UNIDIMENSIONAL INTERPRETATION OF MULTIDIMENSIONAL TESTS

Dissertation zur Erlangung des Doktorgrades
der Philosophischen Fakultät
der Christian-Albrechts-Universität
zu Kiel

vorgelegt von Steffen Brandt

Kiel
September 2015

Erstgutachter: Prof. Dr. Gabriel Nagy
Zweitgutachter: Prof. Dr. Andreas Frey (Universität Jena)

Tag der mündlichen Prüfung: 2. Februar 2016

Durch den zweiten Prodekan, Prof. Dr. John Peterson, zum Druck genehmigt: 3. Februar 2016

Zusammenfassung

Traditionell wurden Fragebogendaten und Daten aus Leistungstests[1] mit Hilfe der klassischen Testtheorie (KTT) ausgewertet, etwa durch Bildung von Summen- und Mittelwerten. Die Verwendung der KTT hat jedoch entscheidende Nachteile zur Folge: (a) Die KTT umfasst keine Theorie zur Berechnung von Aufgabenschwierigkeiten, wodurch die Untersuchung von Testeigenschaften deutlich eingeschränkt ist sowie auch keine Verknüpfung von Tests über Aufgabenteilmengen möglich ist; und (b) die KTT beinhaltet sehr starke Annahmen hinsichtlich der Eigenschaften eines Tests (siehe, z.B., Moosbrugger & Kelava, 2007; Rost, 1996).

Die großen internationalen Leistungsstudien wenden aus den genannten Gründen daher nicht die KTT an, sondern Auswertungsverfahren basierend auf der probabilistischen Testtheorie, üblicherweise bezeichnet als Item Response Theorie (IRT). Auch im Rahmen der IRT geht man von verschiedenen statistischen Annahmen aus. Eine sehr wichtige und grundsätzliche Annahme ist dabei die zur Dimensionalität eines Tests. Es gilt, dass ein Test eindimensional sein muss, um eindimensional ausgewertet werden zu können. Diese Feststellung mag selbstverständlich erscheinen, in der Realität führen jedoch praktische Zwänge häufig dazu, dass diesem Grundsatz widersprochen wird. Im *Programme for International Student Assessment* (PISA), zum Beispiel, wird für Mathematik einerseits ein eindimensionaler Leistungswert berechnet, andererseits werden jedoch auch Leistungswerte in den Subskalen, oder Subdimensionen, *Quantität*, *Raum und Form*, *Veränderung und Beziehungen* und *Unsicherheit und Daten* berichtet. Das heißt, einerseits wird angenommen, dass Mathematik ein eindimensionales Konstrukt ist, andererseits jedoch, dass es ein mehrdimensionales Konstrukt ist. Dieser Widerspruch findet sich in gleicher Weise bei der Auswertung des Lese- und des Naturwissenschaftstest in PISA (OECD, 2012b) und auch in den anderen großen, internationalen Vergleichsstudien, wie der *Trends in International Mathematics and Science Study* (TIMSS) und der *Progress in International Reading Literacy Study* (PIRLS), sind die Auswertungen diesbezüglich widersprüchlich (Martin & Mullis, 2012). Es darf angenommen werden, dass praktische Zwänge, das heißt unter anderem Vorgaben der Auftraggeber, die Ursache für diesen Widerspruch in der Auswertung der Studien sind. Auffallend ist aber, dass in keinem der genannten Fälle das Problem diskutiert wird.

[1] Im Weiteren werden Fragebögen und Leistungstests unter der Bezeichnung Tests zusammengefasst. Die Bezeichnung „Aufgaben“ bezieht sich entsprechend auch immer vergleichbar auf „Fragen“.

Einen anderen Ansatz zur Auswertung verfolgt die in den USA bekannteste Schulleistungsstudie, das *National Assessment of Educational Progress* (NAEP). In NAEP wird zum Beispiel Lesen, wie in den zuvor genannten Studien auch, als mehrdimensionales Konstrukt angenommen, zusammengesetzt aus den Subdimensionen *Lesen als Literarische Erfahrung* („Reading for Literary Experience"), *Lesen zum Informationsgewinn* („Reading to Gain Information") und *Lesen zur Bewältigung von Aufgaben* („Reading to Perform a Task") (Donahue & Schoeps, 2001). Die Leistungswerte für diese drei Subdimensionen von Lesen werden mit Hilfe eines mehrdimensionalen IRT-Modells berechnet. Der Gesamtwert für Lesen wird jedoch nicht mit Hilfe eines eindimensionalen IRT-Modells berechnet, sondern durch einen gewichteten Mittelwert auf Basis der mehrdimensionalen Leistungswerte (Allen, Carlson, & Donoghue, 2001).

Alle erwähnten Studien haben gemeinsam, dass sie in den letzten Jahrzehnten regelmäßig durchgeführt wurden und von starkem politischen und öffentlichen Interesse begleitet werden. Die Studien erfahren daher auch im wissenschaftlichen Bereich große Aufmerksamkeit und stehen unter besonderem Druck ihre Auswertungen gemäß dem aktuellen Stand der Forschung durchzuführen. Hinsichtlich der Berechnung eindimensionaler Leistungswerte für als mehrdimensional angenommene Daten verfolgen die Studien jedoch trotzdem wie beschrieben zwei unterschiedliche Ansätze. Dies kann schon als erstes Indiz dafür angesehen werden, dass sowohl der in PISA, TIMSS und PIRLS verfolgte Ansatz als auch der in NAEP verfolgte Ansatz Vor- und Nachteile mit sich bringt. Ziel des folgenden Abschnitts ist es daher zunächst, die Vor- und Nachteile der beiden bisherigen Ansätze anhand verschiedener Gesichtspunkte zu verdeutlichen (siehe auch Tabelle 1), bevor im Anschluss daran ein IRT-Modell vorgeschlagen wird, das als eine Art Kombination der beiden Ansätze betrachtet werden kann: das Generalisierte Subdimensionsmodell (GSM). Im GSM wird die Schätzung des mehrdimensionalen IRT-Modells dabei so restringiert, dass zusätzlich zur Schätzung der mehrdimensionalen Leistungswerte auch ein eindimensionaler Leistungswert geschätzt wird, der dem eines gewichteten Mittelwerts über die Subdimensionen entspricht. Nach der Darstellung verschiedener Anwendungen wird eine Einordnung des Modells in Bezug zu anderen bereits bestehenden Modellen gegeben und abschließend ein Ausblick auf die zukünftige Anwendung gegeben.

Tabelle 1

Vor- und Nachteile eines als Mittelwert berechneten Gesamtwerts und einer eindimensionalen IRT-Auswertung für mehrdimensionale Daten

Eindimensionale IRT-Auswertung	
Vorteile	Nachteile
• Berechnung von "Maximum Likelihood Estimates" (WLE, MLE, ...) ist möglich.	• Überschätzung der Reliabilität durch die Vernachlässigung von lokalen Abhängigkeiten • Fragwürdige Validität des mehrdimensionalen Konstrukts, da der Test eindimensional konstruiert wurde. • Unklare, implizite Gewichtung des eindimensionalen Leistungswerts
Auf mehrdimensionaler IRT-Auswertung basierender Mittelwert	
Vorteile	Nachteile
• Berücksichtigung lokaler Abhängigkeiten aufgrund von Mehrdimensionalität und damit eine angemessenere Schätzung der Reliabilität • Klare Gewichtung des eindimensionalen Leistungswerts	• Berechnung von "Maximum Likelihood Estimates" (WLE, MLE, ...) ist *nicht* möglich. • Nicht angebracht für das Rasch Modell, da die notwendige Standardisierung der mehrdimensionalen Leistungswerte den Messfehler erhöht.

Vor- und Nachteile bisheriger Ansätze

Lokale Abhängigkeit

Eine grundlegende Annahme von IRT-Modellen ist die der lokalen stochastischen Unabhängigkeit. Diese beschreibt die Annahme, dass die Antworten eines Tests unter Berücksichtigung der Leistungswerte in der zu messenden Dimension vollständig unabhängig voneinander sind. Eine Verletzung dieser Annahme bezeichnet man als lokale Abhängigkeit bzw. im Englischen als „Local Item Dependence" (LID). LID kann verschiedene Ursachen haben. Der wohl am meisten betrachtete Fall ist der von LID aufgrund von

Aufgabengruppen, auch „Testlets“ genannt. Aufgaben in solchen Testlets beziehen sich auf einen gemeinsamen Stimulus, der als Kontext der Aufgaben genutzt wird. Der Vorteil der Verwendung von Testlets liegt in einer effektiveren Nutzung der Testzeit. Dadurch, dass die Personen sich nicht für jede Aufgabe in einen neuen Stimulus einlesen müssen, können sie in der gleichen Zeit mehr Aufgaben bearbeiten. Nachteil ist jedoch, dass, wenn eine Person eine Aufgabe zu einem Stimulus korrekt beantworten kann, es sehr häufig so ist, dass ihre Wahrscheinlichkeit eine Aufgabe zum gleichen Stimulus zu lösen etwas höher ist, als die zu einem anderen Stimulus, bei der sie eine Aufgabe zuvor nicht lösen konnte. Das heißt, die Aufgaben zeigen LID. In der gleichen Weise kann man im Fall von Subdimensionen argumentieren. Nimmt man an, dass eine zu messende Dimension aus unterschiedlichen Subdimension zusammengesetzt ist, so bedeutet dies, dass Aufgaben, die zur gleichen Subdimension gehören, stärker miteinander verbunden sind als solche, die unterschiedlichen Dimensionen angehören. Die Auswirkungen von LID auf IRT-Auswertungen wurden von zahlreichen Autoren untersucht. Dabei wurde einheitlich festgestellt, das eine Vernachlässigung von LID zu einer verzerrten Schätzung der Schwierigkeitsparameter führt, einer Überschätzung der Diskrimination der Aufgaben, einer Verzerrung der geschätzten Varianzen und einer Überschätzung der Reliabilität (siehe, z.B., Monseur, Baye, Lafontaine, & Quittre, 2011; Tuerlinckx & De Boeck, 2001; Wainer, Bradlow, & Wang, 2007; Wang & Wilson, 2005; Yen, 1984).

Die betrachteten großen Leistungsstudien gehen sehr unterschiedlich mit möglichen lokalen Abhängigkeiten um. Während in PISA in allen untersuchten Bereichen Testlets eingesetzt werden, wird in NAEP, TIMSS und PIRLS für den Mathematiktest, zum Beispiel, auf Testlets verzichtet und jede Aufgabe besitzt einen eigenen Stimulus. In NAEP werden die Tests zudem generell durch zur Verfügung stehende Indizes auf LID hin überprüft und entsprechend der Ergebnisse werden eigentlich getrennte Aufgaben gegebenenfalls zu einer Aufgabe zusammengefasst, um LID zu vermeiden[1] (Allen & Carlson, 1987, S. 236–237). Für TIMSS, PIRLS und PISA werden dagegen in keiner der technischen Berichte Ergebnisse zur Untersuchung von LID erwähnt. Es ist jedoch aus anderen Veröffentlichungen bekannt, dass zum Beispiel die in PISA verwendeten Testlets LID zur Folge haben (Brandt, 2006; Monseur u. a., 2011).

Hinsichtlich von LID durch Subdimensionen wird in NAEP ebenfalls ein anderer Ansatz als in den übrigen Studien verfolgt. In NAEP werden die Subdimension zunächst mit Hilfe

[1] Durch das Zusammenfassen von Aufgaben kann man LID vermeiden. Ein Nachteil ist jedoch, dass dabei Antwortinformation verloren geht, da nur noch der Gesamtwert über die Aufgaben eingeht (Yen, 1993).

eines mehrdimensionalen IRT-Modells ausgewertet und die Ergebnisse dann über einen gewichteten Mittelwert zu einem Gesamtwert zusammengefasst (Allen u. a., 2001, S. 155). Auf diese Weise werden negative Auswirkungen durch LID (der Subdimensionen) auf die IRT-Auswertung verhindert. In TIMSS, PIRLS und PISA hingegen werden die Subdimensionen zusammengenommen und durch ein eindimensionales IRT-Modell ausgewertet, um einen Gesamtwert zu berechnen. Mögliche Effekte durch LID werden nicht betrachtet.

Gewichtung der Subdimensionen

In NAEP legen Experten der jeweiligen Fachgebiete fest, welche Subdimension mit welchem Gewicht zu berücksichtigen ist (siehe etwa Donahue & Schoeps, 2001) und die Subdimensionen werden dann, wie oben erwähnt, durch einen gewichteten Mittelwert zusammengefasst. Die Gewichtung der Subdimensionen im finalen Leistungswert ist damit eindeutig.

In TIMSS, PIRLS und PISA ist diese Gewichtung nicht so eindeutig und variiert von Test zu Test. Auch hier legen Experten der jeweiligen Fachgebiete eine Gewichtung für die Subdimensionen fest, die tatsächlichen Gewichtungen weichen jedoch von diesen ab, da die Leistungswerte durch eindimensionale IRT-Auswertungen berechnet werden und sich die Gewichtung der Subdimensionen dabei nach der Anzahl der Punkte richtet, die maximal in einer Subdimension erreicht werden kann[1]. Beispielhaft seien hier die Gewichtungen des PISA Mathematiktests von 2003 und 2012[2] betrachtet: Die Tests umfassen jeweils vier Subdimensionen, die gemäß Experten-Vorgabe mit jeweils 25% gewichtet sein sollten (OECD, 2012a, 2004). Tatsächlich jedoch variieren die Gewichtungen für die vier Subdimensionen in 2003 zwischen 22,8% und 30,4% und in 2012 zwischen 23,9% und 26,1%. Für den in 2012 zusätzlich neu eingeführten computerbasierten Test variieren sie zwischen 18,8% und 31,3% (OECD, 2005, 2012b). Grund für die Variation der Gewichtungen liegt in der Schwierigkeit, die maximale Punktzahl je Subdimension zum Zeitpunkt der Testerstellung vorherzusagen, da diese erst nach Analyse der Antwortdaten final festgelegt wird. So wurde in der PISA 2012 Hauptstudie, zum Beispiel, eine Mathematikaufgabe nachträglich aus der Wertung genommen, da man Bedenken hinsichtlich der einheitlichen Kodierung in den verschiedenen Ländern hatte. Für sechs Länder wurde

[1] Dies gilt so genau genommen nur für das Rasch-Modell (Rasch, 1980), das in PISA verwendet wird. Für das 2-PL-Modell (Birnbaum, 1968), das in TIMSS und PIRLS verwendet wird, hängt die Gewichtung außerdem zusätzlich von der Diskrimination der Aufgaben je Subdimension ab.

[2] In diesen beiden Tests war Mathematik der Schwerpunkt und umfasste die Schätzung von Leistungen in den Subdimensionen.

zudem zusätzlich jeweils eine Aufgabe aus der Bewertung herausgenommen (für jedes der Länder eine andere), da die berechneten Aufgabenschwierigkeiten in diesen Ländern überproportional abwichen (OECD, 2012b, S. 231–232). Je nachdem aus welchen Subdimensionen die Aufgaben stammen, verändern sich dementsprechend die Gewichtungen und können sich, wie im zuletzt genannten Fall, dann sogar auch leicht von Land zu Land unterscheiden. Ein weiterer Grund für Gewichtungsverschiebungen kann auch darin liegen, dass Aufgabenbewertungen nachträglich angepasst werden, etwa indem Aufgaben, für die zunächst 3 Antwortkategorien vorgesehen waren (0, 1 und 2 Punkte), im Nachhinein nur mit Hilfe von 2 Antwortkategorien (0 und 1 Punkt) kodiert werden, da die beobachteten Antworten nicht die erwartete Streuung aufwiesen.

Hinsichtlich der Gewichtungen in TIMSS und PIRLS ist anzumerken, dass diese neben den erreichbaren Punktzahlen zusätzlich von der durchschnittlichen Diskrimination der Aufgaben je Subdimension abhängen, da diese Studien auf das 2-PL -Modell (Birnbaum, 1968) vertrauen, das neben der Aufgabenschwierigkeit zusätzlich auch die Aufgabendiskrimination in die Berechnung der Leistungswerte miteingehen lässt (siehe auch Fußnote 1 oben).

Reliabilität

In NAEP werden die Subdimensionen stets separat betrachtet, was insbesondere auch den Feldtest zur Erprobung der Aufgaben einschließt. Die Aufgabenauswahl, die zu einem Großteil nach statistischen Kriterien vorgenommen wird, zielt dadurch auf eine Maximierung der Reliabilität zur Messung der Subdimensionen ab. Ein möglicher Nachteil dieses Ansatzes besteht dabei darin, dass die Reliabilität für den berechneten Mittelwert so geringer ist, denn die Aufgaben wurden nicht speziell ausgewählt, um das übergeordnete eindimensionale Konstrukt zu messen. Größter Nachteil ist jedoch, dass der gewählte Ansatz derzeit keine reliable Berechnung individueller Leistungswerte zulässt. Die in NAEP durchgeführte Berechung der Gesamtwerte basiert auf der sogenannten Plausible-Values-Technik, die eine schätzfehlerbefreite Berechnung von Leistungswerten auf Gruppenebene erlaubt, die Berechnung von reliablen individuellen Leistungswerten mit Hilfe von Schätzern wie WLE, MLE, oder EAP ist jedoch nicht möglich (für weitere Informationen zu diesen Schätzern siehe, z.B., Rost, 1996). Darüber hinaus eignet sich der Ansatz nicht, wenn die Auswertung mit Hilfe des Rasch-Modells erfolgen soll. Im Rahmen der Auswertung mit dem Rasch-Modell ist es nicht möglich, die Varianzen der einzelnen Subdimensionen in der Schätzung auf die gleiche Größe zu restringieren, wodurch die Standardisierung der Leistungswerte

nachträglich mit Hilfe der Punktschätzer der Varianzen durchgeführt werden muss und der Schätzfehler der Varianz damit in jeden einzelnen Leistungswert miteingeht.

In PISA, TIMSS und PIRLS liegt der Fokus der Testentwicklung eindeutig auf der Konstruktion eines eindimensionalen Leistungswerts. Die Aufgabenauswahl im Rahmen der gesamten Testentwicklung basiert auf dem Ziel, den finalen Test entsprechend des eindimensionalen Rasch-Modells zu optimieren[1]. Neben dem Vorteil reliable individuelle Schätzwerte berechnen zu können, bietet dieser Ansatz die Möglichkeit eine höhere Reliabilität für die eindimensionale Auswertung zu erlangen, da der Test dementsprechend optimiert ist. Diese Möglichkeit, eine höhere Reliabilität zu erreichen, wird allerdings dadurch konterkariert, dass bei einer eindimensionalen Auswertung (falls die Aufgaben, die einer gemeinsamen Subdimension angehören, einen größeren Zusammenhang zueinander haben als zu Aufgaben von anderen Subdimensionen - und nur dann macht die Auswertung von Subdimensionen Sinn) LID durch die Subdimensionen unberücksichtigt bleibt und die berechnete Reliabilität für das eindimensionale Konstrukt damit überschätzt wird (siehe Abschnitt „Lokale Abhängigkeiten“).

Validität

Neben den zuvor betrachteten eher technischen Kriterien ist es wichtig auch zu berücksichtigen inwieweit die geschätzten Leistungswerte valide sind, also dem von ihnen erhofften Nutzen entsprechen (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

Von den Subdimensionen des PISA Mathematiktest, zum Beispiel, erhofft man sich, die Schülerinnen und Schüler hinsichtlich der vier verschiedenen Bereiche von Mathematik zu unterscheiden, um Stärken und Schwächen genauer identifizieren zu können. Solch eine Betrachtung macht jedoch nur Sinn, wenn die Leistungen in den Subdimensionen sich tatsächlich unterscheiden. In NAEP wurden verschiedene Analysen zur Untersuchung der Dimensionalität der verwendeten Tests durchgeführt und man hat es als sinnvoll betrachtet, die Mathematik als mehrdimensionales Konstrukt zu betrachten (Allen u. a., 2001, S. 155–156). In den technischen Berichten von PISA, TIMSS und PIRLS werden keine Dimensionalitätsanalysen berichtet, die berichteten Ergebnisse geben jedoch einen Anhaltspunkt für die Unterschiede zwischen den Subdimensionen. Im PISA 2012 Mathematiktest erreichten die Niederlande einen Mittelwert von 532 Punkten in der

[1] TIMSS und PIRLS verwenden das 2-PL-Modell zur Auswertung der Hauptstudie, die Analyse der Aufgabeneigenschaften im Rahmen der Testentwicklung erfolgt jedoch auf Basis des Rasch-Modells (Martin & Mullis, 2012).

Subdimension *Unsicherheit und Daten* und einen Mittelwert von 507 Punkten in der Subdimension *Veränderung und Beziehungen* (OECD, 2014). Unabhängig von der statistischen Signifikanz, die angesichts der in PISA vorhanden Stichprobengrößen als gegeben betrachtet werden kann, entspricht der Unterschied einer Effektgröße von 0,25 gemäß Cohens *d*[1] und kann damit als bedeutsam eingeschätzt werden (Cohen, 1988). Es scheint daher sinnvoll auch für den PISA Mathematiktest Mehrdimensionalität anzunehmen bzw. mehrdimensionale Skalen zu berichten. Der Test wurde jedoch eindimensional konstruiert und von den Pilotierungen über den Feldtest bis hin zur Hauptstudie wurden die Aufgaben gemäß eines eindimensionalen Konstrukts ausgewählt. Hinsichtlich der Validität des Tests stellt sich damit eine wichtige Frage: Sind die im Rahmen der eindimensionalen Testkonstruktion ausgewählten Aufgaben vollständig repräsentativ für die mehrdimensionalen Konstrukte? Es könnte sein, dass die Ergebnisse in den Subdimensionen durch die Testkonstruktion verzerrt sind und die Validität dadurch vermindert ist.

Wenn die Subdimensionen sich wie angenommen unterscheiden, kann dies jedoch auch einen Einfluss auf die Validität des eindimensionalen Konstrukts haben. Alle betrachteten Studien verwenden ein sogenanntes rotiertes Testheftdesign, in dem nicht alle Schülerinnen und Schüler das gleiche Testheft erhalten, sondern jeweils unterschiedliche, bei denen sich immer nur Teilmengen der Aufgaben überschneiden. In PISA 2012, etwa, wurden insgesamt 13 verschiedene Testhefte für die Hauptsstudie eingesetzt, jedes zusammengesetzt aus vier Blöcken („Cluster“) mit jeweils Aufgaben für eine halbe Stunde Testzeit (die Gesamttestzeit je Heft betrug entsprechend 2 Stunden). Für Mathematik enthält der Test insgesamt sieben Blöcke, die auf die 13 Testhefte verteilt sind. Einige Testhefte enthalten dabei lediglich einen Mathematik-Block, andere bis zu drei (OECD, 2012b, S. 31). Betrachtet man einmal nur die Testhefte mit einem Mathematikblock (insgesamt vier Testhefte), so variieren die Gewichtungen der Subdimensionen bei diesen zwischen 7,7% und 40% (siehe Tabelle 2). Ein Schüler, der eine persönliche Stärke in *Veränderung und Beziehungen* hat und eine Schwäche in *Unsicherheit und Daten*, wird dementsprechend, je nachdem ob er zum Beispiel Testheft 8 oder Testheft 15 bearbeit, unterschiedliche Leistungswerte erreichen. Die individuellen Ergebnisse werden also durch das Testheftdesign verzerrt und weisen entsprechend geringe Validität auf. Auf Gruppenebene können die Ergebnisse bei ausreichender Gruppengröße trotzdem als valide betrachtet werden, da sich die Unterschiede in den Testheften dann ausmitteln. Es bleibt jedoch anzumerken, dass die Schätzung der Reliabilität des

[1] Die PISA-Skalen sind auf eine Standardabweichung von 100 standardisiert, ein Unterschied von 25 Punkten entspricht damit genau einem Cohens d von 0,25.

eindimensionalen Konstrukts unter der Annahme erfolgt, dass die beobachteten Antworten in den Mathematikblöcken unabhängig von eventuellen Unterschieden in den Gewichtungen der Subdimensionen sind. Vor diesem Hintergrund scheint es plausibel anzunehmen, dass das rotierte Testheftdesign zu einer zusätzlichen Überschätzung der Reliabilität führt (neben der im vorherigen Abschnitt beschriebenen Überschätzung durch LID). Eine detailliertere Betrachtung dieses Aspekts geht jedoch über das Thema dieser Arbeit hinaus.

Tabelle 2

Gewichtungen der Mathematik-Subdimensionen in den Testheften 2, 8, 12, und 13 in PISA 2012

Testheft	Veränderung und Beziehungen	Quantität	Raum und Form	Unsicherheit und Daten
2	30,8%	30,8%	7,7%	30,8%
8	15,4%	23,1%	23,1%	38,5%
12	18,2%	36,4%	27,3%	18,2%
13	40%	20%	26,7%	13,3%

Anm. Die Gewichte wurden gemäß der Kodierung und Klassifikation der Aufgaben in Anhang A des technischen Berichts zu PISA 2012 berechnet (OECD, 2012b).

Das Generalisierte Subdimensionsmodell

Um die oben beschriebenen Nachteile der jeweiligen Ansätze zu vermeiden, wird die Verwendung des generalisierten Subdimensionsmodells (GSM) vorgeschlagen. Durch eine Restriktion des mehrdimensionalen IRT-Modells, ermöglicht das GSM die Schätzung eindimensionaler Leistungswerte in Form von gewichteten Mittelwerten, die bereits in der Modellschätzung definiert enthalten sind. Damit erlaubt es die Berechnung reliabler Leistungswerte auf individueller Ebene (zum Beispiel von WLE oder MLE, siehe oben) und vermeidet gleichzeitig die Verzerrung der Schätzwerte durch die Vernachlässigung der lokalen Abhängigkeiten aufgrund der Mehrdimensionalität.

Die Entwicklung des GSM erfolgte in zwei Schritten. Das im ersten Schritt zunächst vorgeschlagene Subdimensionsmodell ist ebenfalls eine Restriktion des mehrdimensionalen Modells, in dem der eindimensionale Leistungswert als Mittelwert der Subdimensionen definiert ist. Es enthält jedoch eine implizite Restriktion der Varianzen, die dazu führt, dass das Subdimensionsmodell nur dann eine zum mehrdimensionalen Modell äquivalente

Passung besitzt, wenn die Varianzen der verschiedenen Dimensionen exakt gleich groß sind. Trotz dieser Einschränkung zeigt bereits die Anwendung des Subdimensionsmodells, dass das Modell die Verzerrung der geschätzten Werte verringert. In der Veröffentlichung „Estimating Tests Including Subtests“ (Kapitel 2; Brandt, 2010) wird hierzu zunächst anhand einer kleinen Simulationsstudie, angelehnt an den in PISA genutzten Ansatz, gezeigt, wie die geschätzten Parameter durch eine eindimensionale Skalierung ohne Berücksichtigung der Mehrdimensionalität verzerrt werden. Anhand eines Vergleichs der Schätzwerte des Subdimensionsmodells für den deutschen PISA 2003 Mathematiktest mit denen des ein- und des mehrdimensionalen Modells wird darüber hinaus gezeigt, dass, trotz der impliziten Restriktion der Varianzen, das Subdimensionsmodell eine deutlich bessere Passung als das eindimensionale Modell gewährt (der Unterschied zwischen der Modellpassung des ein- und des mehrdimensionalen Modells wird um mehr als Dreiviertel verringert) und auch die Verzerrung der Parameterschätzungen deutlich vermindert wird.

In der Veröffentlichung „Estimation of a Rasch Model Including Subdimensions“ (Kapitel 3; Brandt, 2008) wird das Subdimensionsmodell auf die Daten des TIMSS 2003 Mathematiktests der USA angewandt. Neben einer ausführlichen Beschreibung der Definition des Modells auf Basis des „Multidimensional Random Coefficients Multinomial Logit Model“ (MRCMLM) (Adams, Wilson, & Wang, 1997) zur Schätzung des Modells in ConQuest (Wu, Adams, & Wilson, 1998) werden die Ergebnisse des Subdimensionsmodells hier mit den Ergebnissen des Rasch Testlet Modells (Wang & Wilson, 2005) verglichen und es wird gezeigt, dass das Subdimensionsmodell für die gegebenen Daten eine bessere Modellpassung gewährt.

Eine weitere Anwendung des Subdimensionsmodells zeigt die Veröffentlichung „Robustness of Multidimensional Analyses Against Local Item Dependence“ (Kapitel 4; Brandt, 2012). In dieser wird im Rahmen einer Simulationsstudie untersucht, wie die Existenz von LID durch Testlets die Ergebnisse von mehrdimensionalen IRT-Analysen verzerrt. Da die Subdimensionen im Rahmen der Simulation mit äquivalenten Varianzen generiert wurden und die Ergebnisse des Subdimensionsmodells somit äquivalent zum mehrdimensionalen Modell sind, erlauben diese besondere Einblicke in die Veränderungen der (ein- bzw. mehrdimensionalen) Varianzanteile bei der mehrdimensionalen Schätzung mit LID.

Eine Anwendung des GSM wird in der Veröffentlichung „Increasing Unidimensional Measurement Precision Using a Multidimensional Item Response Model Approach“ (Kapitel 5; Brandt & Duckor, 2013) gezeigt. Im GSM wird für die Subdimensionen ein zusätzlicher

Parameter geschätzt, durch den die Varianzunterschiede in den Subdimensionen berücksichtigt werden, sodass die Modellpassung des GSM immer identisch zu der des mehrdimensionalen Modells ist. In der angesprochenen Veröffentlichung wird anhand eines Datensatzes zur Qualifikation angehender Lehrer zunächst gezeigt, wie die Parameter bei Verwendung des eindimensionalen IRT-Modells verzerrt werden und, dass die Standardfehler der eindimensionalen Leistungswerte in dem betrachteten Fall bei eindimensionaler Skalierung höher sind als bei der Skalierung mit Hilfe des GSM. Darüber hinaus wird gezeigt, welchen Einfluss die unterschiedlichen Gewichtungen der beiden Skalierungen haben: einerseits die Gleichgewichtung der Subdimensionen im GSM und andererseits die Gewichtung entsprechend der Anzahl der maximal erreichbaren Punkte je Subdimension. In Abhängigkeit von der gewählten Skalierung ergeben sich damit entsprechende Unterschiede für die Ergebnisse der einzelnen Lehramtstudentinnen und -studenten.

Einordnung des GSM im Vergleich zu anderen Modellen

Das Problem der Berechung von eindimensionalen Leistungswerten für als mehrdimensional angenommene Daten hat in den letzten Jahren zunehmend Aufmerksamkeit erhalten und auch zur Definition einer wachsenden Anzahl von Modellen geführt. Zunächst erfolgt daher eine kurze Charakterisierung dieser verschiedenen Modelle bzw. Modellklassen bevor genauer auf die Modelleigenschaften des GSM und seine Einordnung im Vergleich zu den bestehenden Modellen eingegangen wird.

Das Testlet Modell, das Higher-Order Modell und das hierarchische Modell

In Abhängigkeit davon, ob die betrachteten lokalen Abhängigkeiten als durch die Testkonstruktion verursacht angenommen werden oder als durch das zu messende psychologische Konstrukt, werden die Modelle, die versuchen diese lokalen Abhängigkeiten zu berücksichtigen, üblicherweise entweder als Testlet Modelle oder aber als Higher-Order oder hierarchische Modelle definiert.

