben durchgeführt. Bei angestrebten sieben Prozent der Schülerpopulation (S7) wurden alle 254 Items eingesetzt, bei den restlichen angestrebten 93 Prozent (S93) eine Teilmenge der Items (102 Items). Aufgrund der Repräsentativität der Stichproben ist von äquivalenten Gruppen hinsichtlich der Fähigkeitsverteilung auszugehen. Die Testhefte wurden jedoch hinsichtlich verschiedener Kriterien für die beiden Stichproben unterschiedlich zusammengestellt. Hinzu kommt, dass die Stichproben durch jeweils andere Testleitergruppen (interne bzw. externe Testleiter) administriert wurden, so dass in Summe nicht auszuschließen ist, dass einzelne Items in den beiden Stichproben trotz äquivalenter Beschaffenheit der Gruppen unterschiedlich funktionieren. Daher werden zwei getrennte Skalierungen unter der Annahme gleicher Fähigkeitsverteilung in den beiden Stichproben vorgenommen und die beiden resultierenden separaten Skalen auf der Skala der BL verankert.

Innerhalb der Stichproben wurden aufgrund der Anforderungen an das Testdesign Items in mehreren Testheften an verschiedenen Positionen und in unterschiedlichen Kontexten eingesetzt (für eine Erläuterung der Positions- und Kontextbalancierung als Zielgröße des Testdesigns siehe Kiefer & George, in Vorbereitung). Um einem evtl. durch andere Faktoren begründeten unterschiedlichen Itemfunktionieren Rechnung zu tragen, werden Bookleteffekte untersucht und im Falle des Auftretens solcher Effekte bei der Skalierung berücksichtigt.

Darüber hinaus ist jedes Item einer von vier Inhaltskompetenzen (IK1–IK4) und einer von vier allgemeinen mathematischen Kompetenzen (AK1–AK4) zugeordnet. Die Leistungsdaten jeder dieser acht Kompetenzbereiche wurden getrennt skaliert. Um Konsistenz zu gewährleisten, wurde bei der Skalierung innerhalb der Teilbereiche analog zu der Skalierung der allgemeinen mathematischen Kompetenz vorgegangen.

2 Datengrundlage

In die Skalierung gingen die Leistungsdaten von insgesamt 73.655 Schülerinnen und Schülern ein. Davon haben 57 Schülerinnen und Schüler (< 0.1 %) kein Item beantwortet. Von den restlichen Schülerinnen und Schülern wurden im Mittel 4.7 % (SD 6.5 %) der Items ausgelassen und 1.7 % (SD 5.3 %) der Items nicht erreicht. Die Antworten aller Items sind dichotom kodiert. Gemäß der Testkonzeption wurden alle ausgelassenen und nicht erreichten Items als Falschantworten gewertet. Ein einzelnes Item wurde bei der BIST-Ü-M4 von mindestens 794 und höchstens 47.309 Schülerinnen und Schülern bearbeitet.

3 Skalierungsmodell

Um die dichotomen Leistungsdaten der Schülerinnen und Schüler zu analysieren, wurden eindimensionale Modelle basierend auf der Item Response Theory (IRT; siehe z. B. Yen & Fitzpatrick, 2006) eingesetzt. Die Zusammenstellung der Items eines Testhefts spiegelt die Testkonzeption umgesetzt durch das Testdesign wider. Die Items eines Testhefts sind dabei anhand verschiedener Kriterien ausgewählt worden und gingen bei deren Auswahl jeweils mit gleichem Gewicht in den Testinhalt ein (siehe Kiefer & George, in Vorbereitung). Um die Testinhalte nicht zu verzerren, wird darauf verzichtet einen itemspezifischen Trennschärfeparameter einzuführen, was im Allgemeinen zu einer ungleichen Gewichtung der Items (siehe z. B. Kolen, 2006) bei der Personenfähigkeitsschätzung führt. Konsequenterweise wurden Raschmodelle bzw. Modelle ohne itemspezifische Ternnschärfeparameter verwendet.