In Testlet Modellen (Bradlow, Wainer, & Wang, 1999; Wang & Wilson, 2005) wird angenommen, dass die Beantwortung der Aufgaben grundsätzlich von einem eindimensionalen psychologischen Konstrukt abhängt, aber durch eine testlet-basierte Testkonstruktion trotzdem Mehrdimensionalität in den Testdaten enthalten ist, die misst inwieweit ein zu einer Aufgabe gegebener Stimulus einer Person die Antwort erleichtert oder erschwert. Aus Sicht des eindimensionalen Konstrukts entspricht dies LID. Hierarchische und Higher-Order Modelle (de la Torre & Song, 2009; Gibbons & Hedeker, 1992; Sheng & Wikle, 2008) auf der anderen Seite nehmen an, dass der Beantwortung der Aufgaben ein

mehrdimensionales psychologisches Konstrukt zugrunde liegt. Auch in diesem Fall existiert dadurch aus Sicht des übergeordneten eindimensionalen Konstrukts LID in den Daten, allerdings eben nicht aufgrund der Form des Tests, sondern aufgrund des gemessenen Konstrukts.

Trotz dieser Unterschiede in der Betrachtungsweise von LID sind die statistischen Annahmen der Modelle gleich. So haben Yung, Thissen und McLeod (1999) und Li, Bolt und Fu (2006) gezeigt, dass das Higher-Order Modell und das Testlet Modell Restriktionen des hierarchischen Modells sind. Rijman (2010) hat zudem gezeigt, dass das Testlet Modell und das Higher-Order Modell äquivalent zueinander sind. Eine grundlegende Annahme aller hierarchischen Modelle (also auch des Testlet und des Higher-Order Modells) ist dabei, dass Korrelationen zwischen den Subdimensionen (bzw. Testlets) vollständig durch das übergeordnete eindimensionale Konstrukt erklärt werden (Holzinger & Swineford, 1937).

Das GSM und das mehrdimensionale Modell

Eine der wichtigsten Eigenschaften von IRT-Modellen ist die Anzahl der Parameter, die sie nutzen. Desto mehr Parameter ein Modell zur Verfügung hat, desto besser wird es üblicherweise gegebene Daten modellieren können. Eine höhere Anzahl von Parametern führt in der Regel jedoch auch dazu, dass die Interpretation komplexer wird. Modelle mit weniger Parametern lassen daher meist klarere Aussagen zu. Gleichzeitig bedeuten weniger Parameter jedoch auch immer eine stärkere Restriktion des Modells und damit stärkere Annahmen hinsichtlich der Daten.

Aufgrund der zusätzlichen Annahmen, die die hierarchischen Modelle enthalten, werden in diesen weniger Parameter als im GSM und im mehrdimensionalen Modell geschätzt. Im GSM und im mehrdimensionalen Modell hingegen werden genau gleich viele Parameter geschätzt, lediglich die Restriktion der Parameter ist eine andere, wie auch aus Abbildung 1 ersichtlich wird. Im GSM wird die gleiche Anzahl an Dimensionen geschätzt wie im mehrdimensionalen Modell, anstelle einer der Subdimensionen wird im GSM jedoch eine zusätzliche übergeordnete sogenannte Hauptdimension, geschätzt. Die letzte fehlende Subdimension wird im GSM als Restriktion der anderen Subdimensionen geschätzt und zwar so, dass die Summe aller Subdimensionsparameter je Person genau null ist. Auf diese Weise definiert geben die Subdimensionsparameter die Leistungswerte in den Subdimensionen als Differenz zum Mittelwert an. Eine Eigenschaft des Mittelwerts[1] ist, dass die Kovarianz der Mittelwerte zu den Differenzwerten der Subdimensionen genau null ist. Die Kovarianzen

[1] Genauer gesagt eines gleichgewichteten Mittelwerts der über Subdimensionen gebildet wird, die auf identische Varianzen standardisiert sind, wie im Fall des GSM.

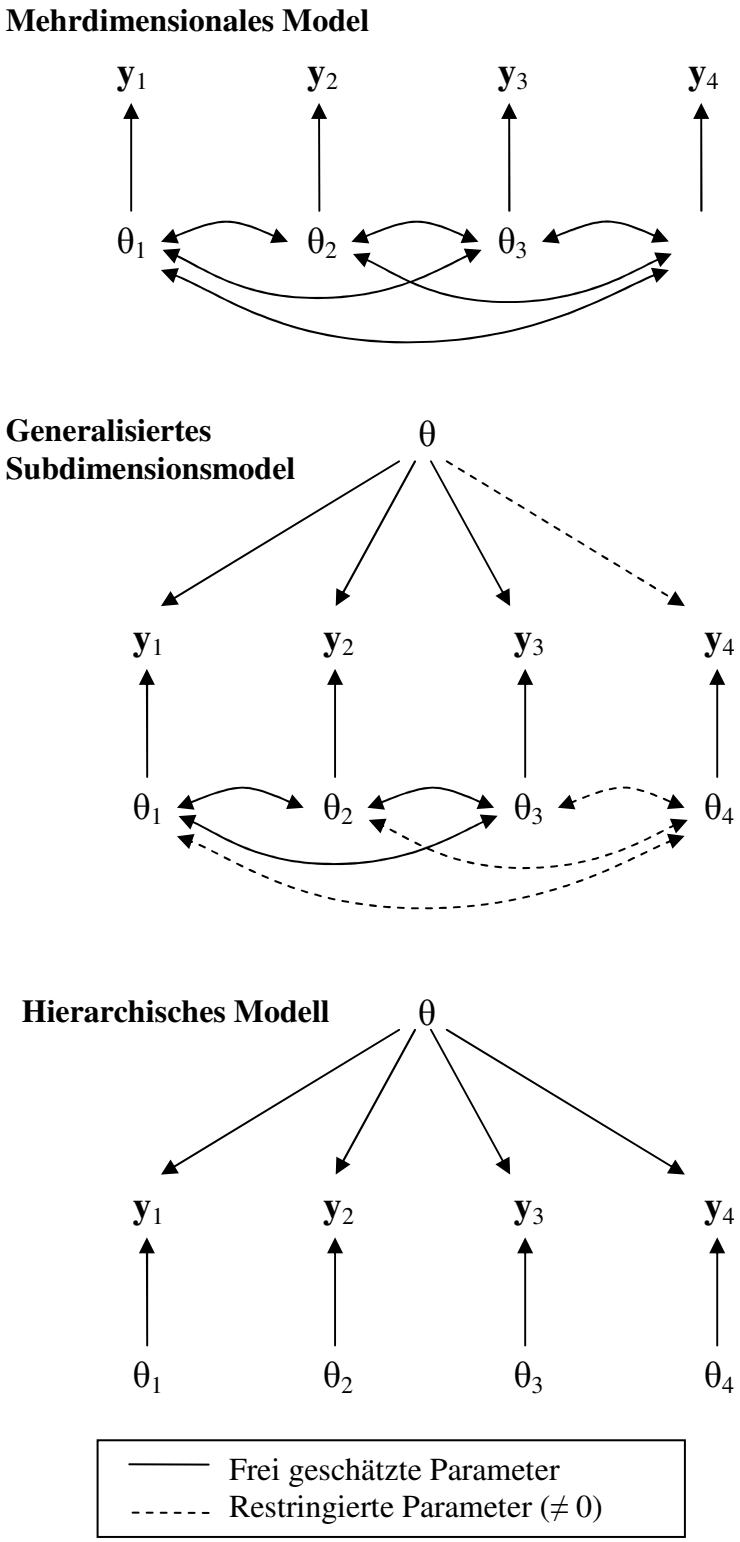

Abbildung 1. Graphische Darstellung des mehrdimensionalen Modells, des generalisierten Subdimensionsmodells und des hierarchischen Modells

zwischen Haupt- und Subdimensionsparametern sind im GSM daher auf null restringiert, was im übrigen auch der üblichen Restriktion in den hierarchischen Modellen entspricht. Im Gegensatz zu den hierarchischen Modellen erlaubt das GSM jedoch Korrelationen zwischen den Subdimensionsparametern. Die Anzahl der geschätzten Kovarianzparameter ist aber geringer als im mehrdimensionalen Modell, da für die letzte (restringierte) Subdimension keine Kovarianzparameter geschätzt werden müssen. Diese im Vergleich zum mehrdimensionalen Modell „frei“ gewordenen Parameter werden allerdings benötigt, um zusätzliche Parameter zur Standardisierung der Varianzen der Subdimensionen zu schätzen. Insgesamt lässt sich zeigen, dass die Anzahl der im GSM geschätzten Parameter immer exakt identisch zu der des mehrdimensionalen Modells ist (siehe Kapitel 6.2) und dass die geschätzten Parameter der beiden Modelle ineinander überführbar sind (siehe Kapitel 6.3).

Das GSM und das eindimensionale Modell

Für den Fall, dass die angenommene mehrdimensionale Struktur in einem Datensatz nicht existiert, sondern dieser tatsächlich eindimensional ist, werden die Varianzen der geschätzten Subdimensionsparameter null und die geschätzten Parameter der Hauptdimension entsprechen genau den Personenparametern der eindimensionalen IRT-Schätzung (siehe Kapitel 6.4).

Das GSM und das hierarchische Modell

Durch die unterschiedliche Restriktion der Parameter ist klar, dass die Berechnung der Parameter der Hauptdimension im GSM und im hierarchischen Model auf sehr unterschiedlichen Annahmen erfolgt. Im Gegensatz zum hierarchischen Modell erlaubt man dabei im GSM insbesondere, dass die Subdimensionen auch nach Kontrolle durch die Hauptdimension noch miteinander korrelieren können. Gemäß der Definition von Holzinger und Swineford (1937) entspricht das GSM damit einem modifizierten hierarchischen Modell, welches einem hierarchischen Modell mit überlappenden subdimensionsspezifischen Faktoren entspricht.

Fazit und Ausblick

Wie oben beschrieben entspricht das GSM einerseits einem restringierten mehrdimensionalen Modell und andererseits – für den Fall, dass keine Mehrdimensionalität in den Daten vorhanden ist – dem eindimensionalen Modell. Das GSM ermöglicht damit die eindimensionale Auswertung mehrdimensionaler Daten, ohne zusätzliche Annahmen über den Zusammenhang der Dimensionen zueinander. Die berechneten eindimensionalen Leistungswerte sind darüber hinaus unverzerrt durch LID aufgrund der Mehrdimensionalität,

die Gewichtung der Dimensionen ist klar definiert sowie unabhängig von Änderungen im Testdesign und die Schätzung erfolgt im Rahmen des IRT-Modells, wodurch auch reliable individuelle Leistungswerte, wie etwa der WLE, berechnet werden können.

Neben den eindimensionalen Leistungswerten bietet das GSM zusätzlich für jede Person einen Schätzwert für die Differenz der Leistung in einer Subdimension im Vergleich zur Leistung im Mittel. Da diese Schätzwerte im GSM standardisiert geschätzt werden, können sie direkt miteinander verglichen werden und sind ebenso schätzfehlerbefreit wie die Werte der übergeordneten Hauptdimension. Für Wissenschaftler sind diese Differenzwerte häufig interessant, um die unterschiedlichen Leistungen in den Subdimensionen unabhängig vom Leistungsniveau zu vergleichen und Leistungsprofile zu betrachten. Zudem besteht eine verbesserte Möglichkeit zu untersuchen, inwieweit sich die beobachteten Leistungen in einzelnen Subdimensionen tatsächlich unterscheiden, dass heißt, inwieweit es sinnvoll ist, die gegebenen Subdimensionen voneinander getrennt zu betrachten. Brandt, Duckor und Wilson (2014), etwa, nutzten das GSM, um auf diese Weise die Dimensionalität des Performance Assessment for California Teachers zu untersuchen.

Die bisherige Definition des GSM bezieht sich lediglich auf das mehrdimensionale Rasch Modell, eine Erweiterung der Definition auf das mehrdimensionale 2-PL Modell (Birnbaum, 1968) ist jedoch ohne Weiteres möglich. Eine interessante Anwendung für die Zukunft wäre daher, zum Beispiel, ein Vergleich der von NAEP berechneten Leistungswerte mit den auf Basis des GSM berechneten Leistungswerten. Ein Problem bei der Auswertung mit Hilfe des GSM war dabei bis vor kurzem, dass die Schätzung des Modells lediglich auf Basis von WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000)[1] möglich war. Neben der Komplexität der Definition von Modellen hat WinBUGS dabei den Nachteil, dass selbst für kleine Datensätze die Schätzung schnell sehr lange dauert. Seit kurzen liegt jedoch ein Update des R-Paketes TAM (Kiefer, Robitzsch, & Wu, 2015) vor, mit dessen Hilfe das GSM genauso schnell und effizient wie ein mehrdimensionales Standardmodell geschätzt werden kann und auch die Schätzung unter Berücksichtigung umfangreicher Regressionsmodelle möglich ist. Die Anwendung des GSM auf Daten der großen Leistungsstudien ist damit nun unproblematisch.

Neben der Anwendung des Modells zur Schätzung zuverlässiger eindimensionaler Leistungswerte ermöglicht das GSM aber wie oben bereits kurz beschrieben auch eine zusätzliche, differenzierte Betrachtung, um die Dimensionalität von Daten zu beurteilen.

[1] Im Gegensatz zum Subdimensionsmodell ist für das GSM die Schätzung auf Basis von ConQuest nicht möglich.

Bisher wurde über die Dimensionalität von Daten in der Regel durch den Vergleich der Passung verschiedener Modelle entschieden. Die zum Vergleich verwendeten Kriterien sind jedoch von der Anzahl der geschätzten Parameter als auch von der Anzahl der Personen abhängig und das Ergebnis des Modellvergleich ist damit letztendlich davon abhängig, wie die verschiedenen Einflussfaktoren zu einem eindeutigen Kriterium verrechnet werden. Je nach Definition bzw. Wahl des Kriteriums legt man in gewisser Weise damit auch das Ergebnis fest und eine wirklich objektive Entscheidung ist dementsprechend häufig nur schwer möglich. Das GSM ermöglicht nun, die Frage der Dimensionalität weg von der Frage der Modellpassung hin zur Frage der Sinnhaftigkeit der Unterscheidung von Subdimensionen umformulieren: Inwieweit gibt es eine relevante Anzahl von Personen, die sich in der Leistung zwischen zwei Subdimensionen unterscheiden bzw. inwieweit ist die Unterscheidung zweier Subdimension nützlich? Durch diese veränderte Betrachtung wird gleichzeitig auch eine direkte Beziehung zur Validität der Dimensionalität hergestellt (American Educational Research Association u. a., 2014; vgl. Brandt u. a., 2014).

Literatur

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

Allen, N. L., & Carlson, J. E. (1987). Scaling Procedures. In A. E. Beaton (Hrsg.), *The NAEP 1983-1984 technical report*. Princeton, NJ: Educational Testing Service.

Allen, N. L., Carlson, J. E., & Donoghue, J. R. (2001). Overview of part II: the analysis of 1998 NAEP data. In N. L. Allen, J. R. Donoghue, & T. L. Schoeps (Hrsg.), *The NAEP 1998 Technical Report* (S. 143–160). Washington, D. C.: National Center for Education Statistics.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Hrsg.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Brandt, S. (2006). *Exploring bundle dependencies for the embedded attitudinal items in PISA 2006*. Gehalten auf der 13th meeting of the International Objective Measurement Workshop (IOMW), Berkeley, CA.

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. In M. von Davier & D. Hastedt (Hrsg.), *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments* (Bd. 1, S. 51–70). Princeton, NJ: IEA-ETS Research Institute.

Brandt, S. (2010). Estimating tests including subtests. *Journal of Applied Measurement*, *11*, 352–367.

Brandt, S. (2012). Robustness of multidimensional analyses against local item dependence. *Psychological Test and Assessment Modeling*, *54*, 36–53.

Brandt, S., & Duckor, B. (2013). Increasing unidimensional measurement precision using a multidimensional item response model approach. *Psychological Test and Assessment Modeling*, *55*, 148-161.

Brandt, S., Duckor, B., & Wilson, M. (2014). *A utility-based validation study for the dimensionality of the performance assessment for california teachers*. Gehalten auf

der 2014 annual conference of the American Educational Research Association (AERA), Philadelphia, PA.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Aufl.). Hillsdale, NJ: Erlbaum.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*(8), 620–639.

Donahue, P. L., & Schoeps, T. L. (2001). Assessment frameworks and instruments for the 1998 national and state reading assessments. In N. L. Allen, J. R. Donoghue, & T. L. Schoeps (Hrsg.), *The NAEP 1998 Technical Report* (S. 255–268). Washington, D. C.: National Center for Education Statistics.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*, 41–54.

Kiefer, T., Robitzsch, A., & Wu, M. (2015). TAM: test analysis modules (Version 1.3) [R]. Abgerufen von http://cran.r-project.org/package=TAM

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*, 3–21.

Lunn, D. J., Thomas, A., Best, N. G., & Spiegelhalter, D. J. (2000). WinBugs - A Bayesian modelling framework: Concepts, structure, and extensibilty. *Statistics and Computing*, *10*(4), 325–337.

Martin, M. O., & Mullis, I. V. S. (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). PISA test format assessment and the local independence assumption. *IERI monograph series–Issues and methodologies in large scale assessments*, *4*, 131–158.

Moosbrugger, H., & Kelava, A. (2007). *Testtheorie und Fragebogenkonstruktion*. Heidelberg: Springer Medizin Verlag.

OECD. (2005). *PISA 2003 technical report*. Paris: OECD.

OECD. (2012a). *PISA 2012 assessment and analytical framework: mathematics, reading, science, problem solving and financial literacy*. OECD Publishing.

OECD. (2012b). *PISA 2012 technical report*. Paris: OECD.

OECD. (2014). *PISA 2012 results: what students know and can do* (Bd. 1, revised edition). Paris: Organisation for Economic Co-operation and Development.

OECD, O. for E. (2004). *The PISA 2003 assessment framework: mathematics, reading, science and problem solving knowledge and skills*. OECD Publishing.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*, 361–372.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion*. Göttingen: Verlag Hans Huber.

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, *68*, 413–430.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, *6*, 181–195.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*, 126–149.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER ConQuest: generalized item response modeling software*. Melbourne, Australia: Australian Council for Educational Research.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.

Yen, W. M. (1993). Scaling performance assessments - strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213.

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model, *64*, 113–128.

Contents

Zusammenfassung ... 3
Figures ... 25
Tables ... 26
1 Introduction ... 27
1.1 Advantages and Disadvantages of the Approaches of Current Large Scale Assessments to Construct Unidimensional Scores ... 29
1.1.1 Local Item Dependence ... 29
1.1.2 Subdimension Weighting ... 30
1.1.3 Reliability ... 32
1.1.4 Validity ... 33
1.2 Merging the Currently Used Approaches: The Generalized Subdimension Model 35
1.3 References ... 37
2 Estimating Tests Including Subtests (Brandt, S. (2010). *Journal of Applied Measurement*, *11*, 352–367.) ... 37
2.1 Problems Considering the Construction and Calibration of Tests Including Subtests 41
2.1.1 Local Item Dependence ... 41
2.1.2 Minimization of Subdomain Differences ... 42
2.1.3 Bias of the Multidimensional Person Parameter Estimates ... 42
2.1.4 Booklet DIF ... 45
2.2 The Rasch Subdimension Model ... 46
2.3 An Empirical Example ... 49
2.3.1 Data and Analysis ... 51
2.3.2 Results ... 51
2.3.3 Discussion ... 51
2.4 Conclusion ... 54
2.5 References ... 56
2.6 Appendices ... 58
3 Estimation of a Rasch Model Including Subdimensions (Brandt, S. (2008). In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments* (Vol. 1, pp. 51–70). Princeton, NJ: IEA-ETS Research Institute.) ... 62
3.1 Introduction ... 62
3.2 A Rasch Model that Includes Subdimensions ... 63

3.2.1 Model Definition ... 64
3.2.2 Discussion of the Model ... 67
3.3 Estimation Using ConQuest ... 69
3.4 An Empirical Example ... 74
3.4.1 Data and Analysis ... 75
3.4.2 Results and Discussion ... 75
3.5 Conclusion ... 78
3.6 References ... 81
4 Robustness of Multidimensional Analyses against Local Item Dependence (Brandt, S. (2012). *Psychological Test and Assessment Modeling*, *54*, 36–53.) ... 84
4.1 Preparatory Considerations for the Design of the Simulation Study ... 85
4.2 Simulation Study Design ... 88
4.3 Data Generation and Analysis ... 89
4.4 Results ... 91
4.5 Discussion ... 95
4.6 Conclusion ... 97
4.7 References ... 99
4.8 Appendix A ... 101
4.9 Appendix B ... 103
4.10 Appendix C ... 104
5 Increasing Unidimensional Measurement Precision Using a Multidimensional Item Response Model Approach (Brandt, S., & Duckor, B. (2013). *Psychological Test and Assessment Modeling*, *55*, 148.) ... 105
5.1 Background and Context of the CAL Scale ... 108
5.2 Method ... 110
5.2.1 Data ... 110
5.2.2 Model Definition ... 111
5.2.3 Estimation ... 112
5.3 Results and Discussion ... 113
5.4 Conclusion ... 117
5.5 References ... 119
6 Discussion ... 122
6.1 The Testlet, the Higher Order, and the Hierarchical Model ... 122
6.2 Estimated Parameters ... 123

6.3 Equivalence With the Multidimensional Model 126
6.4 Correspondence With the Unidimensional Model 127
6.5 Relationship to the Hierarchical Model 128
6.6 Conclusion 129
6.7 Prospect of Current and Future Research 129
6.8 References 131

Figures

Figure 2.1. Item Parameter Distribution for Dimension 1 and 2 of the Simulated Test.....43
Figure 2.2. Relationship between the True Parameters and the Estimated Parameters Using the Unidimensional Model.....44
Figure 2.3. Relationship between the Difficulty Estimates Obtained Using the Multidimensional Model and the Unidimensional Model.....52
Figure 2.4. Relationship between the Difficulty Estimates Obtained Using the Multidimensional Model and the Subdimension Model[1] 52
Figure 4.1. Depiction of a three-dimensional sequential test design (on the left) and a three-dimensional parallel test design (on the right).....88
Figure 4.2. Likelihoods for the unidimensional and multidimensional estimations of the four-dimensional construct with correlations of .5, .7, and .9; a sequential or a parallel test design; and varying extents of item bundle effects.....93
Figure 4.3. Estimated correlations for the four-dimensional construct with generated correlations of .5, .7, and .9; a sequential or a parallel test design; and varying extents of item bundle effects.....94
Figure 4.4. Depiction of the dimensions' unidimensional (main dimension = main) and dimension specific (subdimension = sub) variance components depending on the generated correlations, a sequential or parallel test design, and the extent of item bundle effects.....96
Figure 5.1. The three major domains of the modified assessment triangle framework: Cognition and Learning Targets (CLT), the Assessment Strategies and Tools (AST), and Evidence and Data Interpretation (EDI).....109
Figure 5.2. Comparison of the item estimates for the CLT, AST, and EDI dimension using a unidimensional calibration and a GSM calibration.....114
Figure 5.3. Comparison of the standard errors of the unidimensional person parameter estimates from the unidimensional model and from the generalized subdimension model and from the composed mean score of the multidimensional person parameter estimates.....115

Tables

Table 1.1 Pros and Cons of a Composite Score and Unidimensional Scaling to Obtain a Unidimensional Score for Data Assumed to be Multidimensional 28
Table 1.2 Mathematics Score Distributions by Subdimension for the PISA 2003 Paper-Based Main Survey and the PISA 2012 Paper-Based and Computer-Based Main Survey 31
Table 1.3 Weights of the Mathematics Subdimensions in the Booklets 2, 8, 12, and 13 in PISA 2012 35
Table 2.1 Hotelling's T^2 Tests for the Overall Null Hypothesis of Unbiased Estimation 44
Table 2.2 Parameter Recovery for the Person Parameter Estimates of the Non-Anchored and Anchored 2-Dimensional Calibration 45
Table 2.3 Results of the Re-Analysis of the German PISA 2003 Mathematics Test 50
Table 2.4 Item Parameter Recovery for the Non-Anchored and Anchored 2-Dimensional Calibration for a Subtest Correlation of 0.6 58
Table 2.5 Item Parameter Estimates for the Re-Analysis of the PISA 2003 Mathematics Achievement Test 60
Table 3.1 Results of the re-analysis of the US TIMSS 2003 mathematics achievement test .. 76
Table 4.1 Calibration Results for the German Subsample of the Mathematics Achievement Test of PISA 2003 Using the Rasch Testlet Model 87
Table 5.1 Multidimensional Estimation Results 113
Table 5.2 Weights of the Subdimensions 116
Table 5.3 Comparison of Two Students 116

1 Introduction

Historically questionnaires and achievement tests[1] have been analyzed applying methods based on classical test theory (CTT), using, for example, sum scores or mean scores to interpret results, and a large part of today's analyses still bases on these methods. However, CTT has important disadvantages: (a) CTT does not include a theory about item difficulties and thereby limits the investigation of test characteristics as well as the comparison, or linkage, of test results from tests with different sets of items; and (b) CTT includes very strong assumptions on the characteristics of the test that is analyzed (see, e.g., Moosbrugger & Kelava, 2007; Rost, 1996).

As a way to avoid these disadvantages the item response theory (IRT) was developed (see, e.g., Moosbrugger & Kelava, 2007; Rost, 1996), and today all important national and international studies base on IRT analyses. However, also IRT still includes many assumptions, which are often difficult to meet. An important and very basic assumption is on the given dimensionality of a test. In IRT it is true that a test has to be unidimensional in order to be interpreted unidimensionally. This seems to be a redundant notation, however, as a matter of fact in many cases one and the same test is interpreted unidimensionally as well as multidimensionally. In the Programme for International Student Assessment (PISA), for example, a unidimensional mathematics score is reported and at the same time scores in the four subdimensions *Change and Relationships*, *Quantity*, *Space and Shape*, and *Uncertainty and Data*, assuming that mathematics is a multidimensional construct. The same holds for the reading and for the science construct investigated in PISA (OECD, 2012b) and for similar constructs investigated by other international studies, such as the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) (Martin & Mullis, 2012). Apparently, this is a contradiction to a basic assumption of IRT. Nonetheless, the practical necessity of yielding unidimensional and multidimensional results prevails the necessities of the theory. It is remarkable, however, that neither in PISA, in TIMSS, nor in PIRLS this theoretical contradiction is discussed and that its possible effects are ignored.

A different approach takes the National Assessment of Educational Progress (NAEP). Here, the reading ability, for example, is composed of Reading for Literary Experience, Reading to Gain Information, and Reading to Perform a Task (Donahue & Schoeps, 2001). The scores for these three subdimensions of reading are calibrated using a (three-)multidimensional IRT model. The comprehensive reading score, however, is

[1] In the following both "questionnaires" and "achievement tests" will simply be denoted as tests.

calculated as a weighted mean score based on the estimation results from the calibration of the subdimensions (Allen, Carlson, & Donoghue, 2001). Both the approach of using a scale score from a unidimensional IRT model and the approach of using a composite score based on a multidimensional calibration have their advantages and disadvantages (cf. Table 1.1), which are discussed in the next section. Following this discussion a new approach is presented: the Generalized Subdimension Model (GSM). The GSM represents a combination of the two currently used approaches, a restriction of a multidimensional IRT model that yields a weighted mean score for the comprehensive dimension as an additional parameter of the model.

Table 1.1

Pros and Cons of Obtaining a Unidimensional Score for Multidimensional Data Via a Composite Score Versus a Unidimensional Calibration

Unidimensional Calibration	
Pro	Con
• Calculation of maximum likelihood estimates is possible (WLE, MLE, ...)	• Overestimation of reliability due to neglected local item dependence (LID) • Questionable validity of the multidimensional construct if the test was constructed to be unidimensional • Implicit and unclear weighting of the unidimensional score
Composite Score Based on Multidimensional Calibration	
Pro	Con
• Consideration of LID due to the multidimensional structure and therfore more appropriate reliability estimates • Explicit and clear weighting of the unidimensional score	• Calculation of maximum likelihood estimates is *not* possible • Not appropriate for the Rasch model (standardization of the multidimensional scores leads to increased measurement error)

1.1 Advantages and Disadvantages of the Approaches of Current Large Scale Assessments to Construct Unidimensional Scores

NAEP, TIMSS, PIRLS, and PISA are all large scale assessments that have been conducted on a regular bases over the last two decades. Due to the strong political interest in the results and the considerable attention these studies receive, not only from the public but also from the scientific community, it is a natural obligation for them to conduct the studies always at the current state of the art in measurement. Due to the granted financial resources, the studies are sometimes even capable of contributing significant developments to measurement research. NAEP, for example, was responsible for the introduction of the plausible value approach into measurement, which today is a standard technique to calculate unbiased group-level estimates (Beaton, 1987; Von Davier, Gonzalez, & Mislevy, 2009).

The discussion of the approaches used by these large scale assessments is therefore considered as a good way to provide an overview of the current state of the art in calculating unidimensional scores for multidimensional data.

1.1.1 Local Item Dependence

A basic underlying assumption of IRT models is local item independence (LII). LII describes the fact that the observed answers for a test are assumed to be conditionally independent given the individuals' scores on the latent variable that is measured. The violation of this assumption is denoted as local item dependence (LID). LID can occur due to various reasons, the most commonly considered is probably LID due to testlets, or item bundles. Testlets refer to items that share a common stimulus. They are popular because they allow a more economic use of the testing time; by answering several items to a single stimulus, persons need less answer time per item in comparison to reading a new stimulus for each item. However, there are drawbacks. If a person correctly answers one item of a given stimulus, the probability that he will answer an item of the same stimulus correctly is often slightly higher than the probability of answering correctly to an item from a different stimulus, in which an item has already been answered incorrectly. That is, the items show LID. The same holds for subdimensions; if a given dimension is assumed to comprise subdimensions, it is assumed that the items pertaining to a common subdimension are stronger related with each other than with items from other subdimensions.

The effects of LID in IRT analyses have been investigated by numerous authors. Unanimously, it is stated that ignoring LID leads to a biased estimation of the difficulty parameters, an overestimation of item discrimination, a bias on the variance estimate, and an

overestimation of reliability (see, e.g., Monseur, Baye, Lafontaine, & Quittre, 2011; Tuerlinckx & De Boeck, 2001; Wainer, Bradlow, & Wang, 2007; Wang & Wilson, 2005; Yen, 1984).

The considered large scale assessments take the LII assumption differently into account. In NAEP, TIMSS, and PIRLS the tests for mathematics, for example, simply do not use item bundles but each item is provided via an individual stimulus. Only in PISA item bundles are used for the mathematics test. For the reading ability test, on the other hand, all studies use item bundles. In NAEP the potential LID is, therefore, investigated using available LID indices and when necessary items are collapsed to a single item to avoid local item dependence[1] (Allen & Carlson, 1987, p. 236–237). In the technical reports of TIMSS, PIRLS, and PISA possible local item dependencies are neither mentioned nor discussed. It is known however that the item bundles used in PISA, for example, result in LID for the respective items (Brandt, 2006; Monseur et al., 2011).

Considering LID due to subdimensions NAEP also follows a different approach. In NAEP a multidimensional IRT model is used to calibrate plausible values for each person and each subdimension, and the comprehensive scores across the subdimensions are calculated as weighted means of the plausible values (Allen et al., 2001, p. 155). This way possible negative effects on the IRT calibration due to LID by the subdimensions is avoided. In TIMSS, PIRLS, and PISA the subdimensions are calibrated jointly, using a unidimensional IRT model; possible effects due to LID are not considered.

1.1.2 Subdimension Weighting

In NAEP the subdimensions are calibrated as separate dimensions and subject experts provide a weight for each subdimension within the overarching comprehensive dimension. This way, the score for Reading in Grade 12, for example, is composed by considering *Reading for Literary Experience* with a weight of 35%, *Reading to Gain Information* with a weight of 45%, and *Reading to Perform a Task* with a weight of 20% (Donahue & Schoeps, 2001).

[1] Collapsing items into a single item is one possible strategy to avoid LID, the drawback, however, is a loss in information since only the sum score of the items is considered in the IRT model and not the individual scores of the respective items any more (Yen, 1993).

Table 1.2

Mathematics Score Distributions by Subdimension for the PISA 2003 Paper-Based Main Survey and the PISA 2012 Paper-Based and Computer-Based Main Survey

Subdimension	PISA 2003 Paper-Based	PISA 2012 Paper-Based	PISA 2012 Computer-Based
Change and Relationships	30.4% (28 points)	26.1% (24 points)	29.2 (14 points)
Quantity	23.9% (22 points)	23.9% (22 points)	20.8% (10 points)
Space and Shape	22.8% (21 points)	25.0% (23 points)	31.3% (15 points)
Uncertainty and Data	22.8% (21 points)	25.0% (23 points)	18.8% (9 points)
Total	100% (92 points)	100% (92 points)	100% (48 points)

Note. The scores were calculated based on the item classifications given in Appendix 12 of the PISA 2003 Technical Report and Annex A of the PISA 2012 Technical Report (OECD, 2005, 2012b).

In TIMSS, PIRLS, and PISA the weights of the subdimensions are less clear, and in fact vary from test to test. In order to demonstrate this, the weights for the PISA mathematics test are exemplarily depicted in Table 1.2. The IRT model used in PISA to calibrate the achievement scales is the Rasch model (Rasch, 1980). In the Rasch model, the weight of an item corresponds to the maximum score achievable for this item. Dividing the maximum score for each subdimension by the total score achievable therefore provides the weights of the subdimensions. In PISA 2003 as well as in PISA 2012 the assessment framework for the mathematics tests (in these two PISA cycles mathematics was the focus domain and included the estimation of the subdimensions) specified an equal weighting of 25% for the subdimensions *Change and Relationships*, *Quantity*, *Space and Shape*, and *Uncertainty and Data* (OECD, 2012a, 2004). The actual weightings, however, varied between 22,8% and 30,4% in PISA 2003, between 23,9% and 26,1% for the paper-based test in PISA 2012, and for the computer-based assessment in PISA 2012 between 18,8% and 31,3%. An important reason for the variations in the weightings is the fact that it is very difficult to predict the final total score of a subdimension at the moment the test is administered. The final scores are fixed only after knowing the answer data and the resulting item characteristics. For PISA

2012 for example—even though all items went through an extensive field trial—, a mathematics item included in the main trial was deleted because of concerns regarding the consistency with which the intended coding rule was applied across countries, and for six countries an item was deleted on the national level because the item (in each case a different one) showed a difficulty that was inconsistent with the difficulty observed across the remaining countries (OECD, 2012b, p. 231–232). As a consequence the deleted items' subdimensions will have a smaller weight within the total score. In fact, considering the weightings of the subdimensions for the six countries with national deletions, these will even slightly differ from the weighting for the remaining countries. Another reason for the final maximum score to change might be that an item that was administered to differentiate the persons in three scoring categories (0 points, 1 point, and 2 points) does not show sufficient variability in the answers, so the scoring categories are collapsed to just two (0 points and 1 point).

Considering the final weightings in TIMSS and PIRLS a prior definition of the actual weights is even more complicated since these two studies use a 2-PL IRT model (Birnbaum, 1968). While the Rasch model only estimates a difficulty parameter for each item, the 2-PL model additionally estimates a discrimination parameter for each item, which allows modeling the data more accurate and increasing the reliability. The drawback, however, is that the items obtain different weights for the calibration of the final score, where items with higher discriminations get a higher weight and items with a lower discrimination a lower weight. Correspondingly, the weight of a subdimension also changes if the average discrimination of its items is above or below the average of the total test.

1.1.3 Reliability

In NAEP the IRT analyses consider the subdimensions always separately. That is, also the item fit statistics calculated for the field trial, for example, are based on the results in the respective subdimensions. Hence, the psychometric item selection process is aiming at a maximization of the reliability for the measurement of the subdimensions. A possible disadvantage of this multidimensional test construction might be that the reliability of the overarching comprehensive score is reduced since the items are not fitted to be on a common scale. The more serious disadvantage, however, is probably that the corresponding approach does not allow the calculation of reliable individual scores. The plausible values allow calculating reliable group-level results for the comprehensive scores, but the calculation of individual point estimates, such as the WLE, MLE, or EAP is not possible (for more details

on these estimates see, e.g., Rost, 1996). The approach is therefore only appropriate if the calculation of individual estimates is not necessary. Furthermore, the approach is not appropriate if the test is analyzed using the Rasch model. In contrast to the 2-PL model, the calibration of the Rasch model does not allow constraining subdimensions to equal variances. Therefore, in the case of the Rasch model the plausible values have to be standardized after the calibration of the model, using the estimated variances of the subdimensions. In doing so however, the standard error of the variance estimate of a subdimension will be added to each plausible value of that subdimension, and the plausible values loose their beneficial characteristic of being unbiased due to the estimation.

In PISA, TIMSS, and PIRLS the focus is clearly on a unidimensional test construction. The item selection process is based on the unidimensional Rasch model[1] and aimed at maximizing the reliability of the overarching comprehensive scores. The advantage of this approach is that it provides the option to calculate reliable individual scores. Furthermore, fitting the items to the comprehensive scale might yield a higher reliability for this scale. However, if the items attributed to a particular subdimension have something particular in common—and hence include LID due to the subdimensions—, the observed reliability will be biased and higher than it actually is (see section 1.1.1).

1.1.4 Validity

Besides the technical requirements of measurement considered in the above sections, it is also important to consider to what extent the constructed measures are valid, that is, yield the intended use (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). The intended use for the subdimensions in mathematics, for example, is to differentiate the persons' achievements in specific areas of mathematics in order to identify weaknesses and strengths within the area of mathematics. Such a distinction is only useful, however, if the subdimensions actually differ. In NAEP several dimensionality analyses were conducted, and it was decided that it is reasonable to interpret mathematics as a multidimensional construct (Allen et al., 2001, p. 155–156). The technical reports of PISA, TIMSS, and PIRLS do not include any results from analyses investigating the assumed dimensional structures of the mathematics, reading, or science scales. Here instead the reported results for the subdimensions can be taken as an indication for the given differences between the subdimensions. In the PISA 2012

[1] TIMSS and PIRLS use the 2-PL model to calibrate the final results, the analysis of the item characteristics and the item selection process, however, is based on the Rasch model (Martin & Mullis, 2012).

mathematics achievement test, for example, the Netherlands achieved a mean score of 532 points for the subdimension *Uncertainty and Data* and a score of 507 points for *Space and Shape* (OECD, 2014). Besides the statistical significance, which can be assumed as given considering the sample sizes in PISA, the difference corresponds to an effect size of 0.25 using Cohen's d[1], and can therefore also be considered as meaningful (Cohen, 1988). That is, in this case for the PISA mathematics test, the results indicate that the subdimensions in fact measure different aptitudes and, hence, are multidimensional. However, we have to consider that the test was constructed to be unidimensional; in fact, across all pilot studies, the field trial, and also the main survey, the items were constructed and selected in order to fit on a common unidimensional scale. The important question for the validity of the multidimensional results therefore is: Are the items selected to measure the subdimensions representative for the actually defined aptitudes? It might well be that the results for the subdimensions are biased due to the test construction and that therefore the validity of the subdimension results is reduced.

For a different reason also the validity of the unidimensional results might be affected if the subdimensions actually differ (and only then it makes sense to report them separately). All considered large scale assessments use a rotated booklet design, in which not all items are administered to all persons but each person answers only to a sample of items. In PISA 2012, for example, the main survey includes 13 different booklets, each composed of four, so-called, item clusters, which include items for 30 minutes of testing time (that is, each booklet includes items for 2 hours of testing time). For mathematics the test includes seven different clusters distributed across the 13 booklets, with some booklets including just one of these clusters and others including up to three (OECD, 2012b, p. 31). The booklets 2, 8, 12, and 13 each include only one cluster. Table 1.3 shows the weights of the subdimensions within these clusters. The weights vary from 7,7% to 40%, which makes clear that a person, for example, with a relative strength in *Change and Relationships* and a relative weakness in *Uncertainty and Data* will receive a different score depending on whether he completed booklet 8 or booklet 13. If the test of mathematics includes multidimensionality due to the subdimensions, it will therefore not yield valid results for the comprehensive achievement in mathematics on the individual level. Considering group-level results these can still be considered as valid since the booklet differences will be equaled out if the groups are sufficiently large. However, since the unidimensional calibration assumes that the individual scores in mathematics are

[1] The PISA scales are standardized to a standard deviation of 100; the difference of 25 points therefore corresponds to a Cohen's d of 0.25.

independent of the booklets weightings according to the subdimensions, it is plausible to assume that the rotated booklet design leads to an overestimation of the unidimensional reliability estimate (additional to the overestimation of the reliability due to LID described in the previous section). A more detailed consideration of this aspect is beyond the scope of this work, though.

Table 1.3

Weights of the Mathematics Subdimensions in the Booklets 2, 8, 12, and 13 in PISA 2012

Booklet	Change and Relationships	Quantity	Space and Shape	Uncertainty and Data
2	30.8%	30.8%	7.7%	30.8%
8	15.4%	23.1%	23.1%	38.5%
12	18.2%	36.4%	27.3%	18.2%
13	40.0%	20.0%	26.7%	13.3%

Note. The weights were calculated based on the score and item classifications given in Annex A of the PISA 2012 Technical Report (OECD, 2012b).

1.2 Merging the Currently Used Approaches: The Generalized Subdimension Model

In order avoid the above described disadvantages of the respective approaches currently used in large scale assessments, the Generalized Subdimension Model was developed (GSM). The GSM combines the two approaches by restricting a multidimensional IRT model to yield an additional weighted comprehensive score. This way, the model allows calculating reliable individual scores and at the same time avoids the described problems by inappropriately assuming unidimensionality. The development of the model was conducted in two steps. In the first step, the subdimension model was developed. This earlier version of the GSM also allows calculating a weighted mean score based on a restriction of the multidimensional model, however, it includes a hidden constraint on the variances of the subdimensions. Therefore, the Subdimension Model only shows a fit equal to the multidimensional model if the variances of the subdimensions are equal. In the second step the Subdimension Model was generalized to the GSM by incorporating an additional parameter type considering the subdimensions' variance differences.

In the next chapter, which corresponds to a publication in the *Journal of Applied Measurement* (Brandt, 2010), the characteristics of the Subdimension Model are described.

First, some of the mentioned disadvantages of the unidimensional calibration approach are demonstrated via a small simulation study. Then, the Subdimension Model is applied to the German PISA 2003 mathematics achievement test, and its results are contrasted with those of the unidimensional and the multidimensional model in order to investigate to what extent the Subdimension Model is able to model the LID due to the subdimensions.

Chapter 3 corresponds to a publication in *Issues and Methodologies in Large-Scale Assessments* (Brandt, 2008). Here, the capability of the Subdimension Model to consider LID due to subdimensions is investigated for the US TIMSS 2003 mathematics achievement test, and the results are additionally compared to those from the Rasch testlet model (Wang & Wilson, 2005), in which the subdimensions are considered as testlets. Furthermore, a detailed description to estimate the Subdimension Model using the multidimensional random coefficients multinomial logit model (MRCMLM) (Adams, Wilson, & Wang, 1997) implemented in ConQuest (Wu, Adams, & Wilson, 1998) is given.

Chapter 4 shows a further application of the Subdimension model. Here, it is investigated via a simulation study how LID due to item bundles biases the results of multidimensional analyses. The results of the Subdimension model are particularly useful in this case since it yields valuable insight by separating the different unidimensional and multidimensional variance components. The reported results correspond to a publication in the journal *Psychological Test and Assessment Modeling* (Brandt, 2012).

The Generalized Subdimension Model is introduced in chapter 5, which also corresponds to a publication in the journal *Psychological Test and Assessment Modeling* (Brandt & Duckor, 2013). After its description, an application of the GSM to an empirical example is given, and by comparing the results to those of the unidimensional model, the bias of the item parameter estimates using a unidimensional IRT model is considered as well as the difference in the resulting standard errors for the unidimensional scale scores. Furthermore, the difference in the weightings of the two models and its impact on the individual scale scores are considered.

The last chapter discusses the theoretical characteristics of the GSM and provides information on how to classify the GSM in comparison to other currently existing models. It concludes with some final remarks on current and possible future research.

1.3 References

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

Allen, N. L., & Carlson, J. E. (1987). Scaling Procedures. In A. E. Beaton (Ed.), *The NAEP 1983-1984 technical report*. Princeton, NJ: Educational Testing Service.

Allen, N. L., Carlson, J. E., & Donoghue, J. R. (2001). Overview of part II: the analysis of 1998 NAEP data. In N. L. Allen, J. R. Donoghue, & T. L. Schoeps (Eds.), *The NAEP 1998 Technical Report* (pp. 143–160). Washington, D. C.: National Center for Education Statistics.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Beaton, A. E. (1987). *Implementing the new design: The NAEP 1983-84 technical report*. Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Brandt, S. (2006). *Exploring bundle dependencies for the embedded attitudinal items in PISA 2006*. Presented at the 13th meeting of the International Objective Measurement Workshop (IOMW), Berkeley, CA.

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments* (Vol. 1, pp. 51–70). Princeton, NJ: IEA-ETS Research Institute.

Brandt, S. (2010). Estimating tests including subtests. *Journal of Applied Measurement*, *11*, 352–367.

Brandt, S. (2012). Robustness of multidimensional analyses against local item dependence. *Psychological Test and Assessment Modeling*, *54*, 36–53.

Brandt, S., & Duckor, B. (2013). Increasing unidimensional measurement precision using a multidimensional item response model approach. *Psychological Test and Assessment Modeling*, *55*, 148-161.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Donahue, P. L., & Schoeps, T. L. (2001). Assessment frameworks and instruments for the 1998 national and state reading assessments. In N. L. Allen, J. R. Donoghue, & T. L. Schoeps (Eds.), *The NAEP 1998 Technical Report* (pp. 255–268). Washington, D. C.: National Center for Education Statistics.

Martin, M. O., & Mullis, I. V. S. (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). PISA test format assessment and the local independence assumption. *IERI Monograph series–Issues and Methodologies in Large Scale Assessments*, *4*, 131–158.

Moosbrugger, H., & Kelava, A. (2007). *Testtheorie und Fragebogenkonstruktion [Test theory and questionnaire construction]*. Heidelberg: Springer Medizin Verlag.

OECD. (2005). *PISA 2003 technical report*. Paris: OECD.

OECD. (2012a). *PISA 2012 assessment and analytical framework: mathematics, reading, science, problem solving and financial literacy*. OECD Publishing.

OECD. (2012b). *PISA 2012 technical report*. Paris: OECD.

OECD. (2014). *PISA 2012 results: what students know and can do* (Vol. 1, revised edition). Paris: Organisation for Economic Co-operation and Development.

OECD, O. for E. (2004). *The PISA 2003 assessment framework: mathematics, reading, science and problem solving knowledge and skills*. OECD Publishing.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion [Textbook test theory, test construction]*. Bern; Göttingen; Toronto; Seattle: Verlag Hans Huber.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, *6*, 181–195.

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

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*, 126–149.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER ConQuest: generalized item response modeling software*. Melbourne, Australia: Australian Council for Educational Research.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.

Yen, W. M. (1993). Scaling performance assessments - strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213.

2 Estimating Tests Including Subtests (Brandt, S. (2010). *Journal of Applied Measurement*, *11*, 352–367.)

Many of today's achievement tests, in particular large scale assessments, deal with the measurement of abilities that are assumed to be composed of further more specific abilities. In order to comply with the practical necessities of such assessments, that are to yield a unidimensional as well as multidimensional interpretation of the data, it is usually necessary to accept certain compromises. Thereby, the first and most important decision to be made is on which items to use for the study. Often this decision will be based on the results of a field trial conducted in advance of the main study. Within this decision process, furthermore, the first decision to be made is on whether the item selection is based on the items' (psychometric) properties according to the unidimensional interpretation or else on the items' properties according to the multidimensional interpretation of the data. Typically, it is decided to select the items according to a unidimensional interpretation since an accurate unidimensional analysis of the measured domain is considered more important than the analysis of the specific subdomains via the multidimensional model (cf. Martin, Gregory, & Stemler, 2000; OECD, 2002). When analyzing the data, often, a second compromise is necessary. The simultaneous unidimensional and multidimensional interpretation of the same data results in two different item parameter sets for the data. While from a psychometric point of view this is not particularly problematic, it is problematic when the results will have to be reported and communicated to a public audience. Then, different difficulties for the exact same item depending on whether it is seen, e.g., as a mathematics item or as a geometry item seem to be implausible, and therefore one might have to depend on only one of the two item parameter sets for the analyses. Typically, in accordance with the test construction, this will be the item parameter set of the unidimensional calibration (cf. OECD, 2002). Considering these typical constraints for the construction and analysis of tests that include subtests several issues for the interpretation of the data arise: (1) using the unidimensional model the items of the same subdomain show local item dependence (LID) (otherwise an interpretation of the subdomains will be superfluous); (2) items that make differences between the subdomains particularly visible will show bad unidimensional test characteristics and will therefore with a high probability be eliminated in the item selection process; (3) using the multidimensional model, the usage of the biased unidimensional item parameter estimates can lead to biased multidimensional person parameter estimates when a joint estimation for item and person parameters is used; (4) if a matrix sampling of the items with several different booklets is used so that all persons just get a subset of the items administered, the multidimensional

analysis of subdomains leads to additional differential item functioning (DIF) on the booklet level when the booklets are not balanced across the subdomains.

The following section first provides some more insight into the issues just mentioned. Thereafter, a Rasch model including subdimensions (Brandt, 2007a, 2008) is presented, which might serve as a means to avoid some of the discussed negative impacts. Finally, via an empirical example taken from the mathematics achievement test of Programme for International Student Assessment (PISA) 2003 (OECD, 2003) the model's specific characteristics are described and discussed.

2.1 Problems Considering the Construction and Calibration of Tests Including Subtests

2.1.1 Local Item Dependence

As mentioned above, a unidimensional analysis of a test including subtests means to neglect the assumed local dependencies between the items of the same subtest (or *subdimension*) and to accept the negative impact of local item dependence (LID). As already shown by many authors (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wang & Wilson, 2005a; Yen, 1993), an inappropriate assumption of local item independence results in an overestimation of test information and reliability and an underestimation of the measurement error. Furthermore, LID influences item discriminations: items with LID can show lower or higher discriminations than in the case of no LID (Wilson, 1985; Yen, 1993). Additionally, the variance of the estimated parameters decreases, as also observable in the small simulation study presented below (cf. *Figure 2.2*). Yen as well as Thissen and colleagues have examined these effects of LID by testlets, where a testlet is a subset of items in a test that refers to a common context or stimulus (Wainer & Kiely, 1987). Recently, Wang and Wilson (2005a; 2005b) have shown that it is possible to model LID due to testlets using a Rasch testlet model and, thereby, to obtain more precise and adequate estimates. Similarly, the Rasch subdimension model provides a way to obtain more appropriate and adequate measures for tests including subtests; the definition of the parameters modeling the dependencies between the items of the same subtest is different to the testlet model though. While the testlet model defines the correlation between the testlet specific factors to be zero, the subdimension model allows for a correlation of the subtest specific factors.

2.1.2 Minimization of Subdomain Differences

While the above issue concerns problems when calibrating tests including subtest, this issue relates to the construction of tests. When item characteristics are analyzed via a pilot study or field trial, the items for the main study are usually selected according to a unidimensional analysis even though the same test is supposed to measure several subdomains via a multidimensional analysis (cf. Martin et al., 2000; OECD, 2002). As a consequence, items that exhibit differences between the subdomains particularly well will with a high probability not be selected for the main study since their observed answer probabilities differ more strongly from the expected answer probabilities according to a unidimensional latent trait. Thus, the dilemma of this test construction is that the better the test developers are able to fulfill their task of developing a unidimensional measure, the less value the interpretation of the subdomains will have. The construction of meaningful subtests for the subdomains are necessary though to be able to interpret results in a more qualitative manner, e.g., via profiles for certain types of persons which is often the underlying aim for the construction of the subdomains.

2.1.3 Bias of the Multidimensional Person Parameter Estimates

When results are publicly reported the test analysis might be restricted to the usage of just one set of item parameters in order to avoid plausibility questions that might arise in the public when one and the same item has two different difficulties depending on whether it is, e.g., a mathematics or a geometry item. The well known PISA studies (OECD, 2002, 2005) solve this conflict by anchoring the item parameters for the multidimensional calibration on the item parameters obtained from the unidimensional calibration. First of all, this means that local dependencies between the items due to the different subdomains they refer to are neglected, and this assumption of local independence which is likely to be wrong will then, as described above, result in biased item parameter estimates; more precisely they will show a decrease in variance (cf. Wang & Wilson, 2005a; Yen, 1993). Furthermore though, the application of these biased item parameter estimates for the calibration of the multidimensional person parameter estimates might then also bias these multidimensional results. The small simulation study presented in the following gives an example for this bias.

The test considered for the simulation study is a 2-dimensional test with a correlation of 0.6 between the two dimensions, each of which consist of 25 items. The variances of the person parameter distributions of dimensions were both set to be 2.00 with a mean of 0.00. The item parameter distributions for dimension 1 and 2 are depicted in Figure 2.1. The exact

item parameters used for the generation (denoted as "true parameters") are given in Table 2.4 in the appendices. Furthermore, 1000 persons per test were assumed and 100 replications were generated and calibrated in order to estimate the parameter recovery. The generation of the person parameters as well as the calibration of the item and person parameters was conducted using the software ConQuest (Wu, Adams, & Wilson, 1998), which uses a Marginal Maximum Likelihood (MML) estimation method.

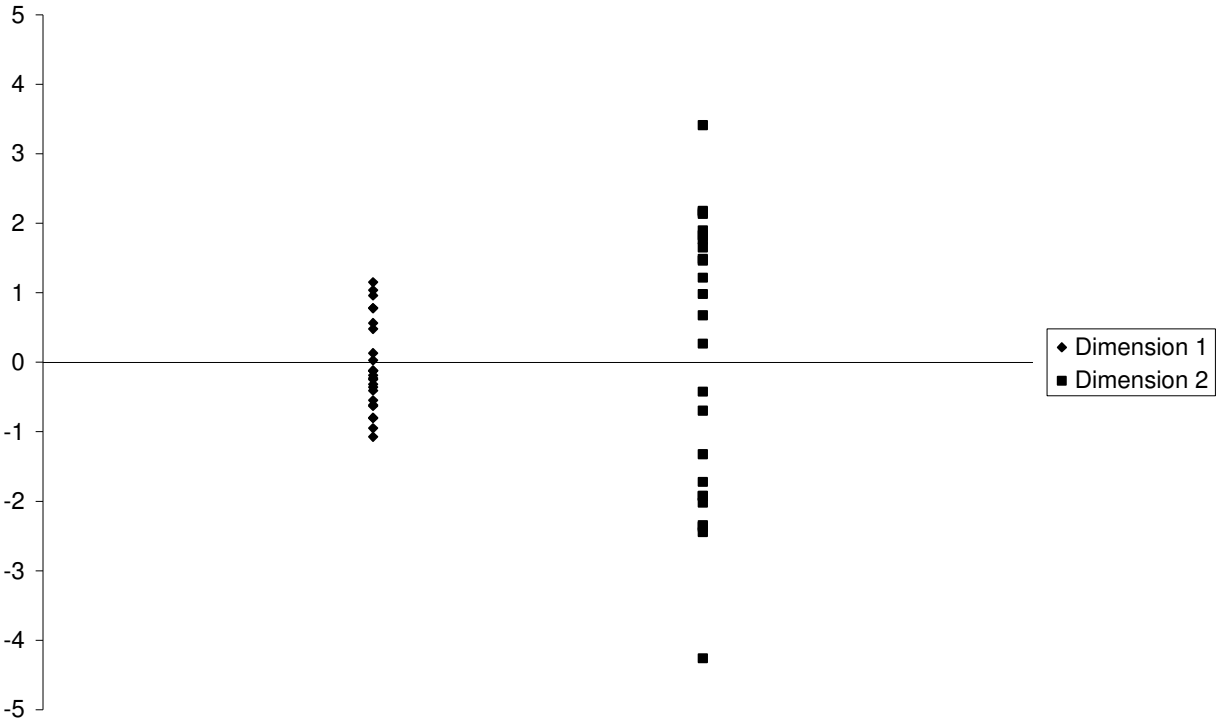

Figure 2.1. Item Parameter Distribution for Dimension 1 and 2 of the Simulated Test

Figure 2.2 depicts the item parameter recovery for a unidimensional calibration of the test, which shows the expected bias of the item parameter estimates. Hotelling's T^2 test was used to test the overall null hypothesis for unbiased estimation of the item parameters; that is $E(\hat{\xi}) = \xi$. The result of the test is given in Table 2.1 and indicates a clear bias for the estimation. The complete results of the calibration and the biases for the single estimates are given in Table 2.4 in the appendices. Likewise to Figure 2.2 these results show that the calibration results in a linear transformation of the true parameters, and that the variance of the estimated parameters is decreased as compared to the generating parameters. In average, the unidimensional calibration leads to an underestimation or overestimation of the item difficulty or easiness, respectively, by about 0.12 logits, which is the average root mean square error (RMSE). Looking separately at the results for the two subtests, it gets apparent

though that the impact due to the local item dependence of the underlying two dimensions is not the same for all items. For the items of subtest 1 (item 1 to 25), the average RMSE for the item parameters is 0.086 and for subtest 2 (item 26 to 50) 0.151; that is the bias is almost twice as strong for the items of subtest 2.

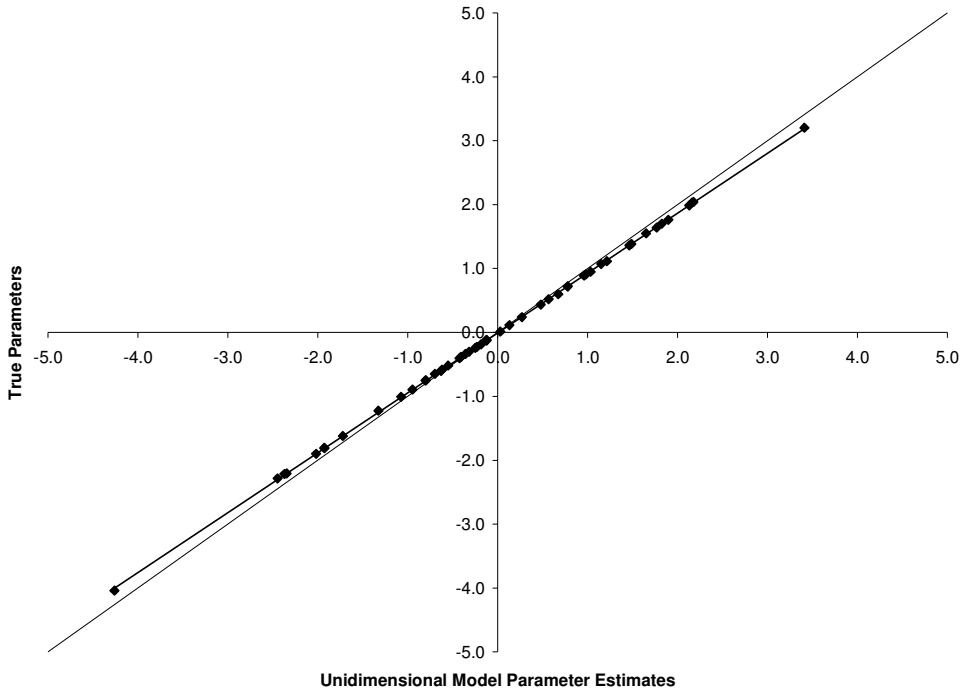

Figure 2.2. Relationship between the True Parameters and the Estimated Parameters Using the Unidimensional Model

Table 2.1

Hotelling's T^2 Tests for the Overall Null Hypothesis of Unbiased Estimation

	F	df_1	df_2	*p*
Unidimensional Calibration	79.30	49	51	0.00

The recovery of the person parameter estimates is given in Table 2.2. The two columns in the center show the recovery of the calibrations with item parameters freely estimated from a two dimensional model ("non-anchored") while the two columns on the right show the recovery for the calibrations with anchored unidimensional item parameters. The results show that ConQuest is very successful in recovering the person parameter estimates when the calibrations are not anchored. The anchored calibrations, though, show a clear bias. While the estimated correlations don't seem to be biased and dimension 1 shows just a relatively small

bias, for dimension 2 the estimated parameter is about 10% smaller than the actual generating parameter. As Figure 2.1 shows, the variance of the item parameter distribution of dimension 2 is larger than that of dimension 1; therefore, the bias for the item parameter estimates of dimension 2 is larger than for those of dimension 1. As a consequence, the anchoring of the item parameter estimates seems to have a smaller effect on dimension 1 than on dimension 2 since the anchored item parameters for this dimension are closer to their generating values.

Table 2.2

Parameter Recovery for the Person Parameter Estimates of the Non-Anchored and Anchored 2-Dimensional Calibration

		Non-Anchored		Anchored	
Parameter	True P.	Est. P.	Bias	Est. P.	Bias
σ^2_1	2.00	2.02	0.016	1.97	-0.030
σ^2_2	2.00	1.99	-0.012	1.80	-0.199
r_{12}	0.60	0.60	-0.003	0.59	-0.005

Note. True P. = true parameter; Est. P. = estimated parameter.