Die hier beschriebenen Modelle wurden zu deren eigentlicher Berechnung als Random Coefficients Multinomial Logit Models (RCML; Adams, Wilson & Wang, 1997) mit einem zugrundeliegenden Populationsmodell spezifiziert und mit dem Softwarepaket TAM (Kiefer, Robitzsch & Wu, 2014) in der Statistiksoftwareumgebung R (R Core Team, 2013) mithilfe der Marginal-Maximum-Likelihood-Methode (MML) geschätzt. Bei der Schätzung wurden die um Ausfallraten adjustierten Stichprobengewichte berücksichtigt.

Die Notation in den folgenden Ausführungen v. a. des Response-Modells orientiert sich an Notationen, wie sie z. B. in Hambleton, Swaminathan und Rogers (1991) zu finden sind, und nicht an der Notation in Adams et al. (1997).

Wir gehen von einem Populationsmodell aus, das die Verteilung der zu messenden Fähigkeit θ – in diesem Fall die Fähigkeit in Mathematik (bzw. in den entsprechenden einzelnen Kompetenzbereichen) auf der 4. Schulstufe – in der Population der zu testenden Schülerinnen und Schüler als normalverteilt annimmt,

$$\theta \sim N(\mu, \sigma^2), \tag{1}$$

mit zugehöriger Dichte $g(\theta; \mu, \sigma)$, wobei μ den Mittelwert und σ die Standardabweichung der Verteilung von θ bezeichnet. Dabei wird zur Identifizierbarkeit der Mittelwert μ auf null fixiert, während σ als unbekannt angenommen wird und anhand der Daten frei zu schätzen ist. Durch die Fixierung von μ werden die Itemparameter normalisiert (siehe z. B. Fischer, 2006).

Für die I Items des betrachteten Tests sei $\boldsymbol{\xi} = (\xi_1, ..., \xi_I)$ der Vektor, dessen Komponente ξ_i die Eigenschaft von Item i beschreibt, und als Itemparametervektor bezeichnet wird. $\boldsymbol{X} = (X_1, ..., X_I)$ sei der Vektor der Zufallsvariablen, bei denen X_i die Ereignisse „falsche Antwort“ und „richtige Antwort“ auf Item i auf die Menge $\{0, 1\}$ abbildet. Für ein Item i sei der Itemparameter ξ_i gegeben. Nach Rasch (1960) wird die Wahrscheinlichkeit, dass eine Schülerin bzw. ein Schüler $p = 1, ..., N$ mit Fähigkeit θ_p auf Item i Antwort $x_i^{(p)} \in \{0, 1\}$ gibt, modelliert durch

$$P(X_i = x_i^{(p)}; \xi_i | \theta_p) = \frac{\exp(x_i^{(p)}(\theta_p - \xi_i))}{1 + \exp(\theta_p - \xi_i)}. \tag{2}$$

Unter der Annahme der lokalen stochastischen Unabhängigkeit lässt sich damit bei gegebenem Itemparametervektor $\boldsymbol{\xi}$ die Wahrscheinlichkeit, einen Antwortvektor $\boldsymbol{x^{(p)}} = (x_1^{(p)}, ..., x_I^{(p)})$ bei Schülerin bzw. Schüler p mit Fähigkeit θ_p zu beobachten, schreiben als

$$P(\boldsymbol{X} = \boldsymbol{x^{(p)}}; \boldsymbol{\xi}|\theta_p) = \prod_{i=1}^{I} P(X_i = x_i^{(p)}; \xi_i|\theta_p). \tag{3}$$