2.1.4 Booklet DIF

The last problem to be considered is one that typically arises in large scale assessments where several different booklets are used for the administration of all test items. Typically, booklet DIF is considered as differential item functioning due to positional effects. That is, because one and the same item occurs in different positions in the different booklets, the difficulty estimates of the items vary due to their position in the booklet; e.g., an item might be easier at the beginning of a booklet as compared to at the end of a booklet. Concerning tests including subtests, a different form of booklet DIF can occur when the booklets are not balanced according to the tested subdomains, or subtests. Assuming a test with two subtests and two persons with the same overall ability but one with a strength in subtest 1 (and a weakness in subtest 2) and the other one with a strength in subtest 2 (and a weakness in subtest 1), these persons will obviously obtain different overall (i.e. unidimensional) ability estimates if the booklet they got administered contains, e.g., a majority of items from subtest 1. To avoid this form of booklet DIF, it is therefore important to either balance the items in

the booklets according to the different subdomains or to explicitly consider differences due to the different subdomains when the unidimensional ability parameter is estimated and thereby to yield an equal weighting of the subdomains post hoc. Such an equal weighting or balancing of the subdomains within the booklets follows Humphreys' (1962; 1970; 1981; 1986) recommendation to control DIF by balancing across items. He has long argued that it is both inadvisable and difficult to try to construct a test of strictly unidimensional items. Therefore, he recommends for certain types of DIF the balancing of items because eventually multidimensionality is what causes DIF; in this particular example, multidimensionality due to the different subdomains. He is supported in his opinion by Wainer and colleagues (1991), who at the same time address the difficulty of this task.

One reason for the difficulties of the equal weighting in practice can be that even though all items were trialed, after the administration of the test it might still become advisable to exclude certain items from the final calibration. Then, even if the items had been balanced when administered, they may not be balanced anymore when calibrated. Other reasons are more subtle, looking at tests administered via multiple booklets according to a matrix sampling of the items, typically used in large-scale assessments, the difficulty in balancing the items is due to practical matters. In these cases, the test designs are usually already constrained by several other factors: (a) the design has to control for effects due to different positions of the items in the booklets; (b) the design has to yield a sufficient linkage between the items measuring different domains; (c) items are not administered as single items but often as part of a testlet (or item bundle), which is a set of items that refers to the same stimulus. An additional constraint to balance the items of different subtests in order to control for DIF due to the subtests will therefore often not be feasible, and the only way to yield more appropriate estimates will then be to adjust the unidimensional estimates according to the subdomains when they are calibrated.

2.2 The Rasch Subdimension Model

The above discussed problems of *Local Item Dependence*, *Bias of the Multidimensional Person Parameter Estimates*, and *Booklet DIF* all have in common that these problems increase the more the included subtests differ. Therefore, a multidimensional test construction according to the subdomains does not seem to be feasible without risking the reliability of the (important) unidimensional construct of the overarching domain. The toll that has to be paid though is the *Minimization of the Subdomain Differences*. The Rasch subdimension model tries to provide a means to avoid this dilemma by giving the opportunity to develop tests that

are truly multidimensional according to the defined subdomains and at the same time do not spoil a reliable overarching unidimensional measure.

In order to achieve this, the subdimension model (Brandt, 2007a, 2008) extends the standard Rasch model (Rasch, 1980) by an additional set of parameters for subdimensions and is based on the assumption that each person has a basic ability in the measured dimension (in the subdimension model denoted as *main* dimension), and strengths and weaknesses in (to be defined ex ante) subdimensions that measure specific abilities within the measured main dimension. Thereby, the model is able to yield person parameters that account for existing LID between the items of the same subdimension.

Assuming a single measured main dimension (e.g., mathematics) composed of a number of defined subdimensions (e.g. different defined subdomains of mathematics) and assuming each person's ability in a subdimension can be characterized by a strength or weakness relative to his/her ability in the measured main dimension, three different sorts of parameters have to be considered when modeling the answers based on a Rasch model approach. The first two sorts of parameters are, analogous to the Rasch model, the item parameters σ_i (with i=1,...,I and I the total number of items) that describe the item difficulties (assuming that the items are dichotomous) and the person parameters θ_v (with v=1,...,V and V the total number of persons) that describe the persons' abilities on the measured main dimension. In addition to these parameters, the parameters γ_{vd} that describe the persons' strengths (or weaknesses) in the measured subdimensions are necessary. A person's actual ability parameter to solve an item of subdimension d (with d=1,...,D and D the total number of subdimensions) is then defined by

$$\theta_{vd} = \theta_v + \gamma_{vd} \,. \tag{2.1}$$

In the following, the parameter θ_{vd} will be denoted as *absolute* ability parameter for the subdimension, while γ_{vd} will be denoted as *relative* ability parameter for the subdimension. Using the definition of Equation 2.1, the probability p_{vi1} for a correct answer from person v to a dichotomous item i in a Rasch subdimension model can then be formulated as

$$p_{vi1} = \frac{\exp\left(\theta_v + \gamma_{vd(i)} - \sigma_i\right)}{1 + \exp\left(\theta_v + \gamma_{vd(i)} - \sigma_i\right)} \tag{2.2}$$

where the mapping $d(i)$ is defined by

$$d(i):\{1,...,I\}\rightarrow\{0,...,D\}\text{, so that }\begin{cases}\gamma_{vd(i)}=\gamma_{vd}\text{ : for item } i \text{ measuring subdimension } d\\ \gamma_{vd(i)}=\gamma_{v0}\text{ : for item } i \text{ not measuring subdimension } d\end{cases}, \tag{2.3}$$

and $\gamma_{v0}=0$. To ensure the identification of the model and that the parameters have the desired properties, the following restrictions are applied:

Restriction 1: $$\sum_{d=1}^{D}\gamma_{vd}=0 \text{ for all } v=1,...,V \tag{2.4}$$

Restriction 2: $$\text{cov}(\theta_v,\gamma_{vd})=0 \text{ for all } d=1,...,D \tag{2.5}$$

Restriction 3: $$\sum_{v=1}^{V}\theta_v=0 \quad \text{and} \tag{2.6}$$

As can be easily seen, Restriction 1 is equivalent to $\sum_{d=1}^{D}\theta_{vd}\Big/D=\theta_v$, that is, Restriction 1 ensures that θ_v is the average of the persons' absolute abilities in the subdimensions (θ_{vd}). This is the essential restriction in order for the model to be correctly identified. Restriction 2, on the other hand, is not necessary in this respect, but this restriction specifies the composition of the estimate for the main dimension; the subdimension model, here, defines the subdimensions to be equally weighted for the composition of the main dimension. All subtest specific factors are defined to have the same covariance with the main factor (namely, a covariance of zero), that is, even if, e.g., one subtest consists of just 2 items and another one of 10, a specific strength in the latter subtest will not result in a higher ability estimate for the main factor as compared to a person with a corresponding strength in the first (small) subtest – even though the first person will with a high probability give more correct answers in total. Finally, Restriction 3 is one of the common restrictions for Rasch models to ensure the identification of the model, in this case, by constraining the person parameters of the main dimension to have a mean of zero. Instead of this restriction, one may also use other variants

like constraining the item parameters to have a mean difficulty of zero or anchoring one or more of the item parameters.

An important characteristic of the subdimension model in order to estimate the model's capability of modeling the local dependencies due to the subtests is that an equivalence of the subtests' variances is a sufficient condition for the model to yield estimates equivalent to those of the multidimensional model (Brandt, 2007c). For practical concerns it will therefore be an important question to know to what extent the subdimension model is capable to deliver results similar to those of the multidimensional model even if the subtests' variances differ, or else, it is possible to construct subtests that show variances of similar size. For the first case, the empirical example presented in the following section will provide some indication.

The extension of other Rasch models to include a subdimension component, such as the partial credit model (Masters, 1982), the rating scale model (Andrich, 1978), and the linear logistic test model (Fischer, 1973) is possible, as well as the extension of the standard multidimensional model by one or more subdimension components (Brandt, 2007a, 2008). Furthermore, it has been shown that the model is a special case of the multidimensional random coefficients multinomial logit model (MRCMLM) (Adams, Wilson, & Wang, 1997). Therefore, the subdimension model can be calibrated using the software ConQuest (Wu et al., 1998).

The following section now demonstrates such a calibration of the subdimension model, and via the given empirical example the potential benefits of the model will be discussed.

2.3 An Empirical Example

The empirical example is based on data taken from the mathematics achievement test of the PISA 2003 study (OECD, 2003, 2005). The test development for the study was conducted via several international item development centers that were coordinated by an international project management center. Test developers prepiloted their items and submitted them to the international project center for application in the field trial. Based on the results of the field trial it was decided which items were administered within the main study. Different selection criteria for the four major domains of the achievement test in PISA 2003, mathematics (which was the main focus), reading, science, and problem solving, were utilized. The main objective for the item selection was to get valid and reliable unidimensional measures for each of the mentioned subjects. Considering validity the set of items was selected so that the items of, e.g., the mathematics test were balanced across the four, defined ex ante,

subdomains within the domain, which were defined to be *Space and Shape*, *Change and Relationships*, *Uncertainty*, and *Quantity*. Considering technical matters DIF analyses as well as item fit measures were used. These psychometric criteria refer to a unidimensional analysis of each of the main subjects. Therefore, looking at the mathematics test, the test is from a qualitative point of view constructed to be multidimensional (with each of the four subdomains loading on its own dimension), but from a quantitative, measurement point of view it is constructed to be unidimensional. Hence, the test displays exactly the dilemma discussed earlier. The following analysis will show to what extent the application of the subdimension model might still (though the test was psychometrically constructed to be unidimensional) help to provide a more appropriate unidimensional measure by modeling the four defined subdomains of mathematics as subdimensions.

Table 2.3

Results of the Re-Analysis of the German PISA 2003 Mathematics Test

	4-Dimensional		Unidimensional		Subdimensional	
Parameter	Estimate	Reliability	Estimate	Reliability	Estimate	Reliability
σ^2_M			1.79	0.863	1.90	0.865
σ^2_1 / σ^2_{S1}	2.27	0.790			0.21	0.199
σ^2_2 / σ^2_{S2}	2.34	0.816			0.11	0.145
σ^2_3 / σ^2_{S3}	1.69	0.798			0.11	0.141
σ^2_4 / σ^2_{S4}*	1.90	0.802			0.13	
μ_M			-0.01		-0.03	
μ_1 / μ_{S1}	-0.25				-0.27	
μ_2 / μ_{S2}	-0.08				0.04	
μ_3 / μ_{S3}	-0.41				-0.56	
μ_4 / μ_{S4}*	0.62				0.78	
r_{12} / r_{S12}	0.89				-0.48	
r_{13} / r_{S13}	0.91				-0.44	
r_{14} / r_{S14}*	0.89				-0.40	
r_{23} / r_{S23}	0.94				-0.05	
r_{24} / r_{S24}*	0.94				-0.23	
r_{34} / r_{S34}*	0.93				-0.34	
Estimated Parameters	103		94		103	
-2 Log Likelihood	128586.5		128892.9		128660.0	

* Calculated via plausible values.

2.3.1 Data and Analysis

The data for the German subsample of PISA 2003, in total 4660 students, was re-analyzed for the mathematics achievement[1]. The mathematics test consists of 84 items, 20 items from the subdomain *Space and Shape (1)*, 22 items from *Change and Relationships (2)*, 20 from *Uncertainty (3)*, and 22 from *Quantity (4)*. Two out of the 84 items are partial credit items with three score categories; one item has four score categories. The data was analyzed via the partial credit model, and the respective multidimensional and subdimensional extensions.

2.3.2 Results

The results of the estimated variances, their reliabilities, the correlations, and the -2 log likelihoods for the three different models are summarized in Table 2.3. The results of the item parameter estimations are depicted in Figure 2.3 and Figure 2.4. As the true parameters are unknown for the empirical data, here, the results from the estimation of the multidimensional model serve as the best proxy and the figures show to what extent the results from the calibrations of the unidimensional and the subdimensional model show bias due to the subdomains. The exact results for the item parameter estimates are given in Table 2.5 in the appendices.

2.3.3 Discussion

Looking at the results of the multidimensional model given in Table 2.3, it gets apparent that the estimated variances for the four subtests differ from 1.69 (subtest 3) to 2.34 (subtest 2), that is, the largest estimated subtest variance is about 38.5% larger than the smallest. As mentioned above, in order for the subdimension model to model the existing LID due to the subdimensions to their full extent, that is, to yield results comparable to the multidimensional model, the chances are best when the subtests have approximately equal variances. Comparing the -2 log Likelihoods of the three calibrated models, the results show that even under these rather unfavorable conditions the subdimension model is able to close the gap between the unidimensional and the multidimensional model by over 75%; while the difference between the unidimensional and multidimensional model is 306.4 the difference between the subdimensional and the multidimensional model is only 73.5.

[1] The data is free to the public and downloadable via http://pisaweb.acer.edu.au/oecd_2003/oecd_pisa_data.html

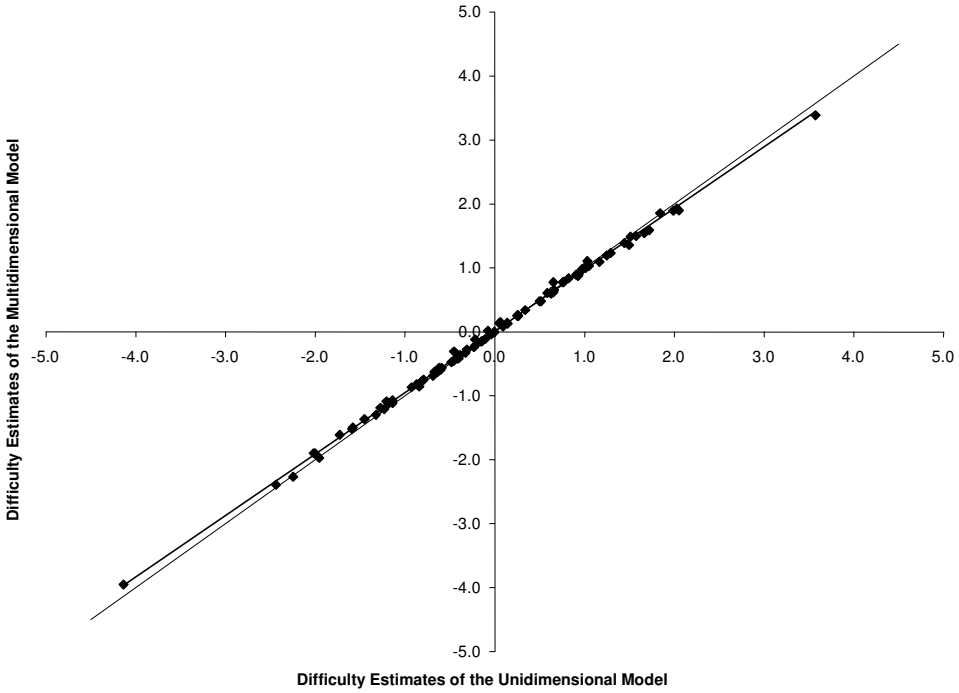

Figure 2.3. Relationship between the Difficulty Estimates Obtained Using the Multidimensional Model and the Unidimensional Model[1]

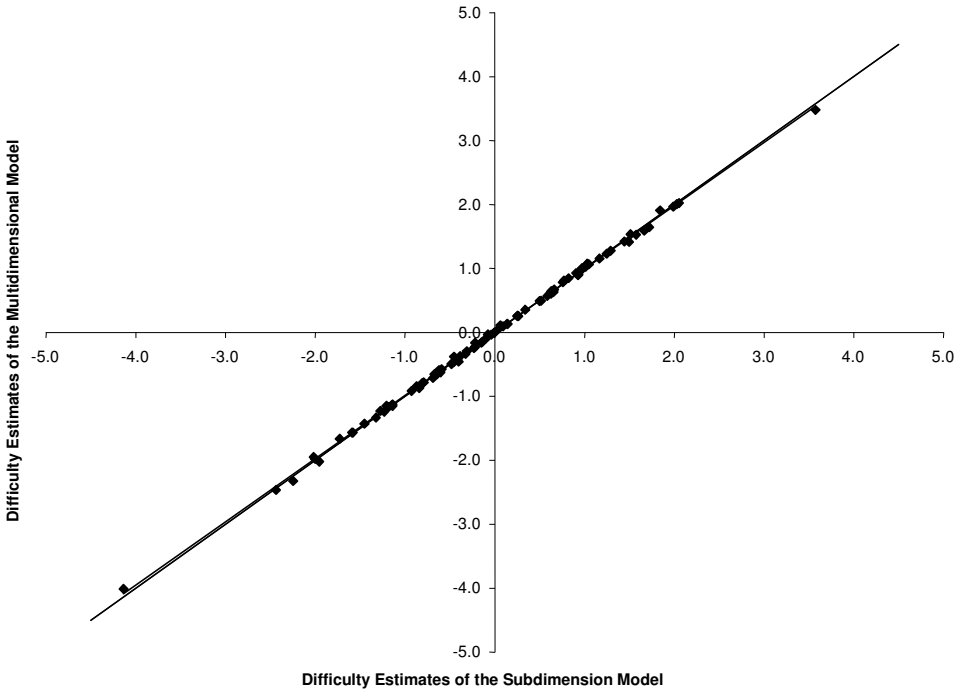

Figure 2.4. Relationship between the Difficulty Estimates Obtained Using the Multidimensional Model and the Subdimension Model[1]

[1] For a more lucid illustration the item parameter estimates of the unidimensional calibration were adjusted to have a mean difficulty of zero for each subtest.

Comparing the estimated variances for the unidimensional and the subdimensional model, one finds that the variance is underestimated when using the unidimensional model. Using the subdimension model the estimated variance increases from 1.79 to 1.90 which is equivalent to an increase by about 6%. This increase is in accordance with existent findings by other authors (e.g., Sireci et al., 1991; Wang & Wilson, 2005b; Yen, 1993), who reported a decrease of the actual variance when the test includes LID. Corresponding to this, a characteristic of tests including LID is the overestimation of the test precision, that is, the reliability is overestimated. Looking at the reliabilities given in Table 2.3 for the two main dimensions, one finds that the reliability of the main dimension estimated via the subdimension model is not lower, as usually observed by other authors, but that these two are practically equivalent; indicating that the gain in variance of about 6% in the subdimension model fully compensates for the overestimation of the reliability in the unidimensional model.

A further indication of to what extent the subdimension model is able to model the subtest LID is given via the comparison of the item parameter estimates of the three models. Assuming the item difficulty estimates of the 4-dimensional model as true parameters, Figure 2.3 provides a similar picture as the results in the simulation study (cf. Figure 2.2 above), depicting a small but clear bias. The estimates of the subdimension model depicted in Figure 2.4, on the other hand, show almost no bias.

Besides the psychometrical benefits of using a subdimension model instead of a unidimensional model, the subdimension model's estimates for the subdimensions might also provide some benefit by allowing for some additional insight into the data. Their interpretation though has to be very careful and is of rather complicated nature: at first, they are no absolute ability estimates but calculated relative to the overall ability, so that their interpretation is not intuitive; and second, for the analysis and reporting of the final results of the subtests, the multidimensional model will (usually) yield more reliable ability estimates. Therefore, the results of the subdimension specific estimates will not be interesting for reporting subtest results; nonetheless, they might be interesting from a researcher's point of view. Because the subdimension specific parameters do not include the overarching and usually dominating main ability, they can help to see differences (or similarities) between subtests more easily. In the above example, e.g., the correlation estimates of the multidimensional model are dominated by the large proportion of common variance and only differ by at most 0.05, from 0.89 to 0.94; the estimated correlations for the subdimension

model though differ by up to 0.34, from -0.13 to -0.47 (cf. Table 2.3). The estimated reliabilities for the subtest specific parameters might serve test developers and analysts as an indication on the test's reliability on distinguishing the defined subdomains. Especially if tests are used to analyze student profiles that are constructed via subtests, these reliabilities provide a measure on how reliable the differentiations of students by these profiles will be and might help to develop tests that provide particularly reliable measures in these terms. For the given empirical example, the analysis with subtest specific reliabilities from 0.139 to 0.211 indicates that interpretations of the differences between the subtests will have to be very cautious. These low reliabilities, however, are not surprising. Due to the unidimensional test construction, the items for the main study were selected so that they match the measurement of main dimension as well as possible, and their measurement characteristics towards the subdomains were not considered. Hence, just very small parts of the variances are attributable to the items' specific subdomains, and the reliabilities are correspondingly low.

2.4 Conclusion

The results of the presented empirical study show that even for tests that are psychometrically constructed to be unidimensional, the inclusion of the defined subdomains via the subdimension model yields more appropriate estimates than the application of the unidimensional model. In doing so the subdimension model corrects for the overestimation of the test reliability in the unidimensional case, which is particularly important to reduce the risk of (mis-)interpreting differences as significant due to overestimation of test precision. Furthermore, the bias of the item difficulty estimates due to the subdomain LID is largely reduced. This characteristic is particularly helpful for studies, like the PISA study, in which the same set of item parameters will have to be used to calibrate the multidimensional model for the subdomain ability estimates. Using the item parameter estimates of the subdimension model then avoids the possible bias on the multidimensional ability estimates induced by the biased unidimensional item parameter estimates.

Considering possible booklet DIF attributable to subtest differences, the subdimension model yields by definition (cf. Restriction 2 of the model above) that for the calibration of the main dimension abilities all four subtests are weighted equally. This equal weighting of the subtests is in accordance with Humphreys' recommendation to control DIF by balancing across items as described above, so that the subdimension model provides via its definition a way of avoiding booklet DIF that arises due to subtest differences. In order to show this

particular characteristic of the model more clearly and to investigate it in more detail, a further simulation study still needs to be undertaken though.

The largest benefit in the application of the subdimension model, however, is to provide a means of avoiding the dilemma that typically leads to the minimization of potential subdomain differences due to a unidimensional test construction. By applying the psychometric selection criteria for each single subtest and developing them as independent unidimensional measures, each of the subtests gets a true chance of displaying its unique characteristics, so that the chance of a more reliable distinction between the subtest specific abilities and more meaningful profiles for different types of persons increases as well. In contrast to the application of the unidimensional model to this type of data which would result in a strong increase of the problems discussed in the beginning, the application of the subdimension model then still provides a way to yield appropriate measures.

The presented application of the subdimension model depicts some important characteristics of the subdimension model. Other applications of the model to further investigate its characteristics will still have to follow up. The application of the model for vertical scaling, that is, for subtests that are given via assessments which are administered at different points in time, is straightforward, and it will have to be investigated how the results using the subdimension model relate to other models used for vertical scaling. Beyond the application to empirical data the subdimension model can also be very useful in simulation studies by providing additional and more subtle information as recently shown by the author (Brandt, 2007b). Finally, another way of using the subdimension model could be to adjust Restriction 2 of the model so that the measured subtests are not balanced within the overall measure but some of them get (per definition) more weight – or relevance – than others, and the characteristics of these models will have to be investigated as well.

2.5 References

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

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Brandt, S. (2007a). *Applications of a Rasch model with subdimensions.* Paper presented at the 2007 Annual Conference of the American Educational Research Association (AERA), Chicago.

Brandt, S. (2007b). *Item bundles with items relating to different subtests and their influence on subtests' measurement characteristics.* Paper presented at the 2007 Annual Conference of the American Educational Research Association (AERA), Chicago.

Brandt, S. (2007c). Modeling tests with subtests. *Under Review*.

Brandt, S. (2008). *Modeling tests with subtests* (Paper submitted for publication).

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist, 17*, 475-483.

Humphreys, L. G. (1970). A skeptical look at the factor pure test. In C. E. Lunneborg (Ed.), *Current Problems and techniques in multivariate psychology: Proceedings of a conference honoring Professor Paul Horst* (pp. 22-32). Seattle: University of Washington.

Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das & N. O'Connor (Eds.), *Intelligence and Learning* (pp. 87-102). New York: Plenum.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology, 71*(2), 327-333.

Martin, M. O., Gregory, K. D., & Stemler, S. E. (Eds.). (2000). *TIMMS 1999 technical report*. Chestnut Hill, MA: Boston College.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

OECD. (2002). *PISA 2000 Technical Report*. Paris: author.

OECD. (2003). *The PISA 2003 assessment framework - mathematics, reading, science and problem solving knowledge and skills*. Paris: author.

OECD. (2005). *PISA 2003 technical report*. Paris: author.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237-247.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*(3), 247-260.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-202.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28*(3), 197-219.

Wang, W.-C., & Wilson, M. (2005a). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*(4), 296-318.

Wang, W.-C., & Wilson, M. (2005b). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126-149.

Wilson, M. (1985). *Measuring stages of growth* (ACER Occasional Paper No. 19). Melbourne, Australia: Australian Council for Educational Research.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER ConQuest: generalized item response modeling software*. Melbourne, Australia: Australian Council for Educational Research.

Yen, W. M. (1993). Scaling Performance Assessments - Strategies for Managing Local Item Dependence. *Journal of Educational Measurement, 30*(3), 187-213.

2.6 Appendices

Table 2.4

Item Parameter Recovery for the Non-Anchored and Anchored 2-Dimensional Calibration for a Subtest Correlation of 0.6

Parameter	True P.	Non-Anchored		Anchored	
		Bias	RMSE	Bias	RMSE
σ_1	-1.07	-0.003	0.088	-0.066	0.101
σ_2	-0.80	-0.007	0.090	-0.051	0.094
σ_3	-0.80	-0.006	0.093	-0.050	0.096
σ_4	-0.63	0.004	0.089	-0.029	0.081
σ_5	-0.62	-0.003	0.093	-0.034	0.088
σ_6	-0.55	-0.004	0.092	-0.030	0.087
σ_7	-0.41	-0.009	0.094	-0.024	0.087
σ_8	-0.32	-0.007	0.080	-0.016	0.073
σ_9	-0.25	-0.005	0.090	-0.010	0.079
σ_{10}	-0.19	0.004	0.089	0.003	0.078
σ_{11}	-0.12	-0.001	0.089	0.003	0.077
σ_{12}	-0.12	-0.012	0.079	-0.007	0.067
σ_{13}	-0.12	-0.009	0.080	-0.004	0.068
σ_{14}	0.03	0.001	0.080	0.016	0.073
σ_{15}	0.13	-0.007	0.078	0.016	0.070
σ_{16}	0.48	-0.005	0.085	0.042	0.086
σ_{17}	0.57	-0.009	0.085	0.045	0.084
σ_{18}	0.78	-0.010	0.090	0.059	0.097
σ_{19}	0.78	-0.009	0.084	0.060	0.096
σ_{20}	0.96	-0.007	0.080	0.074	0.103
σ_{21}	1.03	0.001	0.085	0.086	0.113
σ_{22}	1.15	-0.015	0.105	0.079	0.119
σ_{23}	-0.95	0.000	0.098	-0.054	0.102
σ_{24}	-0.36	-0.009	0.080	-0.021	0.073
σ_{25}	-0.23	0.001	0.081	-0.003	0.072
σ_{26}	-4.26	0.008	0.200	-0.221	0.293
σ_{27}	-2.45	-0.017	0.123	-0.162	0.196
σ_{28}	-2.37	-0.014	0.116	-0.155	0.186
σ_{29}	-2.35	0.003	0.101	-0.138	0.167
σ_{30}	-2.02	-0.001	0.087	-0.123	0.145
σ_{31}	-1.93	-0.006	0.097	-0.123	0.150
σ_{32}	-1.92	0.004	0.102	-0.113	0.144
σ_{33}	-1.72	0.005	0.092	-0.100	0.129
σ_{34}	-1.33	-0.020	0.094	-0.099	0.130
σ_{35}	-0.70	-0.008	0.086	-0.048	0.091

Table 2.4 *(continued)*

Parameter	True P.	Non-Anchored		Anchored	
		Bias	RMSE	Bias	RMSE
σ_{36}	-0.42	0.004	0.088	-0.019	0.080
σ_{37}	0.27	0.005	0.089	0.029	0.083
σ_{38}	0.67	0.025	0.101	0.075	0.115
σ_{39}	0.98	0.001	0.092	0.073	0.110
σ_{40}	1.21	0.014	0.103	0.102	0.138
σ_{41}	1.46	-0.003	0.096	0.102	0.133
σ_{42}	1.49	-0.002	0.095	0.104	0.134
σ_{43}	1.65	-0.017	0.099	0.101	0.133
σ_{44}	1.77	0.002	0.103	0.126	0.157
σ_{45}	1.83	-0.007	0.086	0.122	0.143
σ_{46}	1.90	0.002	0.108	0.134	0.165
σ_{47}	2.13	-0.004	0.109	0.143	0.173
σ_{48}	2.16	-0.019	0.101	0.130	0.158
σ_{49}	2.18	-0.011	0.104	0.139	0.168
σ_{50}	3.41	-0.005	0.157	0.207	0.256

Table 2.5

Item Parameter Estimates for the Re-Analysis of the PISA 2003 Mathematics Achievement Test

Parameter	Subtest	4-Dim	1-Dim	4-Subdim
M033Q01	1	-2.00	-1.69	-2.05
M034Q01T	1	0.26	0.47	0.19
M144Q01T	1	-0.93	-0.65	-0.98
M144Q02T	1	1.25	1.41	1.16
M144Q03	1	-1.14	-0.85	-1.19
M144Q04T	1	0.50	0.70	0.43
M145Q01T	1	-1.58	-1.28	-1.63
M266Q01T	1	1.44	1.60	1.36
M273Q01T	1	-0.79	-0.53	-0.85
M305Q01	1	-1.15	-0.88	-1.20
M406Q01	1	1.29	1.45	1.21
M406Q02	1	1.99	2.11	1.90
M406Q03	1	2.03	2.15	1.94
M447Q01	1	-1.45	-1.15	-1.49
M462Q01T	1	2.05	2.11	1.96
M464Q01T	1	1.17	1.31	1.09
M547Q01T	1	-1.59	-1.31	-1.63
M555Q02T	1	-1.19	-0.91	-1.23
M598Q01	1	-0.81	-0.55	-0.86
M833Q01T	1	0.65	0.82	0.58
M124Q01	2	0.65	0.66	0.71
M124Q03T	2	1.50	1.41	1.50
M150Q01	2	-1.27	-1.14	-1.15
M150Q02T	2	-1.21	-1.04	-1.06
M150Q03T	2	0.14	0.19	0.22
M155Q01	2	-0.87	-0.77	-0.76
M155Q02T	2	-0.62	-0.52	-0.51
M155Q03T	2	1.72	1.64	1.73
M155Q04T	2	-0.39	-0.32	-0.29
M192Q01T	2	0.52	0.53	0.58
M302Q01T	2	-4.14	-3.91	-3.93
M302Q02	2	-2.02	-1.85	-1.87
M302Q03	2	0.93	0.92	0.98
M402Q01	2	-0.67	-0.58	-0.57
M402Q02	2	0.93	0.94	0.99
M446Q01	2	-0.59	-0.52	-0.49
M446Q02	2	3.57	3.43	3.56
M571Q01	2	-0.31	-0.23	-0.21
M704Q01T	2	-1.73	-1.57	-1.58
M704Q02T	2	1.57	1.54	1.61
M810Q03T	2	1.66	1.59	1.67
M828Q01	2	0.63	0.64	0.69
M179Q01T	3	0.77	1.19	0.65
M408Q01T	3	0.25	0.67	0.10
M411Q02	3	-0.01	0.40	-0.17
M420Q01T	3	-0.23	0.17	-0.41

Table 2.5 *(continued)*

Parameter	Subtest	4-Dim	1-Dim	4-Subdim
M421Q01	3	-0.84	-0.45	-1.03
M421Q02T	3	1.84	2.26	1.75
M421Q03	3	0.34	0.75	0.19
M423Q01	3	-1.96	-1.56	-2.18
M438Q01	3	-2.25	-1.86	-2.48
M438Q02	3	-0.04	0.37	-0.20
M467Q01	3	-0.33	0.08	-0.50
M468Q01T	3	-0.14	0.27	-0.31
M505Q01	3	-0.42	-0.01	-0.59
M509Q01	3	0.26	0.67	0.10
M513Q01	3	0.97	1.39	0.84
M564Q02	3	-0.15	0.26	-0.32
M702Q01	3	0.82	1.25	0.69
M710Q01	3	0.76	1.19	0.62
M803Q01T	3	1.03	1.45	0.90
M828Q02	3	-0.69	-0.28	-0.88
M411Q01	4	0.63	0.00	0.77
M413Q01	4	-1.32	-1.92	-1.21
M413Q02	4	-1.23	-1.83	-1.12
M413Q03T	4	0.91	0.27	1.05
M442Q02	4	0.93	0.30	1.07
M474Q01	4	-1.14	-1.73	-1.03
M484Q01T	4	-0.22	-0.83	-0.10
M496Q01T	4	0.25	-0.37	0.38
M496Q02	4	-0.48	-1.09	-0.37
M510Q01T	4	0.97	0.35	1.11
M520Q01T	4	-0.61	-1.21	-0.50
M520Q02	4	1.01	0.38	1.15
M520Q03T	4	0.66	0.04	0.79
M559Q01	4	-0.11	-0.73	0.01
M564Q01	4	0.63	0.00	0.76
M603Q01T	4	0.09	-0.54	0.21
M603Q02T	4	1.05	0.41	1.19
M800Q01	4	-2.44	-3.01	-2.34
M806Q01T	4	0.00	-0.62	0.12
M810Q01T	4	-0.46	-1.07	-0.34
M810Q02T	4	-0.64	-1.25	-0.53
M828Q03	4	1.51	0.87	1.66
M462Q01T step 1	1	0.05	0.13	0.07
M124Q03T step 1	2	-0.45	-0.31	-0.38
M124Q03T step 2	2	0.14	0.13	0.13
M150Q02T step 1	2	-0.22	-0.12	-0.16
M155Q02T step 1	2	1.03	1.11	1.07
M155Q03T step 1	2	0.06	0.15	0.11
M810Q03T step 1	2	-0.08	0.02	-0.03
M179Q01T step 1	3	-0.40	-0.42	-0.46
M520Q01T step 1	4	0.58	0.61	0.57

3 Estimation of a Rasch Model Including Subdimensions (Brandt, S. (2008). In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments* (Vol. 1, pp. 51–70). Princeton, NJ: IEA-ETS Research Institute.)