Die Wahrscheinlichkeit, von einer beliebigen Schülerin bzw. einem beliebigen Schüler (d. h. zufällig aus der Population gezogenen Schülerin bzw. Schüler; für eine Ausführung dieser „random sampling"-Sichtweise siehe z. B. Holland, 1990) den Antwortvektor $\boldsymbol{x}$ zu beobachten, errechnet sich mithilfe des Populationsmodells (1) durch

$$P(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\xi}) = \int_{-\infty}^{+\infty} P(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\xi}|\theta) g(\theta; \mu, \sigma) d\theta \tag{4}$$

bei gegebenen Populationsparametern μ und σ.
Seien nun $\boldsymbol{x^{(1)}}, ..., \boldsymbol{x^{(N)}}$ die beobachteten Antwortvektoren der N Schülerinnen und Schüler. Die Likelihood dieser beobachteten Daten berechnet sich für beliebige, aber feste Parameter $\boldsymbol{\xi}, \mu$ und σ durch

$$L(\boldsymbol{x^{(1)}}, ..., \boldsymbol{x^{(N)}}; \boldsymbol{\xi}, \mu, \sigma) = \prod_{p=1}^{N} \int_{-\infty}^{+\infty} P(\boldsymbol{X} = \boldsymbol{x^{(p)}}; \boldsymbol{\xi}|\theta) g(\theta; \mu, \sigma) d\theta. \tag{5}$$

Darauf basierend lassen sich der Itemparametervektor $\boldsymbol{\xi}$ und die Parameter des Populationsmodells μ und σ mithilfe des EM-Algorithmus (Dempster, Laird & Rubin, 1977) als MML-Schätzer bestimmen. Für eine Ausführung des Schätzverfahrens im Kontext des Raschmodells siehe z. B. Thissen (1982) oder Bock und Moustaki (2006). Das MML-Schätzverfahren im Kontext der RCML ist in Adams et al. (1997) beschrieben.

Neben den Parametern erhält man bei dem MML-Schätzverfahren für jedes Antwortmuster $\boldsymbol{x}$ die bedingte a-posteriori-Verteilung von θ mit Dichte

$$g(\theta; \mu, \sigma|\boldsymbol{x}) = \frac{P(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\xi}|\theta) g(\theta; \mu, \sigma)}{P(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\xi})}. \tag{6}$$

Gemäß dieser bedingten a-posteriori-Verteilung werden für jede Schülerin und für jeden Schüler jeweils zehn Plausible Values gezogen, um auf Eigenschaften in der Populationsverteilung Rückschlüsse zu ziehen. Insbesondere werden Mittelwerte und Varianzen in der Population auf diese Weise geschätzt. Für eine Motivation und eine Erläuterung dieses Vorgehens siehe Mislevy, Beaton, Kaplan und Sheehan (1992) oder auch von Davier, Gonzalez und Mislevy (2009).

Während für Rückschlüsse in der Population die bedingten a-posteriori-Verteilungen der Schülerinnen und Schüler verwendet werden, werden den Argumenten in McDonald (2011) folgend für individuelle Rückmeldungen Schätzer basierend auf Maximum Likelihood Estimators (MLE) berechnet. Dabei wird der durch das eben beschriebene MML-Schätzverfahren gewonnene Itemparametervektor $\boldsymbol{\xi}$ in das Response-Modell in Gleichung (2) eingesetzt. Somit lässt sich für jedes Antwortmuster $\boldsymbol{x}$ eine Likelihood bestimmen, die

ihr Maximum erreicht, wenn die Gleichung gilt

$$l(\boldsymbol{x};\boldsymbol{\xi}|\theta) \equiv \frac{\partial \ln \prod_{i=1}^{I} P(X_i = x_i;\xi_i|\theta)}{\partial\theta} = 0. \tag{7}$$

Die so erhaltenen MLEs von θ weisen allerdings eine Verzerrung auf, deren Größenordnung von θ abhängt (siehe Lord, 1983). Daher werden Weighted Likelihood Estimators (WLE; Warm, 1989) verwendet, bei deren Berechnung in der Schätzgleichung (7) diese Verzerrung bereits berücksichtigt wird.

Die Schätzgleichung der WLEs im Kontext der RCMLs ist z. B. in Adams und Wu (2007) beschrieben.

4 Booklet- und Administrationseffekte

In der S7 wurden die Testhefte 1–3 und 7–30 eingesetzt. Dabei wurden die Testhefte 1–3 auch in der S93 administriert und die Testhefte 28–30 kamen bereits in der BL-Testung 2010 in unveränderter Form zum Einsatz. In der S93 wurden die Testhefte 1–6 eingesetzt, wobei die Testhefte 4–6 dieselben Items enthalten wie die Testhefte 1–3 und lediglich die Positionen und Kontexte der Items verändert wurden.
Insgesamt untersuchen wir 4 Testheft- bzw. Bookletgruppen (BG1–BG4), bei denen wir hinsichtlich der unterschiedlichen Zusammenstellung der Testhefte hinsichtlich des Zeitpunkts der Testheftzusammenstellung (Baseline) und der Zielgrößen des Testdesigns (z. B. unterschiedliche Gewichtung des Antwortformats in der S7 und S93) ein differenzielles Bookletfunktionieren und damit mittlere Bookletgruppeneffekte nicht ausschließen können:

BG1 Testhefte 7–27; diese Testhefte sind nur in der S7 administriert worden,

BG3 Testhefte 28–30; diese Testhefte sind originale BL-Testhefte und sind nur in der S7 administriert worden,

BG2 Testhefte 1–3; diese Testhefte sind sowohl in der S7 als auch in der S93 administriert worden,

BG4 Testhefte 4–6; diese Testhefte sind nur in der S93 administriert worden.

Um die einzelnen Bookleteffekte zu untersuchen, werden jeweils Zweigruppen-Modelle (siehe z. B. Bock & Zimowski, 1997) angepasst. Das heißt, wir gehen jeweils von zwei Gruppen aus und für die Gruppe $g = 1, 2$ wird Gleichung (1) zu

$$\theta \sim N(\mu_g, \sigma_g^2). \tag{8}$$

Dabei stellt ohne Beschränkung der Allgemeinheit Gruppe 1 immer die Referenzgruppe dar, deren Mittelwert μ_1 aus Gründen der Identifizierbarkeit auf null fixiert wird. Die Parameter μ_2, σ_1 und σ_2 des Populationsmodells sind dagegen frei zu schätzen.

Ein nicht zu vernachlässigender Bookleteffekt liegt bei dieser Analysemethode dann vor, wenn sich die geschätzten Mittelwerte $\hat{\mu}_g$ oder die geschätzten Varianzen $\hat{\sigma}_g^2$ der beiden a-priori-Verteilungen im Populationsmodell (8) signifikant unterscheiden.

Dafür werden jeweils zwei getrennte Hypothesentests formuliert, wobei sowohl

$$t_\mu = \hat{\mu}_2 \text{ (für den Mittelwertvergleich)}$$

als auch

$$t_{\sigma^2} = \ln(\hat{\sigma}_1^2) - \ln(\hat{\sigma}_2^2) \text{ (für den Vergleich der Varianzen)}$$

unter der Nullhypothese (keine Gruppenunterschiede) jeweils gleich null angenommen wird.

Um die Standardfehler sd_{t_μ} und $sd_{t_{\sigma^2}}$ der interessierenden Größen t_μ bzw. t_{σ^2} zu schätzen, wird das im Large Scale Assessment übliche Jackknifeverfahren (siehe z. B. Gershunskaya, Jiang & Lahiri, 2009) verwendet unter Berücksichtigung der vorliegenden Stichprobenstruktur (stratifizierte Klumpenstichprobe). Dabei wird für jede Jackknife-Stichprobe eine Schule aus der originalen Stichprobe entfernt, die übrigen Schulen im jeweiligen Stratum regewichtet und für jede so neu erhaltene Stichprobe die Teststatistik neu berechnet. Sei n_h die Anzahl der für die jeweiligen Bookletgruppen interessierenden Schulen im Stratum $h = 1, ..., H$ (d. h. die Schulen aus Stratum h, an denen Testhefte aus der einen oder der anderen Bookletgruppe administriert wurden), t die interessierende Größe in der originalen Stichprobe und $t^{(hj)}$ die interessierende Größe berechnet basierend auf der Stichprobe ohne Schule j mit $j = 1, ..., n_h$ in Stratum h mit entsprechend adjustierten Gewichten. Die Stichprobenvarianz wird dann geschätzt durch

$$\widehat{sd_t^2} = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} (t - t^{(hj)})^2. \tag{9}$$