3.1 Introduction

Many of today's achievement tests, in particular those used within large-scale assessments, deal with measuring abilities that are themselves assumed to be composed of other more specific abilities. As such, a common approach taken by large-scale cross-national assessments like TIMMS and PISA when endeavoring to yield the necessary ability estimates is to analyze the same data-set once using a unidimensional model and once using a multidimensional model (cf. Martin, Mullis, & Chrostowski, 2004; OECD, 2005). This approach, however, has two major downsides. First, from a theoretical point of view, the assumption that the data fits both unidimensional and multidimensional models seems to make model-fit tests obsolete and the application of a particular model somewhat arbitrary—or simply determined by pragmatic needs. Second, and this time from a practical point of view, neglecting the assumed local dependencies among the items of the same subtest (or *subdimension*) that measure a more specific ability means accepting the negative impacts of local item dependence (LID).

As already shown by many authors (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wang & Wilson, 2005a; Yen, 1993), an inappropriate assumption of LID results in an overestimation of test information and reliability and an underestimation of the measurement error. Furthermore, because LID influences item discriminations, items showing LID also show lower discriminations than is the case with items showing no LID (Yen, 1993), and, finally, the variance of the estimated parameters decreases for items with LID.

Yen and Thissen and his colleagues have examined these effects of LID through the use of "testlets"—a subset of items in a test that have a common structural element. An example is bundles that have a common stimulus (H. Wainer & Kiely, 1987). More recently, Wang and Wilson (2005a; 2005b) showed that it is possible to model LID in relation to testlets by using a Rasch testlet model and thereby obtaining more precise and adequate estimates. The Rasch testlet, as well as the Rasch subdimension model proposed below, are special cases of the group of so-called bi-factor models. These models are characterized by the fact that each item loads on at least two dimensions, on a general factor, and on one or more group—or

method-specific—factors, such that the loading on the general factor is non-zero (Holzinger & Swineford, 1937).

In order to analyze these types of models, Gibbons and Hedeker (1992) developed a full-information item bi-factor analysis for binary item responses. The development of appropriate models and estimation procedures relative to graded response data has, however, been less successful (cf. Muraki & Carlson, 1995). The additional computational complexity associated with graded response data leads to the introduction of additional model constraints in order to estimate the model. One restriction commonly applied is that the method- or group-specific factors (in the case of the model presented in the following section, denoted as the latent traits of the *subdimensions*) are constrained so that the factors are independent from the general factor (termed the "main dimension" in the following). While this constraint is appropriate in that the specific factors measure only the residual associations of the items beyond those due to the general latent trait and although this constraint is a common feature of bi-factor models, the computational complexity seems to make a second model constraint necessary for graded response data. As a consequence, an additional assumption in regard to the Rasch testlet model, as well as in regard to the recently proposed full-information item bi-factor analysis for graded response data (Gibbons et al., 2007), is that the specific factors are also independent of one another.

The model that I propose in this paper tries to loosen this latter—rather strong—constraint through application of a different constraint but one that still allows for correlation of the specific factors. I discuss the possible consequences of these different assumptions in somewhat more detail after defining the model in the next section. I then show how to calibrate the model using the software ConQuest. This is followed by an empirical example that depicts the differences between the unidimensional, the (unrestricted) multidimensional, and the subdimension models.

3.2 A Rasch Model that Includes Subdimensions

To resolve the theoretical problem of unidimensionality versus multidimensionality and to reduce negative impacts on measurement precision due to LID, the model proposed here is a Rasch subdimension model (Brandt, 2007a, in preparation). The model extends the standard Rasch model (Rasch, 1980) by using an additional set of parameters for subdimensions, and it is based on the assumption that each person has a general ability in the measured dimension (which in the subdimension model is denoted as the *main* dimension) as well as strengths and weaknesses (to be defined *ex ante*) in the subdimensions that measure specific abilities within

the measured main dimension. This way, the model is able to yield person parameters that account for existing LID among the items of the same subdimension. The model is also a special case of the multidimensional random coefficients multinomial logit model (MRCMLM) (Adams, Wilson, & Wang, 1997), so that it can be directly estimated through use of the software ConQuest (Wu, Adams, & Wilson, 1998).

3.2.1 Model Definition

Assuming we have a single measured main dimension (e.g., mathematics) that is composed of a number of defined subdimensions (e.g., differently defined areas of mathematics) and assuming that we can characterize each person's ability in a subdimension according to a strength or weakness relative to his or her ability in the measured main dimension, we end up with three different sorts of parameters to consider when modeling the answers based on a Rasch model approach. The first two sorts of parameters are, analogous to the Rasch model, the item parameters b_i (with i=1,...,I and I the total number of items) that describe the item difficulties, and the person parameters θ_v (with v=1,...,V and V the total number of persons) that describe the persons' abilities on the measured main dimension. In addition to these parameters, we need parameters γ_{vd} that describe the persons' strengths (or weaknesses) in the measured subdimensions. The persons' actual ability parameter to solve an item from subdimension d (with d=1,...,D and D the total number of subdimensions) is thus defined by

$$\theta_{vd} = \theta_v + \gamma_{vd} . \tag{3.1}$$

While the parameter θ_v denotes the *overall* ability parameter across subdimensions, the parameter γ_{vd} denotes the *specific* ability parameter for the subdimension. If we use the definition in Equation 3.1, the probability p_{vi1} of person v giving a correct response to a dichotomous item i in a Rasch subdimension model is then given as

$$p_{vi1} = \frac{\exp\left(\theta_v + \gamma_{vd(i)} - b_i\right)}{1+\exp\left(\theta_v + \gamma_{vd(i)} - b_i\right)} \tag{3.2}$$

where $\gamma_{vd(i)} = d(i) \cdot \gamma_{vd}$ and $d(i)$ is equal to 1 when item i measures subdimension d, and 0 otherwise. To ensure that the parameters have the needed properties, further restrictions of the parameters have to be introduced (cf. Brandt, 2007a, in preparation):

Restriction 1: $$\sum_{d=1}^{D} \gamma_{vd} = 0 \text{ for all } v = 1,...,V \tag{3.3}$$

Restriction 2: $$\text{cov}(\theta_v, \gamma_{vd}) = 0 \text{ for all } d = 1,...,D \tag{3.4}$$

Restriction 3: $$\sum_{v=1}^{V} \theta_v = 0 \tag{3.5}$$

Restriction 1 is equivalent to $\sum_{d=1}^{D} \theta_{vd} \Big/ D = \theta_v$; that is, it assures that θ_v is the average of the persons' absolute abilities in the subdimensions (θ_{vd}). This restriction is essential for correctly identifying the model. Restriction 2, however, is not necessary in this respect; rather, it specifies the composition of the estimate for the main dimension. By constraining all subdimension specific factors to have the same covariance with the main dimension (namely zero); the subdimensions are defined to be equally weighted for the composition of the main dimension. This practice accords with the common assumption inherent with the bi-factor models described above. It also accords with Humphreys' (1962; 1970; 1981; 1986) recommendation to control DIF (which arises in the here considered case by the subtests measuring different specific abilities) by balancing across items. Humphreys is supported in this opinion by Wainer Sireci, and Thissen (1991), who also address the difficulty of this task.

Finally, Restriction 3 is one of the common restrictions that ensure correct identification of the model. The shown restriction in this case represents the constraint of the mean of the person parameters of the main dimension to zero. However, as an alternative to this restriction, we can constrain the item parameters to have a mean of zero, or anchor one or more of the item parameters.

By using Equation 3.2, we can also formulate the log-odds form of the subdimension model. This results in

$$\log(p_{vi1} \big/ p_{vi0}) = \theta_v + \gamma_{vd(i)} - b_i, \tag{3.6}$$

where p_{vi0} denotes the probability of person *v* giving an incorrect answer to item *i*, and requires application of Restrictions 1 to 3, described above. Furthermore, the equations stated above for dichotomous items can be extended to

$$\log(p_{vij} / p_{vi(j-1)}) = \theta_v + \gamma_{vd(i)} - b_{ij}, \tag{3.7}$$

for polytomous items, where p_{vij} and $p_{vi(j-1)}$ are the probabilities of scoring *j* and *j*-1 (where $j = 1,...,K_i - 1$ and K_i is the number of categories for item *i*) to item *i* for person *v*, respectively, and b_{ij} is the *j*th step difficulty of item *i*. By introducing a parameter b_i, called overall item difficulty, and a parameter τ_{ij}, called *j*th threshold of item *i*, where

$$b\sigma_{ij} = b_i + (b_{ij} - b_i) = b_i + \tau_{ij}, \tag{3.8}$$

we can express Equation 3.7 as

$$\log(p_{vij} / p_{vi(j-1)}) = \theta_v + \gamma_{vd(i)} - (b_i + \tau_{ij}), \tag{3.9}$$

which reduces to the partial credit model (Masters, 1982) when $\gamma_{vd(i)} = 0$. Extending other Rasch models to include a subdimension component, such as the rating scale model (Andrich, 1978) or the linear logistic test model (Fischer, 1973), is straightforward.

By defining weights

$$q_{id} = \frac{u_{id}}{\sum_{d^*=1}^{D} u_{id^*}}, \tag{3.10}$$

where u_{id} is an indicator variable that is 1 if item *i* is within dimension *d* and is zero otherwise, and by inserting Equation 3.1, we can further express Equation 3.2 as

$$p_{vi1} = \frac{\exp\left(\left(\sum_{d=1}^{D} q_{id}\theta_{vd}\right) - b_i\right)}{1 + \exp\left(\left(\sum_{d=1}^{D} q_{id}\theta_{vd}\right) - b_i\right)} \tag{3.11}$$

thereby matching the multidimensional Rasch model (Carstensen, 2000; Rost, 1996). We can, in fact, see the subdimension model as a re-parameterized multidimensional model, somewhat similar to Masters' partial credit model, which re-parameterizes the Rasch model for polytomous items. Note, however, that in the case of the subdimension model, it is not the item parameters but the person parameters that are re-parameterized.

3.2.2 Discussion of the Model

To provide more insight into the subdimension model, I now discuss Restriction 1 and Restriction 2 of the model in more detail.

As I mentioned above, the subdimension model allows for correlations between specific abilities, in contrast to (for example) the Rasch testlet model. Rather, the model incorporates a restriction on the sum of the estimates for the specific abilities (Restriction 1)—a characteristic that can constrain the size of the measured variances for the subdimensions, particularly if the differences in the measured variances are very large. For tests with subdimensions of equal variance, however, it has been shown that the subdimension model provides results equivalent to those of the unrestricted multidimensional model (Brandt, 2007a, in preparation), so allowing the subdimension model to be derived through variable transformation.

This attribute is particularly noteworthy in relation to large-scale assessments such as PISA and TIMSS that utilize detailed background information on the students to impute values for the proficiency variable even though a large portion of item responses are missing due to the matrix-sampling of items administered to each student. Within the calibration process of the person parameter estimates, analysts can use this background information as regression parameters on the estimated latent traits, a process that leaves the residual variances of the latent traits reflecting only those parts of the variances that are not attributable to the regression parameters. This situation, in turn, leads to a decrease in the size of the residual (conditional) variance[1] for the latent traits. It also typically results in variances

[1] The actual variances for the latent trait, including the explained variances due to the regression parameters, are calculated post hoc.

that are closer to one another in size. Use of the subdimension model can therefore be particularly beneficial in these cases.

An important difference between the subdimension model and other bi-factor models such as the Rasch testlet model is evident in the assumptions each holds about the covariances between specific abilities. To make the resulting differences more obvious, let us consider an example of a science assessment consisting of four testlets, each with five items that refer to a common stimulus, and let us additionally assume that the single measurement of each testlet results in the same variance for the distribution of the measured latent trait. (In other words, the subdimension model will yield results equivalent to the unrestricted multidimensional model.) Let us further assume that the used stimuli relate to the following different application areas of science— agriculture, medicine, electronics, and environmental pollution. And then let us suppose, for the purposes of this test, that application of the Rasch testlet model leads to variances of var_{T1} to var_{T4} for the testlet specific dimension and that application of the subdimension model leads to variances of var_{S1} to var_{S4}.

While, for a given test, we may not find it easy to recognize how these two differently measured variances for the same testlet differ, the difference becomes transparent when we assume that a fifth testlet has been added to the test and that this testlet has a higher correlation with one of the already existing testlets than with the remaining three, perhaps because the used stimulus relates to the same application area of (say) agriculture as Testlet 1 does. As a consequence, that part of the testlet specific variance var_{T1} attributable to the application area used is equivalent to that of the new Testlet 5. However, because the Rasch testlet model assumes these variances are independent, applying the Rasch testlet model to the test that has all five testlets will not model the specific variance attributable to the application area agriculture because of the need to comply with the independence assumption concerning the testlet-specific effects. In short, the model will not account for LID because of the application area in question. Hence, in the Rasch testlet model, var_{T1} will be smaller in the test with all five testlets than in the test with just four testlets. An analysis of item-bundle effects for the mathematics achievement test of PISA 2003 showed that the size of the testlet-specific variances for the item bundles included in all tests differed to a considerable extent according to whether certain item bundles were included or excluded from the analysis (Brandt, 2006).

In the subdimension model, however, the variance of var_{S1} will be the same in the test with four and five testlets (subdimensions) if the testlets yield variances of equal size. When

the testlets do not show equal variances, the model usually becomes less capable of modeling the local dependencies among the items of the same testlet.[1] The measured specific effects are comparatively stable, however, and do not depend on the content of the other subdimensions (testlets) in the test. Rather, they depend solely on the size of the variances of these subdimensions, which essentially is due to a normalization problem. Because the subdimension model does not assume the independence of the subdimension-specific factors, the model is less sparse than the testlet model.

In the Rasch testlet model, only one parameter (the variance of the testlet specific effects)[2] has to be estimated, but in the subdimension model, the covariances for all other existing subdimensions have to be estimated as well. Therefore, it is possible to calibrate the Rasch testlet model for even large numbers of testlets. The number of parameters to be estimated for the subdimension model, however, increases in the same way as occurs with those within the unrestricted multidimensional model. In fact, for both models, the same numbers of parameters always have to be estimated given that the subdimension model is essentially a variable transform of the unrestricted multidimensional model. Comparison of the unrestricted multidimensional model and the subdimension model shows that the number of estimated dimensions is equivalent in both models (see the definition of the scoring matrix above). So, if we assume that the ability estimates in both models are constrained to a mean of zero for a test with n dimensions, we will not need to estimate the parameters for n-1 covariances in the subdimension model due to Restriction 2. However, unlike the situation with the unrestricted model, we would have to estimate the n-1 additional parameters for the means of the subdimension-specific latent traits. This is because in the subdimension model only the ability estimates of the main dimension are constrained to a mean of zero. Thus, the number of parameters to be estimated is always the same for both models.

3.3 Estimation Using ConQuest

Because the MRCMLM includes the subdimension model as a special case (Brandt, 2007a, in preparation), the software ConQuest (Wu et al., 1998) can be used to estimate the model. Although the mathematical definition of the subdimension model given via the MRCMLM is provided in the proof that the subdimension model is a special case of the MRCMLM, it is

---

[1] In certain cases, the subdimension model might be able to yield results equivalent to those of the unrestricted multidimensional model when the variances of the testlets/subdimensions differ (cf. Brandt, 2007a, in preparation).

[2] Because ConQuest uses a marginal maximum likelihood approach for the parameter estimation, and assuming a standard normal distribution for the measured latent traits, only the mean and the variance of the distribution are estimated.

still necessary to fully understand the notations used for the definitions of the scoring and design matrices and, furthermore, to be able to define the resulting constraints using ConQuest syntax. Given the complexity of the MRCMLM notation, as well as the need for knowledge about ConQuest, models like those described above, and also the Rasch testlet model (Wang & Wilson, 2005b), are barely accessible to people interested only in applying adequate models to their data and who are less interested in understanding the theoretical definitions and concepts of particular models. Therefore, a main goal of this paper is to fill in this gap relative to the subdimension model and to give a detailed description for calibrating it. To do this, I begin by briefly describing (in the next paragraph) the different given ways of defining or constraining a model via ConQuest.

Basically, ConQuest offers five different ways of defining or constraining a specific model within the MRCMLM:

1. Through the definition of the design matrix, which describes the linear relationship among the items;
2. Through the definition of a scoring matrix, which assigns the items to specific ability dimensions and assigns scores to their response categories;
3. Through the anchoring of the item difficulty parameters, which can be used not only for linking to other tests but also for identification purposes;
4. Through specification of mean abilities for the population distribution[1] (within ConQuest denoted as so-called regression parameters); and
5. Through anchoring of the variance-covariance matrix.

Although the definition of the scoring matrix is embedded in ConQuest's command language, the remaining four types of specifications are done via imported text files. For standard unidimensional or multidimensional calibrations, ConQuest's command language provides the means by which the analyst can automatically generate the appropriate design matrices and anchorages. It is through the command `model` the simple Rasch model (`model item;`), the rating scale model (`model item + step;`), the partial credit model (`model item + item*step;`), and other multifaceted models can be defined, and it is through the `set constraint` command that identification of the model can be set to `items` (i.e., the mean of the item difficulties is set at zero) or `cases` (i.e., the mean of the ability distribution is set at zero).

[1] Because ConQuest uses a marginal maximum likelihood estimation method, the ability distributions for a given population are assumed to be normal.

On completion of other necessary commands concerning the data file to be calibrated and the output files that are to be generated, a ConQuest command file for the calibration of a unidimensional partial credit model looks like this:

```
datafile estimation.dat;              /* Definition of the data file with
                                      the students' answers */

format responses 1-40;                /* Columns in the data file that
                                      represent the students' answers */

codes 0,1,2;                          /* Definition of valid answer codes
                                      - all other codes will be
                                      interpreted as missing by design */

score (0,1,2) (0,1,2)!items (1-40); /* Definition of the scoring matrix
                                      (here, according to a unidimensional
                                      model)*/

model item + item*step;               /* Definition of the design matrix
                                      (here, according to the partial
                                      credit model */

set constraints = items;              /* Constraint for the identification
                                      of the model */

export designmatrix >> estimation.dsm; /* Export of the design matrix to
                                      a data file with the given name */

estimate;                             /* Start of the calibration using
                                      the standard settings /*

show >> estimation.shw;               /* Export of the calibration results
                                      to a data file with the given name
                                      */
```

This example assumes that the answer data provided by the data file has already been scored and that a single digit represents each coded answer. Thus, each column represents the students' answers to a particular item, scored with 0, 1, or 2 credits (cf. also the `code` statement above).[1] In the unidimensional case, the scoring matrix reduces to a simple vector (with 40 elements) that maps all scores of all items to the same dimension.

In regard to the definition of the design matrix via the `model` command, note that if the model is constrained to have a mean item difficulty of zero (as in the above case), the design matrix will have to be changed accordingly. Therefore, the design matrix will not be generated until the start of the calibration in order to comply with the given `set constraint` command. As for the unidimensional calibration conducted by the above command file, the `export designmatrix` command is not necessary. Nevertheless, this statement is valuable

[1] ConQuest also provides a way of scoring the data via the command language; more information about these commands can be found in the ConQuest Manual.

here because the design matrix generated for the calibration is exactly the design matrix needed in order to define the subdimension model presented below.[1] Finally, the `estimate` command starts the calibration with the standard algorithm and convergence criteria of ConQuest, and the `show` command generates a standard output for the results of the calibration, written to a text file named "estimation.shw".

Defining models like the subdimension model requires somewhat more effort since there is no ConQuest command to automatically generate and set the necessary constraints according to the model definitions. This has to be done manually instead by providing appropriate import files. The main focus of this section, therefore, is to describe the construction and definition of these import files as well as the definition of the specific scoring matrix needed for the model.

```
datafile estimation.dat;              /* See above */
format responses 1-40;                /* See above */
codes 0,1,2;                          /* See above */

score (0,1,2) (0,1,2) (0,1,2) () () !items (1-10);
score (0,1,2) (0,1,2) () (0,1,2) () !items (11-20);
score (0,1,2) (0,1,2) () () (0,1,2) !items (21-30);
score (0,1,2) (0,1,2) (0,-1,-2) (0,-1,-2) (0,-1,-2) !items (31-40);
                                      /* Definition of the scoring matrix
                                      */

model item + item*step;               /* Pseudo-definition of the design
                                      matrix */

import designmatrix << estimation.dsm;       /* Actual definition of the
                                      design matrix */

import anchor_covariance << estimation.cov;  /* Setting of the
                                      constraints for the variance-
                                      covariance matrix */

estimate !method=montecarlo,nodes=2000;      /* Start of the calibration
                                      using a Monte Carlo method with 2000
                                      nodes and standard convergence
                                      criteria */

show >> estimation.shw;               /* Export of the calibration results
                                      to a data file with the given name
                                      */
```

The first three commands of the command file correspond with the unidimensional calibration above; that is, the same data-set as the one above is calibrated. Here, it is assumed that the test includes four subtests with 10 items each, with Items 1 to 10 referring to Subtest

[1] If the design matrix has not been generated at the time the command is processed, ConQuest exports the file as soon as the design matrix is generated; that is, after the start of the calibration.

1, Items 11 to 20 to Subtest 2, Items 21 to 30 to Subtest 3, and Items 31 to 40 to Subtest 4. In order to account for the assumed local dependencies between the items of the same subtest, the subdimension model is used for estimation. As the definition of the scoring matrix above shows, the subdimension model is a multidimensional model; in the above example, it has four dimensions. The first dimension, comparable to the unidimensional case above, refers to the unidimensional latent trait that all 40 items commonly measure. The second and fourth dimensions, however, refer to the specific parameters of the subdimensions that are to be estimated. According to Restriction 1 of the definition of the subdimension model, the subdimension-specific parameters must add up to zero for each single student. In order to comply with this restriction, the parameter estimates of the fourth subdimension cannot directly be estimated but rather defined as constrained parameters. When the sum of the four subdimension-specific parameters is zero, then each person's fourth parameter (or any single other of the four) always equals the negative of the sum of the other three parameters. What this means, in essence, is that the subdimension model actually contains only *d-1*-estimated *specific* dimensions, and one final specific dimension, which is totally determined by the negative sum of the previous *d-1*. Therefore, Items 1 to 10 load (in addition to the main dimension) on dimension 2, Items 11 to 20 on Dimension 3, Items 21 to 30 on Dimension 3, and Items 31 to 40 negative on Dimensions 2 to 4.

The model statement follows the process involved in defining the scoring matrix. This statement has only a dummy function, which exists for programming reasons, given that ConQuest's `estimate` command must always be preceded by a `model` command. The design matrix generated according to this standard statement cannot be used because it is problematic in two ways. First, for items that load on more than one dimension, ConQuest adjusts the design matrix in order to keep the items' difficulty estimates in proportion to the size of the ability estimates. In the case of the subdimension model, this step simply results in estimates that are exactly half the size of the unidimensional estimates, thereby making comparisons just that little bit more difficult. Secondly, ConQuest does not adjust the design matrix according to the necessary constraint of the item parameters needed for the subdimension model. The needed constraint is correctly defined, though, in the design matrix generated for the corresponding unidimensional calibration. Furthermore, by using this design matrix, the software renders the parameter estimates of the calibration for the subdimension model comparable to those of the unidimensional model, and it does this without any further linear transformation. Therefore, the easiest and (probably) least error-

provoking way to obtain the correct design matrix is to generate it with the corresponding unidimensional model, as shown in the example above.

Once the correct design matrix is imported, all that remains is correctly anchoring the variance–covariance matrix according to Restriction 2 of the model. This step requires creation and importation of an appropriate text file. For the above example, the import file "estimation.cov" has the following format:

```
1    2    0.0000
1    3    0.0000
1    4    0.0000
```

The first two figures in a row define which covariance is to be set. Thus, in the first row above (the covariance of Dimensions 1 and 2), the third figure sets the value for the given covariance. Here, all listed covariances are set at zero, a practice that aligns with the definition of the subdimension model that requires the covariances between the main dimension and the subdimensions to be constrained to zero.

The empirical example presented in the next section was calibrated using ConQuest, as described in this section.

3.4 An Empirical Example

The empirical example given here is based on data taken from the mathematics achievement test used for TIMSS 2003 (Mullis, Martin, Gonzales, & Chrostowski, 2004; Mullis et al., 2003). This test was developed according to two different aspects—a content domain and cognitive domains. While the latter domain consisted of *knowing facts and procedures*, *using concepts*, *solving routine problems*, and *reasoning*, the analysis presented in the following refers to the five defined content domains, which were *number*, *algebra*, *measurement*, *geometry*, and *data*. To select appropriate items for the main study, the TIMSS researchers conducted a full-scale field trial. They then used the results of this trial to determine which items would be used in the main study. During this selection process, the researchers took care not only to distribute the items across the four cognitive and five content domains according to the proportions defined in the assessment framework but also to ensure that the psychometric characteristics of the items were sufficient, particularly in relation to DIF effects and discrimination power (Martin et al., 2004).

Because the psychometric criteria chosen refer to a unidimensional analysis of the data, we can consider the test to have been constructed as multidimensional from a qualitative

point of view, via the *ex-ante* defined domains, and as unidimensional from a quantitative measurement point of view. The described test construction displays the dilemma of TIMMS and other large-scale assessments associated with lack of appropriate models (cf. the test construction for the PISA study, for example [OECD, 2005]). The resulting data-sets, therefore, are not good examples of true multidimensionality. Despite this, the assessment results are publicly reported and interpreted. With these considerations in mind, the following analysis shows the extent to which the subdimension model can still help provide more appropriate measures by modeling the five content domains defined for the mathematics test.

3.4.1 Data and Analysis

The analyzed test used data obtained from the United States sub-sample of students who participated in TIMSS 2003. This sub-sample consisted of 8,912 students in total, and the test included 194 mathematics items: 47 items for the content domain *algebra*, 28 for *data*, 32 items for *geometry*, 31 for *measurement*, and 56 for *number*. Nineteen of the 194 items were partial-credit items, each with three score categories. In order to compare and discuss the results obtained via the subdimension model (more precisely its extension to the partial credit model), I also analyzed the data using the unidimensional model, the testlet model, and the (unrestricted) multidimensional model.

3.4.2 Results and Discussion

Table 1 summarizes the results of the estimated means[1] and variances for the distributions, their reliabilities, the correlations, and the –2 log likelihoods for the different models. The index M (main dimension) refers to the unidimensional latent trait; the indices 1 to 5 refer to the content domains algebra, data, geometry, measurement, and number, respectively.

On comparing the variance obtained for the main dimension of the subdimension model with the variance obtained via the unidimensional model, we find that the actual variance is underestimated in the unidimensional case because of the local dependencies of the items of the same content domain. Although the test was constructed to be unidimensional, the subdimension model shows an increase in measured variance. The variance rises from 1.19 to 1.25, which is equivalent to an increase of about 5%. The increase in variance accords with findings by other authors (e.g., Sireci et al., 1991; Wang & Wilson, 2005b; Yen, 1993).

[1] The means and correlations for the testlet model given in Table 3.1 are not estimated but instead display the anchor values of the parameters; the testlet model is constrained on the cases given that this constraint is the only one that yields an optimum model fit.

Table 3.1

Results of the re-analysis of the US TIMSS 2003 mathematics achievement test

	Unidim.		Testlet		Subdim.		Multidim.	
Parameter	*Estimate*	*Reliability*	*Estimate*	*Reliability*	*Estimate*	*Reliability*	*Estimate*	*Reliability*
σ^2_M	1.19	0.820	1.22	0.812	1.25	0.816		
σ^2_1 / σ^2_{S1}			0.21	0.141	0.14	0.148	1.45	0.767
σ^2_2 / σ^2_{S2}			0.19	0.103	0.13	0.113	1.74	0.757
σ^2_3 / σ^2_{S3}			0.15	0.096	0.15	0.145	0.86	0.722
σ^2_4 / σ^2_{S4}			0.08	0.053	0.10	0.107	1.48	0.781
σ^2_5 / σ^2_{S5}*			0.07	0.062	0.05		1.41	0.800
μ_M	0.02		0.00		0.00			
μ_1 / μ_{S1}			0.00		0.12		-0.04	
μ_2 / μ_{S2}			0.00		0.29		0.43	
μ_3 / μ_{S3}			0.00		-0.29		-0.15	
μ_4 / μ_{S4}			0.00		-0.20		-0.29	
μ_5 / μ_{S5}*			0.00		0.09		0.12	
r_{12} / r_{S12}			0.00		-0.32		0.88	
r_{13} / r_{S13}			0.00		-0.26		0.85	
r_{14} / r_{S14}			0.00		-0.49		0.87	
r_{15} / r_{S15}*			0.00		-0.03		0.92	
r_{23} / r_{S23}			0.00		-0.36		0.84	
r_{24} / r_{S24}			0.00		-0.26		0.90	
r_{25} / r_{S25}*			0.00		-0.14		0.90	
r_{34} / r_{S34}			0.00		-0.14		0.90	
r_{35} / r_{S35}*			0.00		-0.51		0.89	
r_{45} / r_{S45}*			0.00		0.09		0.95	
Estimated Param.	214		219		228		228	
-2 Log Likelihood	275738.4		275380.2		275311.4		275022.6	

Note. * Calculated via plausible values.

If we look at the given reliabilities for the main dimensions, it becomes even clearer that the reliability given for the ability estimates is overestimated in the unidimensional case. Despite the subdimension model allowing for a gain in measured variance, the given reliability of its estimates is still lower than that of the unidimensional estimates. Essentially, the true reliability of the ability estimates calculated via the unidimensional model is *smaller* than that given for the main dimension of the subdimension model.

A difference between the multidimensional model and the subdimension model that becomes apparent on looking at the results is that the absolute variances of the latent traits measured by the subtests are closer to one another when the subdimension model is used than when the multidimensional model is used. While use of the multidimensional model shows subtest variances ranging from 0.86 to 1.74, the (absolute) variances obtained using the subdimension model range from only 1.35 to 1.40. This difference reflects the inability of the subdimension model to fully model the differences between the subtests due to their different variances. The estimated likelihoods of the two models provide a further indication of the extent to which the subdimension model is capable of modeling the differences between the subtests. The likelihood deviances (-2 log likelihood) of using the multidimensional and the subdimensional models are 275,022.6 and 275,311.4, respectively. The likelihood deviance for the unidimensional model, however, is 275,738.4; its difference of just 715.8 within the multidimensional model reflects the unidimensional construction of the measure. Nevertheless, the sub-dimension model does close the gap between the unidimensional and the multidimensional by about 50%.

Besides the differences in measurement precision and model fit, the interpretational differences of the measures provided by the two models should be of particular interest to test developers and analysts. In the case of the subdimension model, it is not the reliabilities of the *total* subtests that are measured but the reliabilities of the differences *between* the subtests. This is particularly interesting if the tests are being used to analyze, for example, student profiles constructed via the subtests. Here, the reliabilities provide a measure of how reliable differentiating these students according to these profiles will be and so help develop tests that provide especially reliable measures in these terms.

For the given empirical example, the results with subtest-specific reliabilities ranging from 0.107 to 0.148 indicate that an interpretation of the subtest-specific variances—that is, of differences between the subtests—need to be interpreted with caution. On the other hand, the correlation estimates for the subtest-specific variances provided by the subdimension model bring greater transparency to the differences between the subtests. In the multidimensional model, the large proportion of common variance dominates the correlation estimates and these differ, at most, by 0.11 (from 0.84 to 0.95), and the estimated correlations for the subdimension model differ by up to 0.60 (from -0.51 to 0.09). Nevertheless, the interpretation of the usually negative correlations provided by the subdimension model (resulting from the applied constraint) is not as intuitive. This is because a correlation of close to 0 for the relative subdimension-specific parameters is usually equivalent to a very high

correlation of the corresponding absolute ability estimates. An example of this relationship is provided via Dimensions 4 and 5 above. Although their estimated correlation in the multidimensional model is given as 0.95, the corresponding correlation in the subdimension model is 0.09. This example is a very unusual case of positive correlation, and, when compared with the other subtest correlations within the test, it represents a particularly high correlation between the two dimensions.

As a further comparison, and in order to show other differences, I also applied the testlet model to the data. The comparison of the likelihood deviances showed that, even given the very unfavorable conditions for the subdimension model due to the large difference between the smallest and the largest estimated variances, the model under discussion outperformed the testlet model. The difficulties for the testlet model to appropriately model the given data are best displayed by the relationship between Dimensions 4 and 5. As the results of the multidimensional analysis show, the correlation between these two dimensions is, on average, over .05 higher than the correlations between the remaining dimensions. While the subdimension model allows for any specific variance the dimensions have in common, the testlet model constrains the covariance of the respective testlet dimensions to zero. In other words, the large common part of their specific variances is not modeled and, in turn, the modeled variance is comparatively small; in the above example, it is less than half that modeled for the other dimensions.

In summary, the results of the re-analysis show that the application of the subdimension model allows for an increase in measurement precision for the students' unidimensional parameter estimates despite the very unfavorable conditions. Furthermore, the above results indicate that, for the analyses conducted above, the parameter estimates from the multidimensional model yield higher measurement precision for researchers endeavoring to interpret a person's abilities relative to the subtests.

3.5 Conclusion

The presented subdimension model offers test developers and analysts a way of handling the common conflict between theory and practice that arises whenever both unidimensional and multidimensional ability estimates of the same test are needed. Hitherto, tests were usually constructed in a unidimensional manner even if they included subtests that supposedly incorporated different characteristics. This practice meant that expected differences between these subtests due to test construction were minimized. Thus, any items particularly adept at showing differences between the subtests would probably not comply with the

(unidimensional) psychometric criteria used within the selection process after field trial of the items. Therefore, in order to gain interpretational value for the analysis of the subtests, psychometric criteria need to be based on a model that explicitly accounts for the differences in the subtests. The subdimension model provides exactly this opportunity. By allowing for correlations between the subtest-specific factors the model is particularly effective in accounting for differences and is able to outperform more restrictive models, like the testlet model (see discussions above).

Due to the restriction of the subdimension-specific parameters to yield a mean of zero, the correlations obtained under the subdimension model cannot be compared directly with those of the multidimensional model. For tests with large differences in subtest variances, this restriction also hinders the ability of the subdimension model to model, to full extent, the LID brought about by the different subtests. The advantages of the model become particularly apparent, however, when the variances of the measured subdimensions are approximately equal. In these cases, the subdimension model yields results almost equivalent to those of the unrestricted multidimensional model. With large-scale assessment studies that use matrix-sampling for administering the items and detailed background information for estimating person parameters, the chances of obtaining favorable conditions for the subdimension model are particularly high. Additionally, and/or in other cases, it might also be possible to provide more favorable conditions by adjusting the subtests for the differences in variances apparent after the field trial and by, for example, using different numbers of items for each subtest.

Another benefit of the subdimension model becomes apparent in regard to large-scale assessments when the mentioned matrix sampling for items is used. In these cases, each student receives only one booklet containing a subset of items, which means that several different booklets are needed to administer all items. (TIMSS 2003 used 12 different booklets.) The construction of these booklets typically endeavors to link the items that measure the same construct and to balance item-difficulty differences due to positional effects. An additional balancing of the booklets according to the number of items from the same subtest is usually not feasible. In this instance, the various booklets frequently end up including more items of a particular subtest and fewer of another. A student who performs particularly well in one subtest and poorly in another will effectively get different scores for the overall test depending on the booklet he or she completed. More specifically, this is because the common unidimensional Rasch model does not account for differences in sets of items due to different subtests. The subdimension model, however, accounts for these differences, and thereby yields more adequate individual measures. Although researchers

conducting large-scale assessments are usually not interested in achievement scores for single students, estimation of adequate ability estimates for single students is important because the calculation of adequate correlations (e.g., between a person's achievement score and his or her socioeconomic background) depends on adequate scores at the single-person level.

In addition to its ability to analyze tests measuring a general domain and multiple subdomains at the same time, the subdimension model seems to provide benefits for other applications as well. The application for vertical scaling, for example, is straightforward with the subtests representing tests given at different points in time; however, future research in this area needs to investigate how results arising from use of the subdimension model relate to other models used for vertical scaling. Furthermore, and beyond its application to empirical data, the subdimension model could be very usefully employed in simulation studies because of its ability to provide additional and more subtle information, as some of my recent work shows (Brandt, 2007b, in preparation).

Finally, another way of using the subdimension model could be to adjust Restriction 2 of the model so that the measured subtests are not balanced within the overall measure but instead are "assigned" (per definition) more weight—or relevance—than others, which means the characteristics of such models would have to be investigated as well. By providing a detailed description on how to calibrate the subdimension model using ConQuest, I hope that the gap between the development of new models and their application in practice becomes somewhat smaller and that a larger community than at present finds conducting research and practice via a model like the subdimension model a considerably more accessible proposition.

3.6 References

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

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.

Brandt, S. (2006). *Exploring bundle dependencies for the embedded attitudinal items in PISA 2006.* Paper presented at the International Objective Measurement Workshop (IOMW), Berkeley, CA.

Brandt, S. (2007a). *Applications of a Rasch model with subdimensions.* Paper presented at the 2007 Annual Conference of the American Educational Research Association (AERA), Chicago, IL.

Brandt, S. (2007b). *Item bundles with items relating to different subtests and their influence on subtests' measurement characteristics.* Paper presented at the 2007 Annual Conference of the American Educational Research Association (AERA), Chicago, IL.

Brandt, S. (in preparation). The impact of local item dependence on multidimensional analyses.

Brandt, S. (in preparation). Modeling tests with subtests.

Carstensen, C. H. (2000). *Mehrdimensionale Testmodelle mit Anwendungen aus der pädagogisch-psychologischen Diagnostik [Multidimensional test models with applications from the educational and psychological diagnostic]*. Kiel: Leibniz Institute for Science Education (IPN).

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., et al. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*(4), 4–19.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*, 41–54.

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist*, *17*, 475–483.

Humphreys, L. G. (1970). A skeptical look at the factor pure test. In C. E. Lunneborg (Ed.), *Current problems and techniques in multivariate psychology: Proceedings of a*

conference honoring Professor Paul Horst (pp. 22–32). Seattle, WA: University of Washington.

Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning* (pp. 87–102). New York: Plenum.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Social Psychology*, *71*, 327–333.

Martin, M. O., Mullis, I. V. S., & Chrostowski, S. J. (Eds.). (2004). *TIMSS 2003 technical peport*. Chestnut Hill, MA: Boston College.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Mullis, I. V., Martin, M. O., Gonzales, E. J., & Chrostowski, S. J. (Eds.). (2004). *TIMMS 2003 international mathematics report*. Chestnut Hill, MA: Boston College.

Mullis, I. V., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzales, E. J., et al. (2003). *TIMMS assessment frameworks and specifications 2003* (2nd ed.). Chestnut Hill, MA: Boston College.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*(19), 73--90.

Organisation for Economic Co-operation and Development (OECD). (2005). *PISA 2003 technical report*. Paris: Author.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion [Text book test theory, test construction]*. Bern, Göttingen, Toronto, Seattle, WA: Verlag Hans Huber.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237–247.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*(3), 247–260.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–202.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, *28*(3), 197–219.

Wang, W.-C., & Wilson, M. (2005a). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, *29*(4), 296–318.

Wang, W.-C., & Wilson, M. (2005b). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER ConQuest: Generalized item response modeling software*. Melbourne, Australia: Australian Council for Educational Research.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213.

4 Robustness of Multidimensional Analyses against Local Item Dependence (Brandt, S. (2012). *Psychological Test and Assessment Modeling*, *54*, 36–53.)

An essential assumption of item response theory is local item independence; that is, beyond the variance due to one or several latent traits, the items of a test are supposed to measure, the items show no additional common variance. The negative impact of a violation of this assumption, which is denoted as local item dependence (LID), has been reported by many authors, and it has been shown that an inappropriate assumption of local item independence results in an overestimation of test information, model fit and reliability and an underestimation of the measurement error (see, e.g., Rosenbaum, 1988; Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wainer, Bradlow, & Wang, 2007; Wen-Chung Wang & Wilson, 2005a; Yen, 1984, 1993).

It is reasonable to assume that these effects that have been investigated based on unidimensional analyses apply to the single dimensions of a multidimensional analysis in the same way. That is, without appropriate statistical modeling of existing LID their respective test information and reliability is overestimated, and the measurement error is underestimated. However, the generalization in multidimensional analyses is not that straightforward since LID can not only occur within a dimension but also across several dimensions. If, for example, two items with the same stimulus measure different constructs in a two-dimensional analysis, these two items are expected to have a higher correlation beyond that of the respective constructs they are supposed to measure. It is therefore expected that this additional correlation has an impact on the estimated covariance for the two dimensions, or more precisely, that the covariance will be overestimated due to the local item dependence. If, on the other hand, a two-dimensional analysis includes local dependence within the dimensions, it is expected that the covariance of the two dimensions is underestimated since the reliability of the respective dimension will be overestimated and the correction for the disattenuation of the covariance due to measurement error will not be appropriate.

The investigation of the impact of the effects of LID on multidimensional analyses and in particular on the corresponding covariance matrices is important since the decision on whether a data set should be interpreted unidimensional or multidimensional relies on these results. Maul (in press), for example, reanalyzed the dimensional structure of a well known measure of emotional intelligence, and found that models without consideration of LID yield a multidimensional structure; models with consideration of LID, however, did not. Wang, Cheng and Wilson (2005) investigated the impact of LID for items across tests connected by common stimuli. After applying different administration designs and models with and

without consideration of LID, they found a significant impact for tests with a "parallel" design, that is, items having a common stimulus but referring to separate psychological constructs. And without consideration of LID the tests had a correlation that was .36 higher than with consideration of LID.

Despite the possible significance, as depicted by the works above, and despite the widespread application of multidimensional analyses in large-scale assessments (e.g., Martin, Gregory, & Stemler, 2000; OECD, 2002), there has not been much emphasis to date on the investigation of the possible impact of LID on multidimensional analyses. The presented simulation study, therefore, was conducted in order to depict the possible impact of LID depending on the size of the LID and the chosen administration design for a given multidimensional construct. Furthermore, the results of the simulation study provide insight into how the differences between the results with and without consideration of LID arise.

4.1 Preparatory Considerations for the Design of the Simulation Study

A typical source of LID are item bundles. An item bundle is a set of items (also denoted as testlet) that is linked to a common stimulus (cf. Wainer & Kiely, 1987). The common stimulus for these items usually results in local item dependencies referred to as item bundle effect. The actual impact of such LID on multidimensional analyses depends on a large variety of factors: the number of dimensions, the numbers of items per dimension, the numbers of items in each item bundle, the correlations between the dimensions, the extents of the item bundle effects, and the extents of the variances of the single dimensions. A simulation study considering just two different conditions for each of these factors will result in a total of 128 different conditions for the overall test. In order to reduce the amount of test conditions, it was therefore chosen to generate different conditions on the bases of an exemplary, given multidimensional construct with a fixed amount of items and item bundles, and only the extents of the item bundle effects, the correlations of the dimensions, and the test design characteristic (see description below) are varied.

The chosen multidimensional construct roughly follows the structure of the mathematics achievement test of the Programme for International Student Assessment (PISA) 2003 (OECD, 2005). The PISA mathematics achievement test comprises four dimensions: Quantity, Change and Relationships, Space and Shape, and Uncertainty. Each dimension is measured by 20, 22, 20, and 22 items respectively. Seventy-six of these eighty-four items are dichotomous, seven have three score categories, and one has four score categories. Forty-two of these items were administered within item bundles. In order to give an impression of the

extents of the item bundle effects in the real data set, the extents of the item bundle effects were investigated by a reanalysis for the German subsample using the Rasch testlet model (Wen-Chung Wang & Wilson, 2005a; 2005b; see description below). The Rasch testlet model is a restricted hierarchical model (Holzinger & Swineford, 1937; Li, Bolt, & Fu, 2006) that bases on the testlet model by Bradlow, Wainer, and Wang (1999) and is an extension of the standard Rasch model (Rasch, 1980) by an additional parameter which describes the interaction between persons and items within an item bundle. Wang and Wilson denote this parameter $\gamma_{nd(i)}$, representing the interaction between person n (n=1,..., N, and N the number of persons) and item i within item bundle $d(i)$ (i=1,..., I, and I the number of items; $d(i)$=1,..., D, and D the number of item bundles). The model equation is

$$\log(p_{ni1}/p_{ni0}) = \theta_n - b_i + \gamma_{nd(i)}, \tag{4.1}$$

where p_{ni1} and p_{ni0} are the probabilities of scoring 1 and 0 on item i for person n, respectively, θ_n is the ability of person n, and b_i is the difficulty of item i. For the identification of the model several constraints have to be applied. In order to fix the locations of the scale for the latent trait and those for the item bundle effects, the means of all dimensions are set to zero. For rotational invariance the covariances of the dimension θ_n with the dimensions for the item bundle effects are set to zero. Furthermore, the item bundle effects themselves are assumed to be independent to each other.

The results of the analysis using the Rasch testlet model are given in Table 4.1. They show that the effects differ strongly across item bundles and range from 0.34 to 2.94, with an average variance of 1.20 for the item bundle effects and a variance of 1.95 for the measured overall mathematics achievement.

Table 4.1

Calibration Results for the German Subsample of the Mathematics Achievement Test of PISA 2003 Using the Rasch Testlet Model

Dimension	Items	Variance
Mathematics Achievement	1-84	1.95
Bundle 1	3, 4, 5, 6	1.43
Bundle 2	11, 12, 13	2.94
Bundle 3	21, 22	1.09
Bundle 4	23, 24, 25	0.34
Bundle 5	26, 27, 28, 29	0.52
Bundle 6	31, 32, 33	1.61
Bundle 7	34, 35	1.36
Bundle 8	36, 37	0.75
Bundle 9	39, 40	0.45
Bundle 10	47, 48, 49	0.54
Bundle 11	51, 52	0.68
Bundle 12	64, 65, 66	1.87
Bundle 13	70, 71	2.76
Bundle 14	73, 74, 75	0.38
Bundle 15	78, 79	0.87
Bundle 16	82, 83	1.61

Besides the extents of the item bundle effects, the used test design plays an important role for the impact of the local item dependencies. Following the terminology of Wen-Chun Wang et al. (2005), possible test designs for multidimensional constructs are sequential and parallel test designs. These two test designs are exemplary depicted in Figure 4.1 for a three-dimensional construct comprising six item bundles with three items each. In the sequential test design on the left, each dimension comprises six items from two different item bundles. In the parallel test design on the right, each dimension comprises as well six items but from six different item bundles. That is, in the first case each item bundle measures only a single dimension, whereas in the latter each item bundle measures all three dimensions. An uncountable number of other multidimensional test designs that are mixtures of the parallel and the sequential test design are possible. However, the parallel and the sequential test designs can be considered as the extremes of these possible test designs. In order to investigate the full range of the possible impact on multidimensional analyses, it is therefore useful to consider the results of a simulation study for these extremes. Additionally, parallel and sequential test designs are common in test construction. A well known test using a parallel design, for example, is the multidimensional Self-Description Questionnaire III by Marsh and O'Neill (1984). An example for a sequential test design is given by the PISA

study, in which the domains mathematics, reading, and science are measured with item bundles which are entailing items from one distinct dimension only (cf. OECD, 2005).

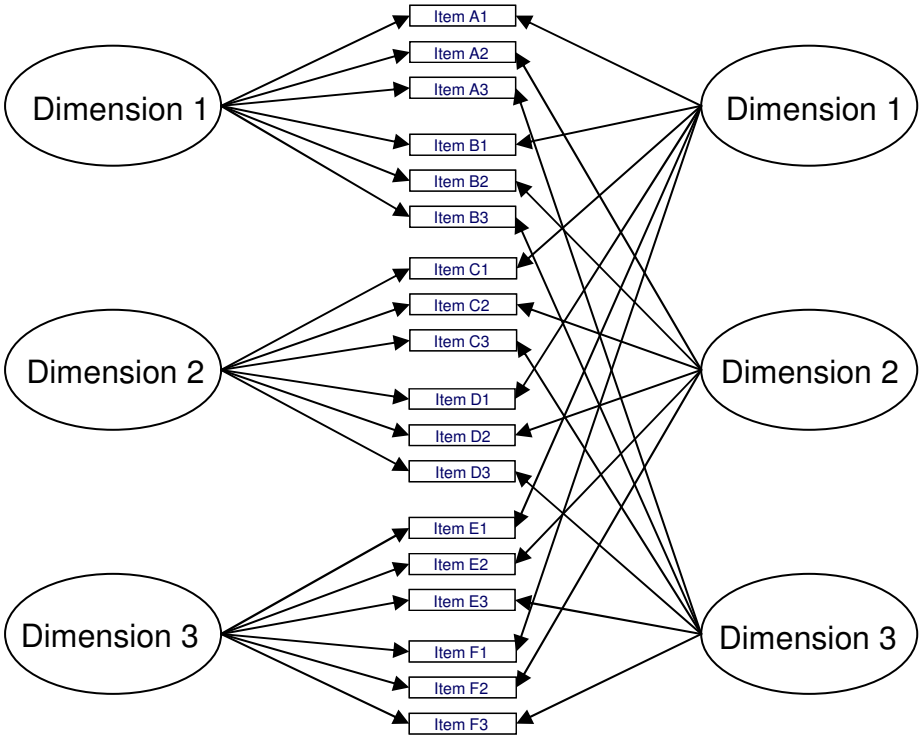

Figure 4.1. Depiction of a three-dimensional sequential test design (on the left) and a three-dimensional parallel test design (on the right).

4.2 Simulation Study Design

The following three test characteristics are varied in the conducted simulation study: (a) the test design, (b) the extent of LID due to item bundles, and (c) the correlations between the measured dimensions.

For the above given reasons, the considered test designs are a test with item bundles in a sequential test design, a test with item bundles in a parallel test design, and additionally a reference test without item bundles, that is, without local item dependencies. Following the definition of small, medium, and large item bundle effects given by Wen-Chung Wang and Wilson (2005b), variances of 0.5, 1.0 and 2.0 for the item bundle effects are considered with variances of 2.0 for the measured latent traits. The considered levels of correlations between the latent traits are .5, .7, and .9, representing medium to high correlations and were chosen

according to the extents of correlations that are typically observed in multidimensional constructs. In particular, the consideration of a correlation of .9, therefore, does not question whether such dimensions might in fact be unidimensional or not but solely corresponds to regularly reported correlations (cf. Martin et al., 2000; OECD, 2005).

As previously mentioned the used multidimensional construct follows that of the mathematics achievement test of PISA 2003. Some test characteristics are modified though, in order to allow for a more lucid presentation of the results. The number of items is adjusted to be 20 per each dimension, and each item is assigned to an item bundle of four items (according to the given test design), resulting in a total of 20 item bundles for the test. The item difficulty parameters are taken from a unidimensional, dichotomous[1] reanalysis of the German PISA 2003 mathematics achievement data and range from -2.79 to 2.56 except for one item with a difficulty of 4.07. Further, the variances for the four dimensions are set to 2.0 with a mean ability of zero. The variance of 2.0 was chosen in order to consider a scale close to that of the empirical data; it coincides with the estimated variance for the above presented example using the Rasch testlet model.

4.3 Data Generation and Analysis

The data generation is based on a multidimensional extension of the Rasch testlet model by Wang and Wilson (2005a; 2005b). The single steps in order to generate the simulation data according to the model are as follows (cf. Wen-Chung Wang & Wilson, 2005b):

1. Person parameters for the four-dimensional multivariate distribution are generated using ConQuest (Wu, Adams, & Wilson, 1998).
2. Normally distributed variables representing the item bundle effects are generated using SPSS for Windows.
3. The generated person parameters (θ) and random variables (γ_k), as well as the predefined item parameters (b) are used to calculate the corresponding answer probabilities using Equation 4.1.
4. The calculated answer probabilities are compared to a random number from the uniform [0, 1] distribution, and the simulated item response is defined as 1 if the random number is less than or equal to the associated probability, and 0 otherwise.

[1] Only answers in the highest score category received a score of 1; all other answers received a score of 0.

For each of the 21 test conditions with the characteristics given in the previous section, one hundred data sets with 1000 cases each are generated. Each data set is analyzed using the unidimensional Rasch model, the multidimensional Rasch model (Adams, Wilson, & Wang, 1997; Rost, 1996), and the Rasch subdimension model (Brandt, 2008, 2010). While the analyses using the multidimensional model show the extents of the observed bias due to the generated item bundle effects, the analyses using the subdimension model show the origin of the observed bias. The results of the unidimensional model were included as reference in order to depict possible biases on decisions on the dimensionality of data sets.

The unidimensional Rasch model coincides with the above given Rasch testlet model without the extension by the parameters γ. That is, the model equation is given by

$$\log(p_{ni1}/p_{ni0}) = \theta_n - b_i\,. \tag{4.2}$$

For the given test data, the multidimensional model applied for the analysis can be expressed as the multicategorical, multidimensional Rasch model (Rasch, 1961; cf. Rost & Carstensen, 2002):

$$\log(p_{ni1}/p_{ni0}) = \theta_{nd} - b_{id}\,, \tag{4.3}$$

where θ_{nd} is the ability of person n for dimension d, and b_{id} is the difficulty of item i for dimension d.

The additionally applied subdimension model corresponds to a modified hierarchical model (Holzinger & Swineford, 1937), in which each item loads on a general factor, in the context of the subdimension model referred to as main dimension, and a specific factor, in the subdimension model referred to as subdimension. In contrast to the simple hierarchical model, however, the specific factors (subdimensions) are assumed to correlate. The definition of the subdimension model is given by

$$\log(p_{ni1}/p_{ni0}) = \theta_n + \gamma_{nd(i)} - b_i\,, \tag{4.4}$$

where $d(i)$ is defined as the subdimension of item i; $\gamma_{nd(i)}$ is the strength or weakness of person n in subdimension $d(i)$ relative to its ability in the main dimension; and p_{ni1}, p_{ni0}, θ_n, and b_i are defined as above (cf. Brandt, 2008). For the identification of the model, the person parameters are constrained to a mean of zero, and the covariance between the main

dimension and the subdimensions is set to zero, which is common to all hierarchical models. However, in contrast to the testlet model the covariances between the subdimensions are not constrained to zero but the sum of the specific abilities for each person is constrained to zero; that is, $\sum_d \gamma_{nd} = 0$ for all n = 1, …, N. For the analyses of the simulation data, the subdimensions correspond to the dimensions of the four dimensions of the multidimensional construct, while the main dimension represents the general factor measured commonly by these four dimensions.

All considered models are special cases of the multidimensional random-coefficients multinomial logit model (MRCMLM; Adams et al., 1997) and can therefore be estimated using ConQuest (Wu et al., 1998). The unidimensional estimations were conducted using the Gauss-Hermite quadrature integration method with 100 nodes. Due to their higher complexity the multidimensional and the subdimensional estimations were estimated using the Monte Carlo integration method with 4000 nodes. The convergence criterion for all estimations was 0.01 for the change in parameters.

The estimation results of the models are compared with regard to their deviances (-2 log likelihoods)[1], their correlations, and their variances for the given four-dimensional construct depending on the extent of the generated correlation, the size of the generated item bundle effect, and the chosen test design.

4.4 Results

In order to facilitate a lucid presentation, the results of the 100 calibrated data sets per test condition as well as the dimensions' variances and correlations (which were generated to be equal) were summarized calculating their means. The results of the unidimensional model were not considered separately for the sequential and the parallel test design since the model yields equal results for both test designs.

Figure 4.2 depicts the changes in deviance for the unidimensional and multidimensional analyses in dependence of the size of the item bundle effects. With generated higher correlations between the four dimensions the fit of the unidimensional model is, as expected, closer to that of the multidimensional model. Furthermore, for all analyses the model fit decreases (i.e., the deviance increases) with increasing item bundle effects. However, for the multidimensional analyses the magnitude of the decrease in model fit depends on the chosen

[1] Results for model fit indices such as Akaike's information criterion (Akaike, 1974) or the Bayesian information criterion (Schwarz, 1978) are not reported since the observed differences in the deviances are of such extent that the criteria do not provide additional information.

test design. In the presence of item bundle effects, the data according to a sequential test design yields a better model fit than the data according to a parallel test design. Therefore, the difference between the fit of a unidimensional model and a multidimensional model is not only affected by the correlation between the considered dimensions but as well by the size of the item bundle effects and the chosen test design. The relatively stronger decrease in model fit for the parallel test design is particularly notable for the case of large item bundle effects with generated correlations of .9, in which the model fit of the unidimensional model in fact exceeds that of the multidimensional model. Here, the average deviance for the multidimensional model is 92559.5 while the deviance of the unidimensional model is 92418.8 (cf. Table A3 in the Appendix). For the sequential test design, however, the magnitude of the difference in model fit between the unidimensional and the multidimensional model seems to be comparatively independent of the size of the item bundle effects.

Figure 4.3 depicts the change of the estimated correlations of a multidimensional calibration depending on the generated correlations, the size of the item bundle effects, and the chosen test design. Corresponding to the model fit, the extents of the estimated correlations for the four-dimensional construct depend on the size of the item bundle effects and the chosen test design. The differences between the estimated correlations in dependence of the chosen test design are very similar for all three generated correlations. For small, medium, and large item bundle effects, the sequential test design on average results in correlations that are 0.02, 0.07, and 0.23 lower than for the parallel test design (cf. results in Appendix B). While these differences solely depend on the extent of the item bundle effect, the biases of the respective estimates in comparison to the originally generated correlations depend on the level of the generated correlation. While the bias is of equal magnitude for a generated correlation of .5, for a correlation of .9 only the sequential test design shows bias and the parallel test design seem to provide unbiased correlation estimates, independently of the size of the item bundle effects.

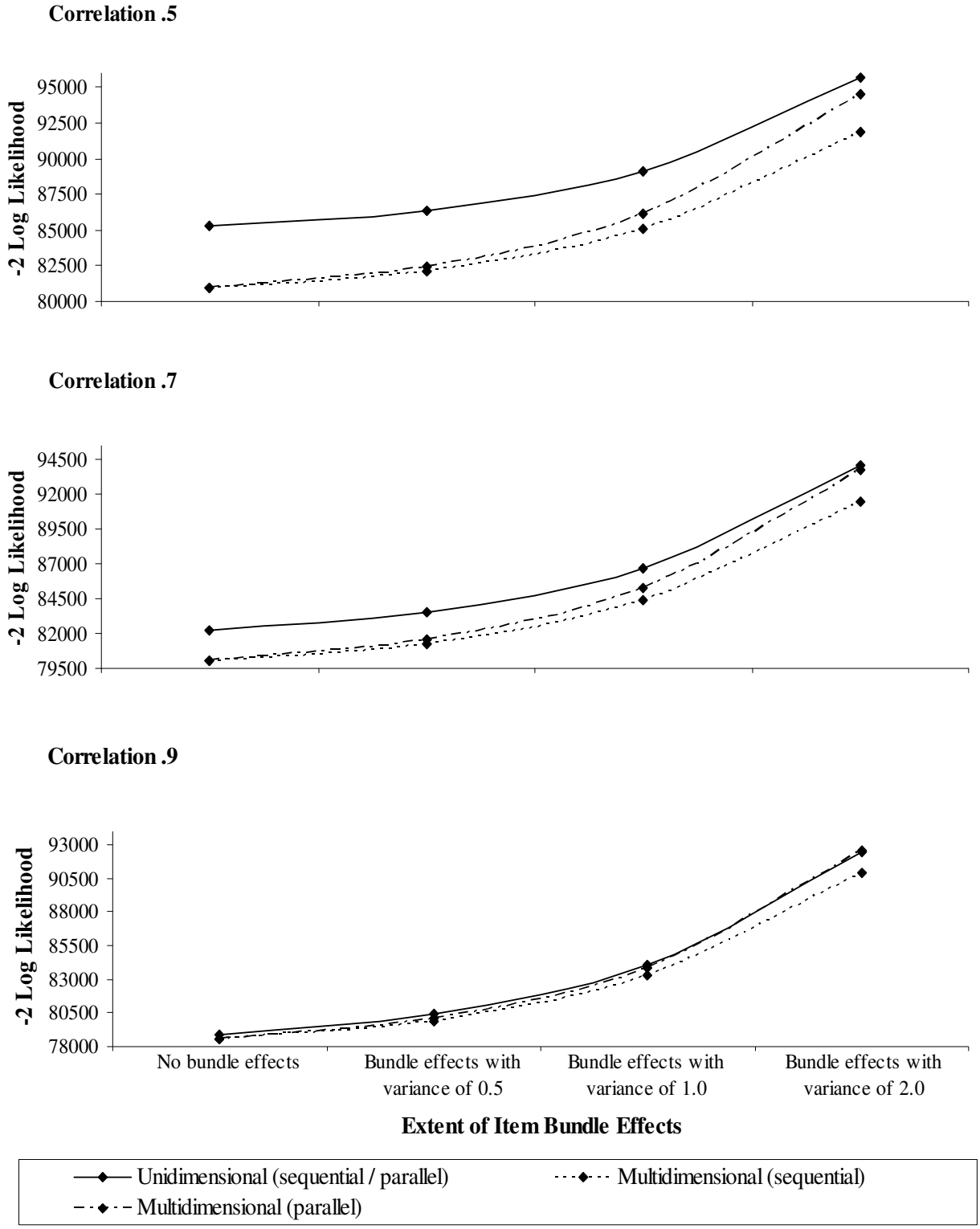

Figure 4.2. Likelihoods for the unidimensional and multidimensional estimations of the four-dimensional construct with correlations of .5, .7, and .9; a sequential or a parallel test design; and varying extents of item bundle effects.