Die daraus resultierenden Teststatistiken

$$T_\mu = \frac{t_\mu}{\sqrt{\widehat{sd_{t_\mu}^2}}} \text{ und } T_{\sigma^2} = \frac{t_{\sigma^2}}{\sqrt{\widehat{sd_{t_{\sigma^2}}^2}}}$$

sind jeweils asymptotisch t-verteilt mit $n - H$ Freiheitsgraden, wobei n die Anzahl aller Schulen in allen H Strata für die jeweiligen Bookletgruppen repräsentiert.

Die Ergebnisse der Analysen sind in Tabelle 1 zusammengefasst. Da bei allen Vergleichen der jeweiligen Gruppen signifikante Unterschiede hinsichtlich Varianz oder Mittelwert im Populationsmodell auftreten, werden Items, die in zwei oder mehreren Bookletgruppen administriert wurden, jeweils als unterschiedliche Items aufgefasst und gehen damit als zwei bzw. mehrere separate Items in das Skalierungsmodell ein.

Eine analoge Untersuchung des Administrationseffekts bei den Testheften 1–3, die sowohl in der S7 als auch in der S93 administriert wurden, ergab keine signifikanten Unterschiede.

Um die nicht vernachlässigbaren Bookleteffekte (siehe Tabelle 1) zu berücksichtigen, wird für eine Schülerin bzw. einen Schüler p mit Fähigkeit θ_p, der bzw. dem ein Testheft aus Bookletgruppe $g[p] \in \{BG1, ..., BG4\}$ administriert wurde, für Item i und Antwort

Tabelle 1: Ergebnisse der Analysen zu Bookleteffekten der Bookletgruppen BG1 - BG4. T_μ ist die Teststatistik für den Mittelwertvergleich, T_{σ^2} die Teststatistik für den Vergleich der Varianzen, $\widehat{sd_{t_\mu}}$ und $\widehat{sd_{t_{\sigma^2}}}$ sind die entsprechenden Standardfehler, p_μ und p_{σ^2} die entsprechenden p-Werte der Hypothesentests und *df* ist die Anzahl der Freiheitsgrade der zugrundeliegenden t-Verteilung.

Bookletgruppen	$\hat{\mu}_2$	$\widehat{sd_{t_\mu}}$	p_μ	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\widehat{sd_{t_{\sigma^2}}}$	p_{σ^2}	*df*
BG1 und BG2 (S7)	0.089	0.030	0.003	0.903	0.855	0.050	0.284	350
BG1 und BG3 (S7)	0.004	0.032	0.890	1.013	0.900	0.046	0.010	350
BG2 und BG4 (S93)	–0.033	0.007	< 0.001	0.831	0.791	0.012	< 0.001	2943

$x_i^{(p)}$ die Responsefunktion in Gleichung (2) zu

$$P(X_i = x_i^{(p)}; \xi_{i,g[p]}|\theta_p) = \frac{\exp(x_i^{(p)}(\theta_p - \xi_{i,g[p]}))}{1 + \exp(\theta_p - \xi_{i,g[p]})}. \tag{10}$$

Das heißt, es werden bookletgruppenspezifische Itemparameter $\xi_{i,g[p]}$ eingeführt, wobei $g[p]$ indiziert, dass Person p ein Testheft aus Bookletgruppe $g \in \{BG1, ..., BG4\}$ bearbeitet hat.