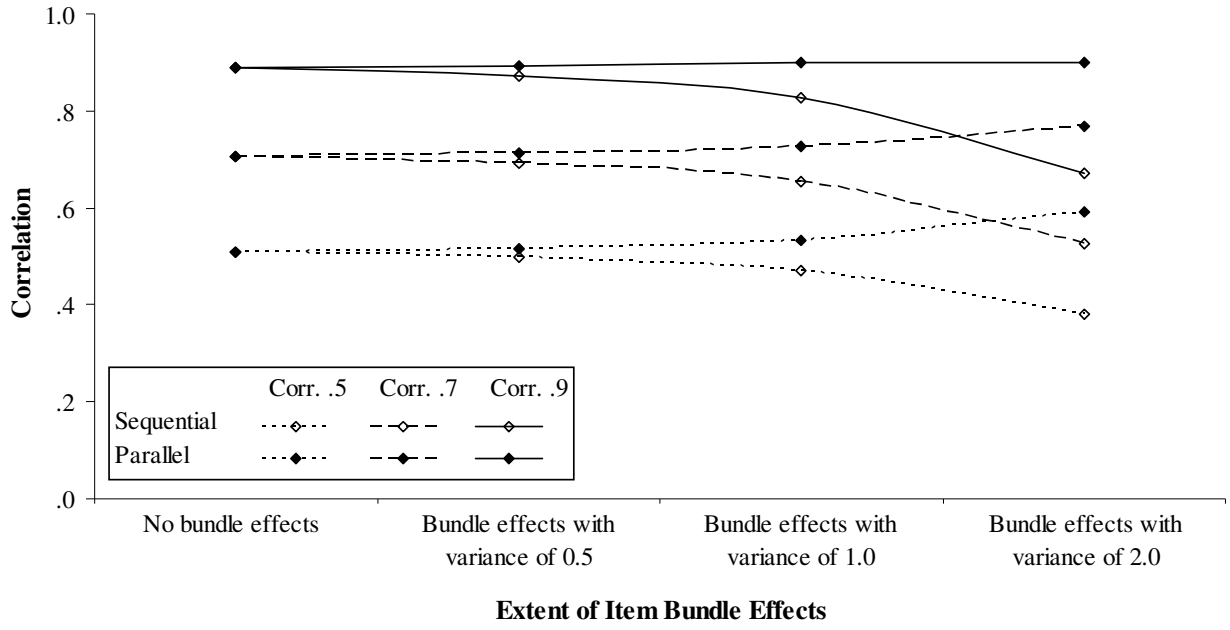

Figure 4.3. Estimated correlations for the four-dimensional construct with generated correlations of .5, .7, and .9; a sequential or a parallel test design; and varying extents of item bundle effects.

In order to consider the origins of the differing biases in more detail, the results of the subdimension model are considered. The model allows separating the variance components of a multidimensional construct into the common variance component (responsible for the correlation of the dimensions) and the dimension-specific variance component. The variance components corresponding to the results presented in Figure 4.3 are depicted in Figure 4.4. Here, larger proportions of unidimensional common variance within the estimated total variance represent higher correlations. Figure 4.4 demonstrates that for all considered test conditions the estimated total variance decreases with increasing item bundle effects. The common unidimensional variance component and the dimension-specific variance component are affected differently, though. For the sequential test design with generated correlations of .7, for example, the absolute variance of the dimensions-specific variance component stays almost unchanged (0.44 to 0.42 for no to large item bundle effects) while the common variance components decreases from 1.56 to 0.77 (again for no to large item bundle effects). Considering the differences between the sequential and the parallel test design, the item bundle effects lead to a larger decrease in total variance when using a parallel test design. The difference in the decrease of the total variance is mainly attributable to the dimension-specific variance components, however. While the unidimensional variance components differ at most

at 0.08, the differences for the dimension-specific variance components extend to 0.30. The difference between the two test designs gets most visible for correlations of .9 and large item bundle effects. Here, the dimension-specific variance component is more than three times as large for the sequential test design as for the parallel test design (0.30 vs. 0.08). Furthermore, a comparison to the result of the calibration without item bundle effects shows that for correlations of .9 the introduction of item bundle effects results for the sequential test design in an increase of the dimension-specific variance beyond the dimension-specific variance that was originally generated. While the generated dimension-specific variance is 0.19, the introduction of item bundle effects results in dimension-specific variances of .20, 0.23, and 0.30 for small, medium, and large item bundle effects, respectively.

4.5 Discussion

For a better understanding of the results, it is necessary to recall the origin of the two test designs. By assigning each item of an item bundle to the same dimension, each item bundle in a sequential test affects a particular single dimension. In the parallel test design, on the other hand, each item of an item bundle loads on a different dimension. From a single dimension's perspective, therefore, not an item bundle is added to the dimension but just a single item. That is, here, the dimensions do not include item bundles in its actual sense. And even though the generated item bundle effects are still present, their impact is not that of LID but that of added independent error variances on the items' answers. This is in contrast to the item bundle effects in the sequential test designs in which the items of an item bundle commonly influence the same dimension and, therefore, not only result in added random error variance but in common error variance (that is, they introduce LID into the data).

The parallel test design's characteristic that it does not include LID in its actual sense is emphasized by the results depicted in Figure 4.2. Here, the multidimensional calibration of the data for the parallel test design for correlations of .9 and large item bundle effects results in a worse model fit than the calibration of the unidimensional model. A result which is theoretically impossible if the test data responds to the assumptions of IRT. It is attributable to the fact that the unidimensional calibration includes LID and therefore overestimates its model fit while the multidimensional calibration (for the parallel test design) does not include the LID and therefore not overestimates its fit. For the same reason the multidimensional calibration always provides a better model fit for the sequential test design than for the parallel test design.

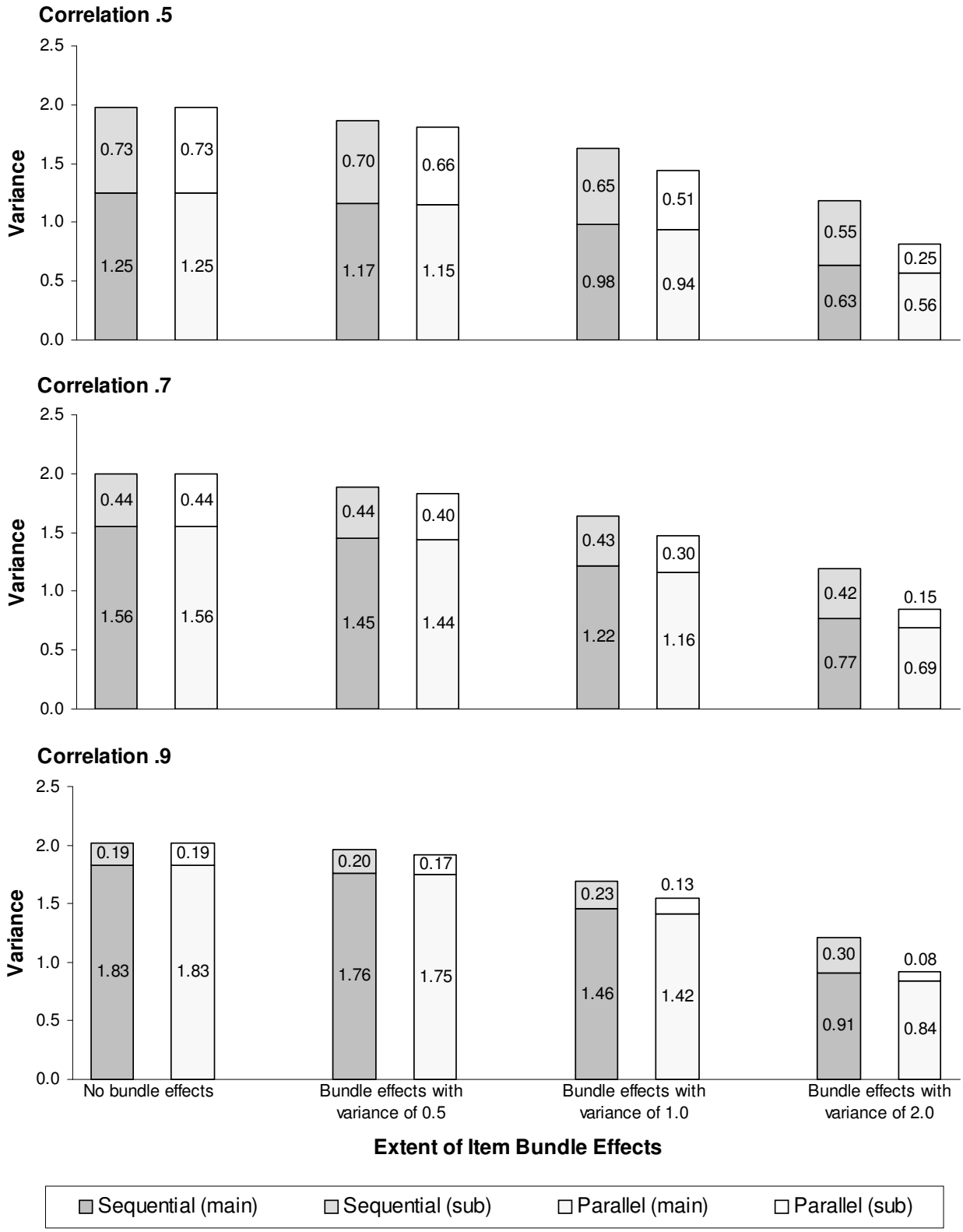

Figure 4.4. Depiction of the dimensions' unidimensional (main dimension = main) and dimension specific (subdimension = sub) variance components depending on the generated correlations, a sequential or parallel test design, and the extent of item bundle effects.

The origin for the biases in the estimation of the correlations is demonstrated via the differing impacts on the variance components depicted in Figure 4.4. In all cases the LID included in the calibration of the data for the sequential test design results in an increase in the unidimensional and the dimension-specific variance component in comparison to the parallel test design. The increase of the two variance components is different, however. Since the single dimensions only include 20 items, the four items of an item bundle have a comparatively large dimension-specific effect; while the effect on the unidimensional variance component that is determined by a total of 80 items is comparatively smaller. The unidimensional variance components for calibrations of the sequential and the parallel test designs, therefore, only differ to a small amount, while the subtest-specific variances components show substantial differences. Particularly the results for correlations of .9, in which the subdimension-specific variances for the sequential test design increase beyond the actually existing[1], hereby, emphasize that variance is introduced into the measures that is solely due to the used item administration form and not due to the variance of the measured constructs. The presence of LID might therefore result in a falsely interpretation of differences between measured dimensions; assuming that differences are solely attributable to differences in the constructs while they might, in fact, partially origin in item bundle effects. The results in Figure 4.4 show that for a sequential test design with correlations of .7 and .9 already medium size item bundle effects result in dimension-specific variances that originate only by 69.7% and 56.5%, respectively, in the measured construct (0.30 of 0.43 and 0.13 of 0.23).

4.6 Conclusion

The results of the presented simulation study emphasize that LID not only biases the results of unidimensional calibrations but additionally biases the covariance estimates in multidimensional calibrations. Moreover, the chosen test design for the measurement of the multidimensional construct interferes with the impact of the LID and defines the direction of the bias. Considering that in practice test designs commonly are not as strict as the designs presented here but might consist of a mixture of item bundles that are attributed sequentially or parallel to different dimensions, the effect of the LID will often be hard to predict. The differences in the results for the two presented test designs, however, show that the effects

[1] The reliabilities of the subdimension-specific components are very low for correlations of .9. The added random variances due to the LID, therefore, only have a small impact here in contrast to the added common variances. For correlations of .7 and .5 the reliabilities of the subdimension-specific variance components increase; therefore, the added random variances there have a relatively larger impact, which prevents increases in the variances as observed for correlations of .9.

due to local item dependencies have to be separated into two different types of effects: (1) their effect as an error variance (visible via the results of the parallel test design) and (2) their effect as a redundantly modeled part of the measured latent trait (visible via the difference between the results for the parallel and the sequential test design).

Furthermore, the results underline the importance of the investigation of LID during test construction and test analysis in order to prevent an interpretation of biased results.

4.7 References

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153-168.

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments* (Vol. 1, pp. 51-70). Princeton, NJ: IEA-ETS Research Institute.

Brandt, S. (2010). Estimating tests including subtests. *Journal of Applied Measurement, 11*, 352-367.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*, 41-54.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*, 3-21.

Marsh, H. W., & O'Neill, R. (1984). Self Description Questionnaire III: the construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement, 21*, 153-174.

Martin, M. O., Gregory, K. D., & Stemler, S. E. (Eds.). (2000). *TIMMS 1999 technical report*. Chestnut Hill, MA: Boston College.

Maul, A. E. (in press). Examining the structure of emotional intelligence at the item level: New perspectives, new conclusions. *Cognition and Emotion*.

OECD. (2002). *PISA 2000 technical report*. Paris: OECD.

OECD. (2005). *PISA 2003 technical report*. Paris: OECD.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4, pp. 321-333). Berkeley: University oc California Press.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Rosenbaum, P. R. (1988). Items Bundles. *Psychometrika, 53*, 349-359.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion*. Göttingen: Verlag Hans Huber.

Rost, J., & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. *Applied Psychological Measurement, 26*, 42-56.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*, 247-260.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-202.

Wang, W.-C., Cheng, Y.-Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement, 65*, 5-27.

Wang, W.-C., & Wilson, M. (2005a). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*, 296-318.

Wang, W.-C., & Wilson, M. (2005b). The Rasch testlet model. *Applied Psychological Measurement, 29*, 126-149.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). ConQuest: Generalized item response modelling software. Camberwell, Victoria: Australian Council for Educational Research.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. M. (1993). Scaling performance assessments - strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

4.8 Appendix A

-2 Log Likelihoods for the Four-Dimensional Constructs With Correlations of .5 (Table A1), .7 (Table A2), and .9 (Table A3).

Table A1

	No bundle effect	Bundle effect with variance of 0.5	Bundle effect with variance of 1.0	Bundle effect with variance of 2.0
Unidimensional				
Sequential	85299.7	86384.0	89134.0	95633.7
Parallel	85299.7	86386.0	89153.5	95673.0
Multidimensional				
Sequential	80921.7	82081.1	85039.0	91878.1
Parallel	80921.7	82408.4	86134.9	94507.4
Subdimensional				
Sequential	80927.4	82088.4	85047.5	91880.8
Parallel	80927.4	82416.2	86142.6	94514.5

Table A2

	No bundle effect	Bundle effect with variance of 0.5	Bundle effect with variance of 1.0	Bundle effect with variance of 2.0
Unidimensional				
Sequential	82229.5	83483.7	86653.0	94060.7
Parallel	82229.5	83472.2	86666.6	94080.5
Multidimensional				
Sequential	80055.4	81272.9	84360.9	91483.2
Parallel	80055.4	81542.6	85312.0	93721.5
Subdimensional				
Sequential	80063.7	81281.3	84367.4	91489.6
Parallel	80063.7	81550.9	85319.5	93736.9

Table A3

	No bundle effect	Bundle effect with variance of 0.5	Bundle effect with variance of 1.0	Bundle effect with variance of 2.0
Unidimensional				
Sequential	78931.9	80410.4	84040.2	92406.3
Parallel	78931.9	80378.6	84033.2	92418.8
Multidimensional				
Sequential	78542.3	79926.4	83301.0	90921.3
Parallel	78542.3	80067.9	83891.7	92559.5
Subdimensional				
Sequential	78567.2	79944.9	83313.7	90930.2
Parallel	78567.2	80095.1	83928.1	92619.6

4.9 Appendix B

Estimated Correlations Using the Multidimensional Model

	No bundle effect	Bundle effect with variance of 0.5	Bundle effect with variance of 1.0	Bundle effect with variance of 2.0
Sequential				
Correlation of .5	0.51	0.50	0.47	0.38
Correlation of .7	0.71	0.69	0.65	0.53
Correlation of .9	0.89	0.87	0.83	0.67
Parallel				
Correlation of .5	0.51	0.51	0.53	0.59
Correlation of .7	0.71	0.71	0.72	0.77
Correlation of .9	0.89	0.89	0.90	0.90

4.10 Appendix C

Estimated Variances for the Generated Simulation Data

	No bundle effect	Bundle effect with variance of 0.5	Bundle effect with variance of 1.0	Bundle effect with variance of 2.0
Sequential				
Correlation .5				
Dimension (MD)	1.98	1.87	1.63	1.18
Subdimension (SD)	0.73	0.70	0.65	0.55
Main Dimension (SD)	1.25	1.17	0.98	0.63
Correlation .7				
Dimension (MD)	1.99	1.89	1.64	1.19
Subdimension (SD)	0.44	0.44	0.43	0.42
Main Dimension (SD)	1.56	1.45	1.22	0.77
Correlation .9				
Dimension (MD)	2.02	1.93	1.67	1.21
Subdimension (SD)	0.19	0.20	0.23	0.30
Main Dimension (SD)	1.83	1.76	1.46	0.91
Parallel				
Correlation .5				
Dimension (MD)	1.98	1.81	1.44	0.81
Subdimension (SD)	0.73	0.66	0.51	0.25
Main Dimension (SD)	1.25	1.15	0.94	0.56
Correlation .7				
Dimension (MD)	1.99	1.83	1.46	0.83
Subdimension (SD)	0.44	0.40	0.30	0.15
Main Dimension (SD)	1.56	1.44	1.16	0.69
Correlation .9				
Dimension (MD)	2.02	1.88	1.52	0.90
Subdimension (SD)	0.19	0.17	0.13	0.08
Main Dimension (SD)	1.83	1.75	1.42	0.84

Note. MD = Result of the multidimensional model; SD = Result of the subdimension model.

5 Increasing Unidimensional Measurement Precision Using a Multidimensional Item Response Model Approach (Brandt, S., & Duckor, B. (2013). *Psychological Test and Assessment Modeling*, *55*, 148.)

In their introduction to multidimensional measurement Briggs and Wilson (2003) note that measuring latent variables in the human sciences is a combination of "art and science." Following Wright and Masters (1982, p. 8) psychometricians in the Rasch IRT tradition describe the four basic scientific requirements for measuring as:

1. The reduction of experience to a *one dimensional* abstraction,
2. more or less comparisons among persons and items,
3. the idea of linear magnitude inherent in positioning objects along a line, and
4. a unit determined by a process which can be repeated without modification over the range of the variable.

The art of measuring, according to Briggs and Wilson, is the non-trivial task of finding the smallest "number of latent ability domains such that they are both statistically well-defined and substantively meaningful" (p. 88). Considering the complexity of this task, the authors acknowledge that "the art of measuring often hands us something that doesn't quite conform to these fundamental rules" (p. 88). Presenting the advantages of the multidimensional item response theory (IRT) approach Briggs and Wilson focused their work on the multidimensional model's capabilities in constructing statistically well-defined dimensions using a smaller number of items.

A fundamental tension with meeting the scientific requirements for measuring, however, entails the task of finding domains that are "substantively meaningful" and statistically well-defined. Too often, content experts can agree on whether a domain is substantively meaningful, though it may not appear to be statistically well-defined by psychometricians. Conversely, measurement experts can agree that a dimension is statistically well-defined, but can not persuade others as to a substantive definition to support its use. This problem is illustrated in large-scale studies such as the Programme for International Student Assessment (PISA). For policy stakeholders an interpretation of their country's student ability estimates in the mathematics dimension "Change and Relationship"[1] might not be substantively meaningful, since from a policy perspective they are being evaluated with the unidimensional results in the overall mathematics dimension on the PISA. More often, in these large-scale testings, the focus for stakeholders is on a particular country's performance (i.e. ranking)

[1] The mathematics framework in PISA differentiates the general mathematics ability on five different subscales: Quantity, Change and Relationships, Space and Shape, and Uncertainty (OECD, 2013).

across all tested dimensions. For educational researchers and practitioners, on the other hand, the results of a multidimensional analysis of the data set are potentially more meaningful and authentic to how children learn. Psychometric findings that inform the multi-dimensional nature of mathematics knowledge and skills acquisition are welcome. For these stakeholders, the focus is more often on a multi-faceted, complex analysis of the internal structure of the score data and making valid inferences about particular dimension or use of sub scores (APA, AERA, NCME, 1999).

In order to cope with these alternate and potentially conflicting needs, measurement specialists have attempted to satisfy different stakeholders by running analyses from two different but related lens. In the first instance, the data set is calibrated using a unidimensional IRT approach to yield global scores on a single scale. In the second instance, the data set is calibrated using a multidimensional approach (OECD, 2009). Due to a lack of plausible alternatives, this approach is common practice in PISA, and in other large-scale assessments such as TIMSS and PIRLS (Martin, Mullis, & Kennedy, 2007; Olsen, Martin, Mullis, Martin, & Mullis, 2008).

A main problem with this "re-run" approach is in the negligence of local item dependence (LID). If the data is multidimensional but interpreted unidimensionally, the neglected LID leads to an overestimation of reliability and biased parameter estimates (see, e.g., Wang & Wilson, 2005; Yen, 1980). In the search for alternatives, a growing variety of item response theory (IRT) models now focus on the estimation of unidimensional abilities for tests including subtests. Depending on whether the suspected LID due to the subtests is based on the type of test construction (e.g., due to the use of item bundles) or on the psychological construct that is to be measured (e.g., the assumption of sub-competencies), these models are typically denoted as testlet models (Bradlow, Wainer, & Wang, 1999; Wang & Wilson, 2005) or as hierarchical or higher-order models (de la Torre & Song, 2009; Gibbons & Hedeker, 1992; Sheng & Wikle, 2008), respectively. However, it has been shown that the testlet model and the higher-order model are formally equivalent and both are restrictions of the hierarchical model (Li, Bolt, & Fu, 2006; Rijmen, 2010; Yung, Thissen, & McLeod, 1999).

Additionally, all the mentioned models assume the existence of a unidimensional latent trait, and in doing so, assumptions regarding the LID or the sub-competencies are introduced in order to yield its identification. That is to say, it is assumed that any common variance between sub-competencies, or groups of items with LID, originates in the unidimensional latent trait to be measured. A further aspect of this approach, however, is that the weighting of the subdimensions (e.g., the testlet dimensions) for the general (overarching) dimension is

undefined. In the hierarchical model it is not clear how the subdimensions are weighted for the calibration of the general person ability estimates. The weighting of the subdimensions will depend on the subdimensions discrimination according to the general latent trait. Comparable to higher discriminating items in the 2-PL model (Birnbaum, 1968), here higher discriminating subdimensions will inadvertently receive higher weights.

The approach presented in this article does not assume the existence of a unidimensional latent trait but rather rests on the assumption of a truly multidimensional construct. Based on the generalized subdimension model (GSM) proposed by Brandt (2012), latent mean abilities are calculated from multidimensional scales in order to yield unidimensional ability estimates (without assuming the existence of a unidimensional trait). In contrast to the above-mentioned testlet and higher order models, the multidimensional latent variables can freely correlate in this modeling approach. Following the framework of Holzinger and Swineford's work (1937) one might conceptualize the GSM as a modified hierarchical model (cf. Brandt, 2012).

Of course one might propose an alternative approach: Why not simply obtain the unidimensional ability estimates, using the ability estimates of the multidimensional model, and then summarize these by a mean score? In order to do so, however, the ability estimates have to be standardized such that the dimensions yield equal variances (assuming an equal weighting of the dimensions), and further, the standardized estimates have to be summarized in a single score. To conduct the necessary calculations for the standardization, the usage of point estimates, for example, leads to additional measurement error: the estimated values of the dimensions' variances given by the multidimensional model include a measurement error. The standardization, that is, the multiplication of each ability estimate with the estimated variance therefore results in an additional inclusion of the measurement error of the variance estimate in each (standardized) ability estimate, and thereby in an increased overall measurement error for each ability estimate. Since in the GSM the necessary parameters are directly estimated without making a detour via point estimates, it avoids such an increase in measurement error.

The aim of this article is two-fold. First, we demonstrate the advantages of a latent mean ability approach for unidimensional estimates by showing its statistical advantages in yielding more precise and more appropriate (i.e., less biased) estimates. Second, we show the differences in interpretation due to an explicit weighting of the subdimensions, and contrast this approach with the implicit weighting of the subdimensions in a traditional unidimensional approach. We demonstrate the advantages of the GSM approach by applying

it to a classroom assessment literacy (CAL) scale currently used to measure pre-service teachers' assessment knowledge at a large public university in Northern California.

5.1 Background and Context of the CAL Scale

In the United States, accountability in the teaching profession is maintained, in part, through licensure process that includes the use of standardized testing batteries and performance assessments to warrant readiness to teach. The intended purpose of these large-scale instruments is to warrant a summative judgment about readiness to teach across a multitude of proficiencies such as planning, instructing, assessing and so forth. In California, as in most states, only a few items or tasks are used to assess pre-service teachers' competency in the domain of classroom assessment itself. State licensing bodies for teacher certification have set minimum standards for "safe beginners" in the area of classroom assessment (National Research Council, 2000) but many of these items/tasks focus narrowly on data interpretation. Information about an individual teacher's ability, skill, and/or knowledge of the principles and practices that can be employed to guide and improve their own classroom assessments is not measured by these large-scale instruments. This poses a problem for measuring classroom assessment literacy at the individual and program level across the teacher population in any meaningful way.

Building on previous research into the development of measurement expertise (B. Duckor, Draney, & Wilson, 2009; B. M. Duckor, 2006), a team of educational researchers and teacher educators have recently begun to develop a substantively meaningful instrument intended to measure teachers' proficiency with the major domains of assessment expertise as defined by national experts (Pellegrino, Chudowsky, & Glaser, 2001). Utilizing a modified version of the Assessment Triangle (Pellegrino et al., 2001) framework, the CAL scale advances a multi-dimensional theory of assessment literacy that draws upon three topics of knowledge to demonstrate proficiency with understanding classroom assessments—their design, use, and interpretation. While the researchers suspected that some of the proficiencies across the topics are strongly related, they nonetheless sought to carefully distinguish between each of the topics in the construct definition phase. A total of three construct maps (Wilson, 2005) were initially developed to represent each of the three major domains shown in Figure 5.1.

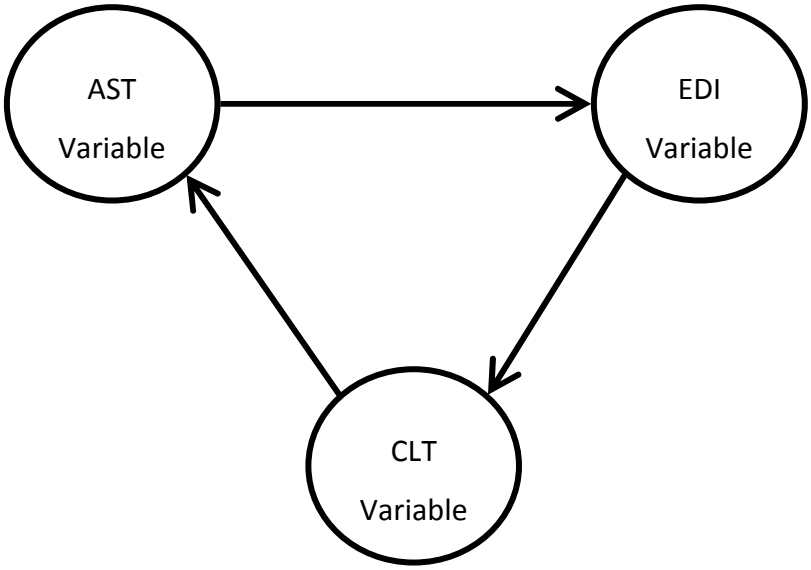

Figure 5.1. The three major domains of the modified assessment triangle framework: Cognition and Learning Targets (CLT), the Assessment Strategies and Tools (AST), and Evidence and Data Interpretation (EDI).

In the first topic domain, there is the Understanding Cognition and Learning Targets (CLT) map, which focuses on the types and quality of the construct map representations the classroom assessor uses to define an assessment target. The second topic domain is the Understanding the Assessment Strategies and Tools (AST) map. This variable focuses on the classroom assessor's knowledge of traditional item formats and uses, in addition to the general rules for constructing "good" items. The third topic domain is the Understanding Evidence and Data Interpretation (EDI) map; it includes the classroom assessor's knowledge and use of the properties of scoring and evaluation strategies, which depend on purpose and use (e.g., grading, feedback, reporting). At the highest levels on each map, the classroom assessor is expected to employ ideas related to validity, reliability, and standardization to evaluate the issues and problems related to, e.g., identification of cognitive learning targets, choice of item types to elicit a range of student skills and abilities, use of different scoring strategies to evaluate patterns of student progress, and so forth.

Initially, Duckor et al. (2013) employed the unidimensional construct modeling approach to evaluate the psychometric properties of the classroom assessment literacy (CAL) variable. The researchers' primary goal was to construct a measure of pre-service student teachers (in terms of latent "proficiency") and calibrate items (in terms of task "difficulty") on a technically sound scale. Towards this end, they examined evidence for validity and reliability scores derived from the classroom assessment literacy instrument. Guided by nationally

recognized standards (AERA, APA, & NCME, 1999) for instrument validation, the internal structure of the scale demonstrates acceptable fit according to a partial credit item response model. Evidence for relations to other, external variables (e.g., PACT, 2007) was strong. The CAL instrument's reliability was high (.94). The researchers also reported that model fit differences between constructed responses and fixed choice item formats provide insight into new directions for modeling the CAL variable.