Durch dieses Vorgehen wird der mittlere Schwierigkeitsunterschied dieser Bookletgruppe im Vergleich zu den anderen Bookletgruppen kontrolliert. Um weiterhin noch die Unterschiede in den mittleren Trennschärfen der verschiedenen Bookletgruppen, die sich in den unterschiedlichen Varianzen der a-priori-Verteilungen ausdrücken, zu berücksichtigen, wird weiterhin das Response-Modell um einen bookletgruppenspezifischen Trennschärfeparameter $a_{g[p]}$ erweitert:

$$P(X_i = x_i^{(p)}; \xi_{i,g[p]}, a_{g[p]}|\theta_p) = \frac{\exp\left(a_{g[p]}(\theta_p - \xi_{i,g[p]})\right)^{x_i^{(p)}}}{1 + \exp\left(a_{g[p]}(\theta_p - \xi_{i,g[p]})\right)}. \tag{11}$$

Die geschätzten bookletgruppenspezifischen Trennschärfeparameter sind in Tabelle 2 dargestellt.

Die Einführung von einem mittleren Trennschärfeparameter pro Bookletgruppe stellt lediglich eine Erweiterung des in Kapitel 3 beschriebenen Skalierungsmodells dar und lässt sich analog in den MML-Ansatz einbetten. Ebenso lässt es sich als RCML spezifizieren und mit dem Softwarepaket TAM schätzen.

5 Linking

Im Folgenden wird die in Haberman (2009) vorgestellte Methode des simultanen Linkings mehrerer Studien skizziert. Dabei werden die Items hinsichtlich ihrer Schwierigkeit und Trennschärfe als konstant über die verschiedenen Erhebungen hinweg angenommen. Weiterhin geht man davon aus, dass die zu messenden Fähigkeiten normalverteilt sind und sich lediglich Mittelwert und Standardabweichung über die Erhebungen hinweg ändern. Die Berechnung wurde mit dem Paket sirt (Robitzsch, 2014) in der Softwareumgebung R durchgeführt.

Tabelle 2: Dargestellt sind die Trennschärfeparameter der einzelnen Bookletgruppen der globalen Skalierung (über alle Teilkompetenzen hinweg) sowie für die einzelnen Teilkompetenzbereiche jeweils für die Skalierung der S7 und der S93.

	S7			S93	
Bereich	BG1	BG2	BG3	BG2	BG4
global	0.949	0.931	1.009	0.916	0.893
IK1	1.021	1.085	1.085	1.048	1.017
IK2	1.014	0.954	1.131	0.939	0.928
IK3	0.932	0.914	1.027	0.907	0.872
IK4	0.931	0.876	0.979	0.853	0.831
AK1	0.980	1.005	1.135	0.984	0.963
AK2	0.860	0.981	0.922	0.930	0.916
AK3	0.959	0.900	1.038	0.931	0.883
AK4	1.024	0.880	1.021	0.863	0.849

Für jede Studie $t \in \{BL, S7, S93\}$ und jede Person p in der jeweilgen Studie t sei $\theta_p^{(t)} \sim N(B_t, A_t^2)$. Aus Gründen der Identifizierbarkeit setzen wir $B_{BL} = 0$ und $A_{BL} = 1$. Eine entsprechende Transformation der BL-Skala wurde anhand von Populationsschätzern basierend auf den Plausible Values durchgeführt.
Sei

$$\theta^{*(t)} = (\theta^{(t)} - B_t)/A_t \tag{12}$$

die Lineartransformation von $\theta^{(t)}$, sodass $\theta^{*(t)} \sim N(0, 1)$.
Den Log Odds Ratio aus Gleichung 11 für ein Item i und eine Person p in Studie t mit Fähigkeit $\theta_p^{(t)}$ schreiben wir dann als

$$\log \frac{P(X_i = 1; \xi_{i,g[p]}, a_{g[p]}^{(t)} | \theta_p^{*(t)})}{P(X_i = 0; \xi_{i,g[p]}, a_{g[p]}^{(t)} | \theta_p^{*(t)})} = a_{g[p]}^{(t)} (\theta_p^{*(t)} - \xi_{i,g[p]}^{(t)}) \tag{13}$$

Setze

$$a_{g[p]}^{(t)} = A_t a_{g[p]} \tag{14}$$

und

$$\xi_{i,g[p]}^{(t)} = (\xi_{i,g[p]} - B_t)/A_t. \tag{15}$$

Dann wird (13) zu

$$\log \frac{P(X_i = 1; \xi_{i,g[p]}, a_{g[p]}^{(t)} | \theta_p^{*(t)})}{P(X_i = 0; \xi_{i,g[p]}, a_{g[p]}^{(t)} | \theta_p^{*(t)})} = a_{g[p]} (A_t \theta_p^{*(t)} + B_t - \xi_{i,g[p]}). \tag{16}$$

$A_t \theta_p^{*(t)} + B_t$ in (16) ist die Umkehrfunktion von (12) und es gilt $A_t \theta_p^{*(t)} + B_t \sim N(B_t, A_t)$. Für alle $t \in \{BL, S7, S93\}$ erhalten wir wie in den Kapiteln 3 und 4 beschrieben die

Parameterschätzer $\hat{\boldsymbol{\xi}}^{(t)}$ und $\hat{\boldsymbol{a}}^{(t)} = (\hat{a}^{(t)}_{BG1}, ..., \hat{a}^{(t)}_{BG4})$ durch separate Skalierungen der entsprechenden Daten.

Für alle $i = 1, ..., I$ und jede Bookletgruppe g_i, in der Item i administriert wurde, bleiben die Parameter ξ_{i,g_i} und a_{g_i} sowie für $t \in \{S7, S93\}$ die Parameter A_t und B_t zu schätzen (beachte, dass $B_{BL} = 0$ und $A_{BL} = 1$ fixiert wird). Die Schätzung erfolgt in zwei Schritten.

Aus Gleichung (14) erhalten wir

$$\log a_g^{(t)} \approx \log A_t + \log a_g$$

und betrachten das Regressionsmodell, in dem die Schätzer $\hat{A}_t$ und $\hat{a}_g$ den Term

$$\sum_{t \in \{BL,S7,S93\}} \sum_{g \in G_t} \left[\log \hat{a}_g^{(t)} - \log \hat{A}_t - \log \hat{a}_g\right]^2$$

minimieren, wobei G_t die Menge der Bookletgruppen ist, die bei Studie t administriert wurden.

Im zweiten Schritt werden unter Verwendung von Gleichung (15) die Schätzer $\hat{B}_t$ und $\hat{\xi}_{i,g}$ so bestimmt, dass

$$\sum_{t \in \{BL,S7,S93\}} \sum_{i \in I_t} \sum_{g_{i,t} \in G_{i,t}} \left[\hat{\xi}^{(t)}_{i,g_i} \hat{A}_t + \hat{B}_t - \hat{\xi}_{i,g_i}\right]^2$$

minimiert wird, wobei I_t die Menge der Items ist, die in Studie t administriert wurden, und $G_{i,t}$ die Menge der Bookletgruppen ist, in denen Item i in Studie t eingesetzt wurde. Die so erhaltenen $\hat{A}_t, \hat{B}_t, \hat{a}_g$ und $\hat{\xi}_{i,g_i}$ sind geeignete Schätzer für die Parameter in Gleichung (16). In Tabelle 3 sind die geschätzten Parameter, die der Transformation dienen, für das globale und das kompetenzbereichsweise Linking dargestellt.

Tabelle 3: Dargestellt sind die Schätzer der einzelnen Transformationsparameter A_{S7}, B_{S7}, A_{S93} und B_{S93}, die sich global und kompetenzbereichsweise aus dem Haberman-Linking ergeben.

Bereich	$\hat{A}_{S7}$	$\hat{B}_{S7}$	$\hat{A}_{S93}$	$\hat{B}_{S93}$
global	1.0204	0.2614	1.0015	0.3381
AK1	1.0037	0.3506	0.9824	0.4293
AK2	1.0286	0.2174	0.9748	0.2947
AK3	0.9859	0.2311	1.0196	0.3081
AK4	1.0173	0.2076	0.9991	0.2535
IK1	0.9529	0.4013	0.9214	0.4679
IK2	1.0231	0.3456	1.0060	0.4296
IK3	1.0096	0.2023	1.0018	0.2722
IK4	1.0268	0.0789	1.0012	0.1295

Mit den erhaltenen Schätzern $\hat{A}_t$ und $\hat{B}_t$ werden die Fähigkeitsschätzer $\theta^{*(t)}$ auf der Baselinemetrik, die auf Mittelwert 500 und Standardabweichung 100 normiert wird, durch die Transformation

$$T(\theta^{*(t)}) = (\theta^{*(t)}\hat{A}_t + \hat{B}_t)100 + 500 \tag{17}$$

verankert.

Literatur

Adams, R. J., Wilson, M. & Wang, W.-C. (1997). The multidimensional random coefficient multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

Adams, R. J. & Wu, M. L. (2007). The mixed-coefficients multinomial logit model: A generalized form of the Rasch model. In M. von Davier & C. Carstensen (Hrsg.), *Multivariate and mixture distribution Rasch models* (S. 57–75). New York: Springer.

Bazinger, C., Freunberger, R. & Itzlinger-Bruneforth, U. (2013). *Standard-Setting Mathematik. Technische Dokumentation – BIST-Ü Mathematik, 4. Schulstufe, 2013.* (Tech. Rep.). Salzburg: BIFIE.

Bock, R. D. & Moustaki, I. (2006). Chapter 15 – Item response theory in a general framework. In C. Rao & S. Sinharay (Hrsg.), *Handbook of statistics psychometrics* (Bd. 26, S. 469–513). Elsevier.

Bock, R. D. & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. Hambleton (Hrsg.), *Handbook of modern item response theory* (S. 433–448). New York: Springer.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38.

Fischer, G. H. (2006). Chapter 16 – Rasch models. In C. Rao & S. Sinharay (Hrsg.), *Handbook of statistics psychometrics* (Bd. 26, S. 515–585). Elsevier.

Gershunskaya, J., Jiang, J. & Lahiri, P. (2009). Chapter 28 – resampling methods in surveys. In C. Rao (Hrsg.), *Handbook of statistics sample surveys: Inference and analysis* (Bde. 29, Part B, S. 121–151). Elsevier.

Haberman, S. J. (2009). *Linking parameter estimates derived from an item response model through separate calibrations* (ETS Research Report RR09-40). Princeton: ETS.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park: Sage.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*, 577–601.

Kiefer, T. & George, A. C. (in Vorbereitung). *Pilotierung und Testdesign. Technische Dokumentation – BIST-Ü Mathematik, 4. Schulstufe, 2013.* (Tech. Rep.). Salzburg: BIFIE.

Kiefer, T., Robitzsch, A. & Wu, M. (2014). *TAM: Test analysis modules.* (R package version 1.0-3)

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Hrsg.), *Educational measurement* (S. 155–186). Westport: Praeger Publisher.

Lord, F. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*(2), 233–245.

McDonald, R. P. (2011). Measuring latent quantities. *Psychometrika*, *76*, 511–536.

Mislevy, R. J., Beaton, A. E., Kaplan, B. & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133–161.

R Core Team. (2013). *R: A language and environment for statistical computing.* Vienna, Austria.

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

Robitzsch, A. (2014). *sirt: Supplementary item response theory models.* (R package version 0.43-70)

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, *47*(2), 175–186.

von Davier, M., Gonzalez, E. & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI monograph series*, *2*, 9–36.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

Yen, W. M. & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Hrsg.), *Educational measurement* (S. 111–154). Westport: Praeger Publisher.