The CAL scale was developed and piloted in order to evaluate the pre-service teachers' proficiency with understanding classroom assessment principles and practices. In accordance with the initial research design, it is assumed that responses to items can be differentiated into three different dimensions. That is, respondents (student teachers) should employ different levels of proficiency with CLT, AST, and EDI constructs. In this case, a calibration and interpretation of the item response data using a multidimensional IRT model would appear to be a straightforward solution in order to match the internal structure of the instrument. However, for the purposes of formative evaluation of respondents in the classroom context, the analyses generated by traditional multidimensional models are typically not at the right grain size to aid the end-user (in this case, teacher educators). In order to decide whether the student teacher has obtained a sufficient degree of knowledge to pass a course, for example, it would be necessary to have a single ability estimate across all three dimensions. Further, if the instrument were included in a state licensure context it is necessary for decision makers to obtain results that are readily interpretable, for example, in order to decide whether the general level of these proficiencies is sufficient to warrant provisional licensure or if additional resources and support (e.g., professional development) are required to improve these proficiencies across a larger population of teachers.

Following the described multidimensional modeling approach using the GSM, this article therefore explores the technical properties of a pilot classroom assessment literacy (CAL) scale for unidimensional ability estimates based on multidimensional latent variables.

5.2 Method

5.2.1 Data

The Classroom Assessment Literacy instrument is a pre- and post-test designed to measure teachers' understanding and use of the modified version of the National Research Council's "Assessment triangle" framework with particular focus on the three topic domains "Cognition and Learning Targets", "Assessment Strategies and Tools", and "Evidence and

Data Interpretation" (Pellegrino et al., 2001). The test consists of 55 items: 13 constructed response and 42 fixed choice questions. We analyzed 13 constructed response items from the CAL instrument, which were all coded as partial credit item with three different score categories each, ranging from 0 to 2. There are three items on the CLT sub-scale, four items on the AST sub-scale, and six items on the EDI sub-scale.

A sample of 72 respondents consisting of pre-service teachers who participated in a post baccalaureate course, titled "EDSC 182: Classroom Assessment and Evaluation" was obtained for this study. The 182 course was taught at a large California State University by the second author with concurrently Phase II/III student teaching field placements in diverse middle and high school classrooms. Respondents in the 182 course completed four course exhibitions, including the pre- and post-test described above. The data used in this study is taken from the post-test.

5.2.2 Model Definition

The applied partial credit extension of the generalized subdimension model (Brandt, 2012) is given by

$$\log\left(\frac{p_{nij}}{p_{ni(j-1)}}\right) = d_{k(i)}\left(\theta_n + \gamma_{nk(i)}\right) - b_{ij} \quad , \tag{5.1}$$

where p_{nij} is the probability of person n to give an answer corresponding to answer category j of item i; p_{ni0} the corresponding probability of giving an answer matching category (j-1); b_{ij} is the difficulty of step j of item i; θ_n is person n's ability on the constructed unidimensional dimension (denoted as main dimension); $\gamma_{nk(i)}$ is the person's subtest specific ability for (sub-) dimension k (with item i referring to dimension k) relative to the ability on the main dimension; and $d_{k(i)}$ is the translation parameter that translates the different multidimensional (or subdimensional) scales to a common one. Corresponding to hierarchical models, it is assumed that each item loads on exactly one subdimension. In order to identify the model several restrictions on the parameters have to be applied. First, the mean of the ability estimates θ and γ_k have to be constrained to zero, and the correlations between the main dimension and the K subdimensions have to be set to zero. Further, for each person the sum of the subtest specific parameters has to be constrained to zero ($\sum_k \gamma_{nk} = 0$), and the square

of the parameters d_k are constrained to the sum of K with each d_k additionally constrained to be positive ($\sum_k d_k^2 = K$).

The latter two constraints result from the characteristics of a mean score, and it can be shown that the given definition results in the main ability estimate to be the (equally weighted) mean of the specific abilities (Brandt, 2012).

5.2.3 Estimation

The estimation of the unidimensional partial credit model (Masters, 1982) and the generalized subdimension model was conducted following a Bayesian approach (Gelman, Carlin, Stern, & Rubin, 2003) using the computer program WinBUGS 1.4 (Lunn, Thomas, Best, & Spiegelhalter, 2000). In the Bayesian approach, prior distributions are assigned to the model parameters, and these along with the model definition and the observed data are used to produce a joint posterior distribution for the parameters. WinBUGS uses Markov Chain Monte Carlo techniques based on the Metropolis-within-Gibbs algorithm, a modified Metropolis-Hastings algorithm (Chib & Greenber, 1995), in order to simulate the joint posterior distribution.

For the presented analyses each item parameter is estimated based on a normal prior with mean 0 and variance 0.0001. The used priors for the variance estimation of the person parameters base on uniform and inverse gamma distributions. More precisely, the estimated person parameter variance in the unidimensional model and the variance of the main dimension in the generalized subdimension model are estimated using priors with uniform distributions from 0 to 100, and the variances and covariances of the subdimensions in the generalized subdimension model are estimated using an inverse-Wishart prior. The used hyperparameters for the inverse-Wishart prior are the identity matrix and the number of dimensions as degrees of freedom.

Further, both models are estimated using five Markov chains with different initial values. A total number of 11,000 iterations is calculated for each estimation with the first 1,000 iterations used as burn-ins. Every tenth iteration the simulated draws are saved, resulting in 1000 saved simulation draws for the calculation of the estimated parameters. The convergence of the chains was checked using the potential scale reduction factor (Brooks & Gelman, 1998; Gelman & Rubin, 1992).

5.3 Results and Discussion

All calibrations converged well and the potential scale reduction factor for all variables is close to one[1]. The calibrations of the generalized subdimension model and of the unidimensional model result in deviances of 1,324 and 1,376, respectively; that is, the unidimensional model yields a lower likelihood, and a multidimensional calibration is supported. The latent correlations, which range from .74 to .82, and the variances, which range from 1.08 to 2.63, (cf. Table 5.1) as well suggest the measurement of a heterogeneous construct including multiple dimensions.[2] A further argument for the heterogeneity of the data yields the comparison of the item parameter estimates from the unidimensional model and from the GSM (which are equivalent to those of the multidimensional model). Figure 5.2 shows that the variance of the item parameters for the dimension Cognition and Learning Targets is clearly reduced when estimated within the unidimensional model, whereas the estimates of the other two dimensions are more closely related for the unidimensional and GSM estimations.

Table 5.1

Multidimensional Estimation Results

Dimension	Variances and Correlations		
	CLT	AST	EDI
CLT	2.63	.74	.82
AST		1.28	.79
EDI			1.08

Note. Entries on the diagonal represent variances; entries above the diagonal represent correlations.

[1] For all variables the scale reduction factors' differences to one were below 0.002.

[2] In the above mentioned large scale assessments even such different domains such as reading and science typically show a higher correlation (>.9) and more similar variances than the here observed results (cf. OECD, 2009).

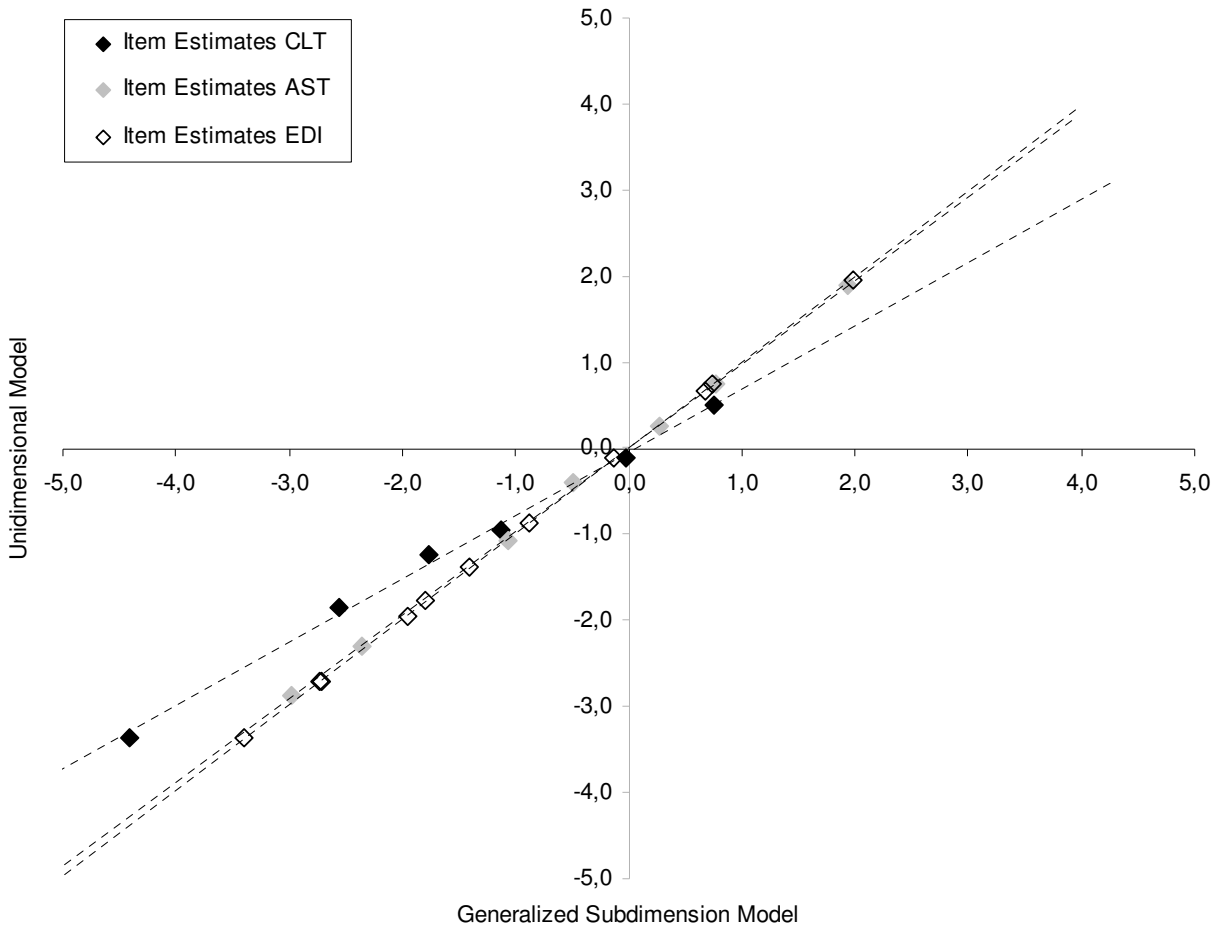

Figure 5.2. Comparison of the item estimates for the CLT, AST, and EDI dimension using a unidimensional calibration and a GSM calibration.

The calibration of the unidimensional model results in a variance of 1.01, and the corresponding (main dimension) variance in the generalized subdimension model is equal to 1.39. In order to compare the precision for the unidimensional ability estimates, the Expected a Posteriori (EAP) Estimates and their posterior standard deviations, which serve as standard errors, are depicted in Figure 5.3. It demonstrates that the GSM yields smaller standard errors for the ability estimates than the unidimensional model. The GSM yields a mean standard error in standard deviation of 51.8% for the unidimensional ability estimates while the unidimensional model yields 53.6%[1]. The resulting difference of 1.8% corresponds to an increase in measurement precision by 3.4%.

[1] In comparison to the standard deviation, the standard errors might seem high. However, in a large scale sample that includes a variety of different universities and programs, the achievement of the student teachers are assumed to vary to a larger extent, which will result in a larger standard deviation and therefore in smaller standard errors in comparison to the standard deviation.

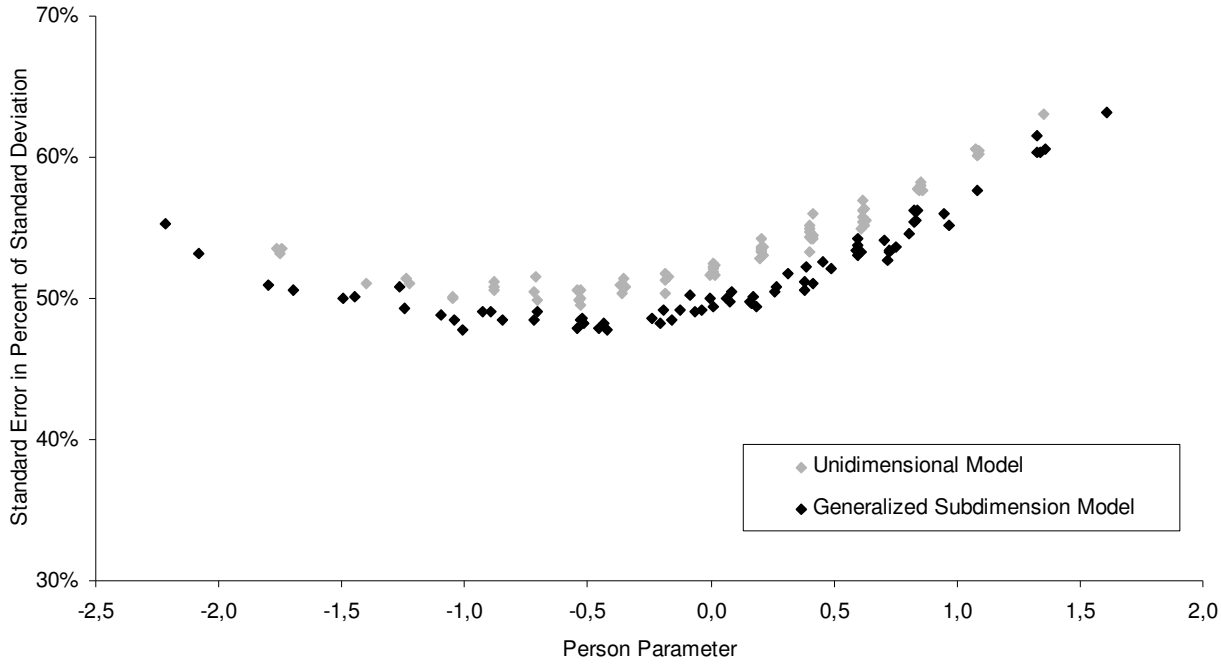

Figure 5.3. Comparison of the standard errors of the unidimensional person parameter estimates from the unidimensional model and from the generalized subdimension model and from the composed mean score of the multidimensional person parameter estimates.

A further characteristic of the generalized subdimension model is that it explicitly defines the subdimensions to be of equal weight[1]. In the unidimensional model the weighting of the subdimensions is implicit and is based on the total score that can be achieved within each subdimension. The total score depends on the number of items and on the number of scoring categories of each item. For the given data set the unidimensional model, therefore, results in a weighting of 23% (CLT), 31% (AST), and 46% (EDI) for the respective subdimensions (cf., Table 5.2). In the above given definition of the GSM, on the other hand, the subdimensions are of equal weight.

[1] Brandt (2012) also describes the extension of the model by a weighting parameter, which is not considered here.

Table 5.2

Weights of the Subdimensions

Dimension	Items	Score	Weight Unidimensional Model	Weight GSM
CLT	3	6	.23	.33
AST	4	8	.31	.33
EDI	6	12	.46	.33

If students show varying strengths and weaknesses in the subdimensions, their individual total score clearly depends on the applied weighting. Table 5.3 demonstrates the resulting differences by comparing the achievement of two single students included in the data set. The shown IRT ability estimates were standardized with a mean of 500 and a standard deviation of 100 (a commonly used scale, e.g., in the PISA study (OECD, 2009)). While according to the unidimensional model the first student outperforms the second student by 26 points (i.e., 26% of a standard deviation), according to the generalized subdimension model the second student outperforms the first by 8 points. The students' differences in the sum scores for the single subdimensions explain the origin of these contradictory results. Since the first student has a strength in the subdimension EDI, which has a high weight in the unidimensional model, and a weakness in CLT, which has a corresponding low weight, this student benefits from a calibration using the unidimensional model; while the contrary is true for the second student with a strength in CLT and a weakness in EDI. There are no *a priori* grounds for accepting one interpretation over the other. The stakeholder must decide whether the results according to the unidimensional model or the GSM are more appropriate and useful in making a decision about student progress and/or achievement.

Table 5.3

Comparison of Two Students

Student	Score CLT	Score AST	Score EDI	Total Score	Ability Unidimensional Model	Ability Generalized Subdimension Model
A	3	7	12	22	598	575
B	6	6	9	21	572	582

5.4 Conclusion

The results demonstrate that the multidimensional approach using the GSM allows the definition of an overall unidimensional ability estimates with increased measurement precision. In this case, the gain in precision (6.7%) was smaller than for the large-scale data set reported by Brandt (2012). Additionally, however, the further empirical analyses presented underscore the importance of utilizing an explicit weighting when approaching the problem of arriving at a "substantively meaningful" and statistically well-defined solution.

As Ackerman (1992) pointed out two decades ago: "because ordering is a unidimensional concept, researchers cannot order examinees on two or more abilities at the same time, unless they base their ranking on, for example, the weighted sum of each skill being measured" (see also Briggs & Wilson, 2003). The implicit weighting of the unidimensional model, however, is not transparent at first sight and may lead to invalid inferences about person proficiency or ability estimates. Additionally, the unidimensional model does not allow for a change in the implicit weighting, unless the number of items or scoring categories in an item is changed, which adds complexity to the test design and arguably less parsimony. The GSM, on the other hand, allows for an explicit weighting of the subdimensions and, thereby, makes the weighting transparent to stakeholders. Further, for policy makers interested in measuring trends with constructs weighted equally over time, it may also reduce the complexity of the "at scale" test design to invite more parsimonious interpretation of results.

A further characteristic of the generalized subdimension model in comparison to the unidimensional model is that it directly provides estimates for individual strengths and weaknesses in the different domains (by the gamma parameters). Although not directly addressed in this analysis, an additional benefit of the GSM approach is that it can provide estimates in educational contexts envisioned by the developers of the CAL instrument. The GSM approach allows the university instructor to differentiate teacher candidates (in this case, pre-service students) not only on a linear scale but also according to different types of proficiency profiles. These profiles might detect weakness in a topic area such as Cognition and Learning Targets (CLT): diagnostically, the instructor may want to review instruction related to defining and representing student thinking with concept maps or taxonomies; formatively, the instructor might reinforce instruction activities with timely, specific, addressable feedback on assignments and activities in the CLT unit; summatively, the instructor is likely most interested in the single scale score and may simply wish to obtain a precise measure before issuing a grade. An innovation of the GSM is that it integrates both

the formative and the summative information in a coherent, theoretically sound modelling approach.

From the instructors' perspective, educational interventions leading to decisions such as re-teaching the unit or redesigning a lesson or deploying more feedback should be guided by reliable score information. The multidimensional approach, using the GSM, provides a way for making better decisions about individual learners' needs and performance, for different stakeholders and contexts. We offer a modeling strategy with explicit weightings that directly addresses the tension between the non-trivial task of finding the smallest "number of latent ability domains such that they are both statistically well-defined and substantively meaningful."

5.5 References

Ackerman, T. A. (1992). A Didactic Explanation of Item Bias, Item Impact, and Item Validity From a Multidimensional Perspective. *Journal of Educational Measurement*, *29*(1), 67–91.

AERA, APA, & NCME. (1999). *The Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. & Novick, M.R. (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison–Wesley.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Brandt, S. (2012). Definition and classification of a generalized subdimension model. *2012 annual conference of the National Council on Measurement in Education (NCME)*. Vancouver, BC.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of applied measurement*, *4*(1), 87–100.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Chib, S., & Greenber, E. (1995). Understanding the Metroplis-Hastings algorithm. *Statistician*, *49*(4), 327–335.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*(8), 620–639.

Duckor, B., Draney, K., & Wilson, M. (2009). Measuring measuring: toward a theory of proficiency with the constructing measures framework. *Journal of applied measurement*, *10*(3), 296–319.

Duckor, B., Draney, K., & Wilson, M. (2013). Assessing assessment literacy: An item response approach. Presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Duckor, B. M. (2006). *Measuring Measuring: An Item Response Theory Approach*. University of California, Berkeley.

Gelman, D., Carlin, J., Stern, H., & Rubin, D. (2003). *Bayesian Data Analysis*. Boca Raton: CRC Press.

Gelman, D., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, 2, 41-54.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3–21.

Lunn, D. J., Thomas, A., Best, N. G., & Spiegelhalter, D. J. (2000). WinBugs - A Bayesian modelling framework: Concepts, structure, and extensibilty. *Statistics and Computing*, *10*(4), 325–337.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2007). *Progress in International Reading Literacy Study (PIRLS): PIRLS 2006 Technical Report*. TIMSS & PIRLS International Study Center. Boston College, Chestnut Hill, MA.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.

National Research Council. (2000). *Tests and teaching quality: Interim report*. Washington, D.C.: National Academies Press.

OECD. (2009). *PISA 2006 technical report*. Paris: OECD.

OECD. (2013). *PISA 2012 Assessment and Analytical Framework*. Paris: OECD.

Olsen, J. F., Martin, M. O., Mullis, I. V. S., Martin, M. O., & Mullis, I. V. S. (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: Boston CollegeOlsen, J.F.

PACT. (2007). A Brief Overview of the PACT Assessment System. Retrieved November 16, 2012, from http://www.pacttpa.org/_files/Main/Brief_Overview_of_PACT.doc

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. National Academies Press.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*(3), 361–372.

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, *68*(3), 413–430.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Mesa Press.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for 2 latent trait models. *Journal of Educational Measurement*, *17*(4), 297–311.

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *64*, *2*(113-128).

6 Discussion

The problem of calculating unidimensional scale scores for data assumed to be multidimensional has received substantial attention and had led to the formulation of a growing variety of IRT models to cope with this issue. In this chapter, the different groups of existing IRT models will therefore, at first, shortly be characterized before the special characteristics of the GSM and its relationship to these models are considered in more detail.

6.1 The Testlet, the Higher Order, and the Hierarchical Model

Depending on whether an assumed LID originates in the test construction or in the psychological construct that is to be measured, the models are typically denoted as testlet models, or as hierarchical or higher-order models, respectively.

Testlet models (Bradlow, Wainer, & Wang, 1999; Wang & Wilson, 2005) assume that the answers on a test depend on a single psychological construct. Additionally though, they assume that due to a testlet based test construction the test includes multidimensionality that corresponds to each person's familiarity with the stimulus given for a testlet. From the perspective of the unidimensional latent trait that is to be measured this multidimensionality corresponds to local item dependence (LID).

Hierarchical or higher-order models (de la Torre & Song, 2009; Gibbons & Hedeker, 1992; Sheng & Wikle, 2008), on the other hand, typically assume that the answers in a test depend on multiple psychological constructs that are related by a common underlying construct. That is, from a unidimensional perspective, the test includes local item dependencies not due to the test construction but due to the nature of the psychological construct.

While the reasons for the necessity to model multidimensionality (or LID) are different for the two mentioned groups of models, the statistical assumptions made are very similar. Yung, Thissen, and McLeod (1999) and Li, Bolt, and Fu (2006) have shown that both the higher-order model and the testlet model are restrictions of the hierarchical model. Furthermore, Rijman (2010) has shown that the testlet model is formally equivalent to a second-order model (i.e., a higher-order model with a second order as the highest order). A general assumption of hierarchical models (i.e., as well of testlet and of higher-order models) is that existing correlations between the subtest (or testlet) factors fully originate in the underlying common latent trait they measure (Holzinger & Swineford, 1937). That is, constrained on the common latent trait, often denoted as g-factor and in the GSM denoted as main dimension, the subdimension factors are independent (see Rijmen, 2010; Yung et al.,

1999). To what extent this theoretical requirement of hierarchical models and the resulting constraints applied for the estimation lead to a significantly worse model fit in comparison to the common multidimensional model depends on the given multidimensional construct that is measured (or on the stimuli used for the testlets). In chapter 3, it is shown that for the TIMSS 2003 mathematics achievement test, for example, the difference in the deviance of the testlet model and the multidimensional model is about 50% of the difference in deviance between the multidimensional model and the unidimensional model. That is, the testlet model is only partially able to model the multidimensionality in the given data set. The ability of testlet models, or more general of hierarchical models, to model multidimensionality will be different for each data set depending on the given covariance structure of the subdimensions. Due to the restrictive assumptions, however, it is very unlikely that they will be able to fully model the multidimensionality, and the resulting model fit will therefore be significantly worse, as in the analysis using the TIMMS data.

6.2 Estimated Parameters

A basic characteristic of all IRT models is the number of parameters they use. The more parameters a model comprises, the better it will typically be able to model a given data set. A higher number of parameters, however, usually means a more complicated interpretation, whereas models with fewer parameters often provide clearer interpretations. On the other hand, fewer parameters typically correspond to stronger assumptions, that is, more constraints for the calibration of the model. Comparing the number and types of estimated parameters is therefore a useful way to discuss the characteristics of different models.

The GSM's constrains the means of the person abilities to zero, which is a standard constraint in IRT models and is necessary to fix the scales on the latent continuum. Furthermore, it is a characteristic of mean scores calculated from distributions (here dimensions) with equal variances to yield a covariance of zero between the mean scores and the distributions of the difference values, that is the distribution values minus the respective mean scores. Hence, constraining the covariance of the main dimension and the subdimensions to zero does not constrain the estimation of the mean score. This constraint is still in accordance with the hierarchical models, the remaining constraints of the GSM are different, though. They are necessary in order to allow for correlations between the subtest specific parameters, which are constrained to be independent in hierarchical models. The differences between these assumptions of the GSM and the hierarchical model are depicted in Figure 6.1. Here, $\mathbf{y}_k$ denotes the response vector pertaining to subtest k, θ the

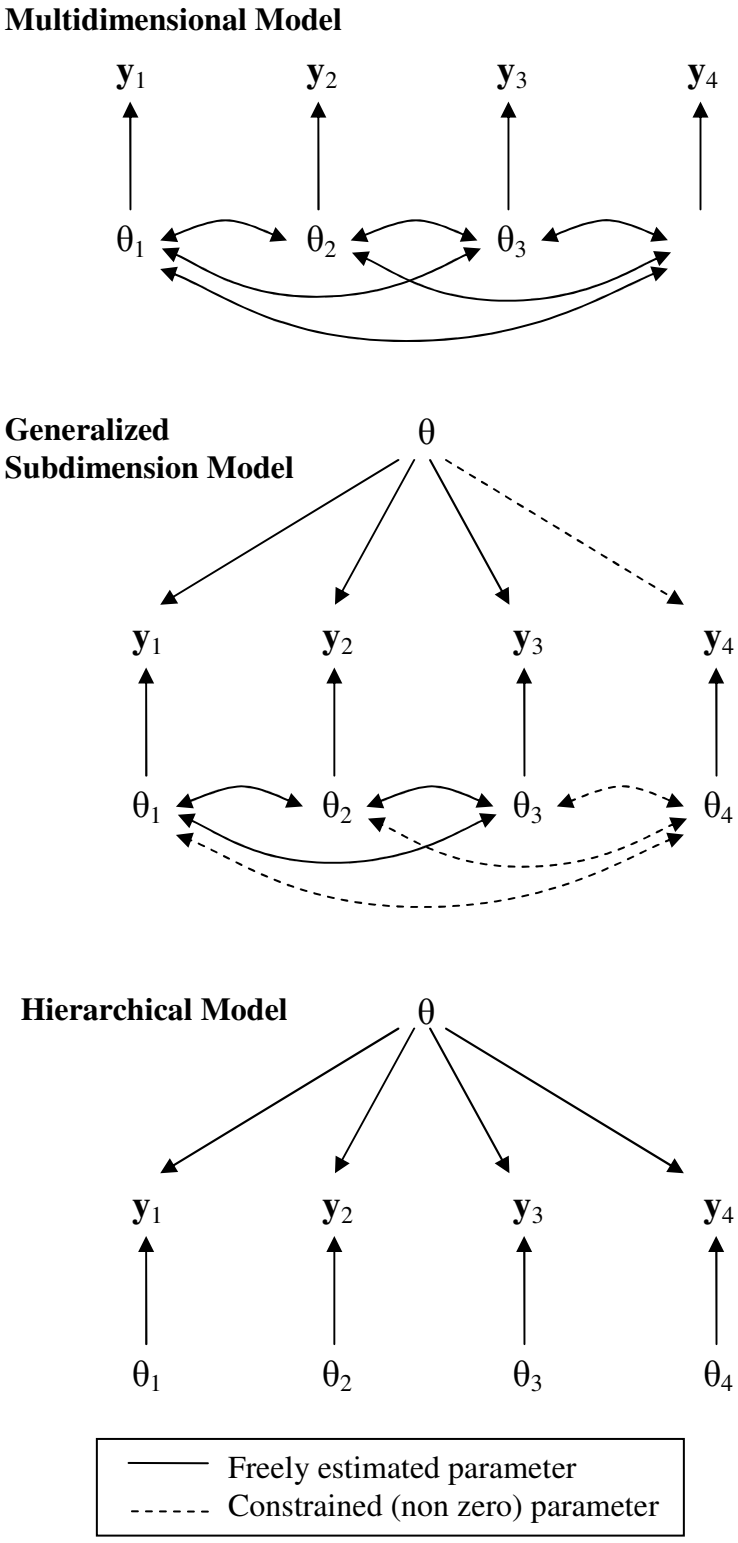

Figure 6.1. Graphical representations of the estimated parameters of the multidimensional model, the generalized subdimension model, and the hierarchical model.

latent variable of the main dimension, and θ_k the latent variable of subdimension *k*. Additionally, the characteristics of the two models are contrasted to the multidimensional model that is depicted as well, and it can be shown that the number of (freely) estimated parameters in the GSM equals that of the multidimensional model.

The multidimensional model (cf. Rost, 1996) is defined as

$$\log\left(\frac{p_{ni1}}{p_{ni0}}\right) = \mathbf{a}_i\left(\boldsymbol{\theta}_n - b_i\mathbf{1}\right), \tag{6.1}$$

where $\boldsymbol{\theta}_n = \left(\theta_{n1},...,\theta_{nK}\right)^{\mathrm{T}}$ is the ability vector of person *n* for his/her abilities in the *K* dimensions; $\mathbf{a}_i = \left(a_1,...,a_K\right)^{\mathrm{T}}$ is a vector with values of only 0 and 1, indicating whether an item loads on a dimension or not; $\mathbf{1}$ is the unity vector with *K* elements; and the remaining variables are defined as above. Without loss of generality, we assume that the variances in the multidimensional model are constrained to a mean of zero in order to identify the model. That is, *K* parameters for the estimation of the dimensions' variances have to be estimated. In the generalized subdimension model as well *K* variances have to be estimated: *K*-1 subdimension specific variances (one subdimension specific variance, typically that of subdimension *K,* is not estimated since its parameters result via the constraint $\sum_k \gamma_{nk} = 0$) and the variance for the main dimension (cf. Equation 5.1 in the previous chapter). That is, as well *K* parameters for the dimensions' variances have to be estimated.

Considering the covariance estimates, these are all estimated freely in the multidimensional model, resulting in $\sum_{k=1}^{K}(k-1)$ covariance parameters to be estimated. In the GSM the covariance matrix includes the main dimension, of which the covariances are constrained to zero, and *K*-1 subdimensions (one subdimension results from constrained parameters; cf. above) with freely estimated covariances; that is, here $\sum_{k=1}^{K-1}(k-1)$ covariance parameters have to be estimated, which are *K*-1 covariance parameters less than in the multidimensional model. However, there are exactly *K*-1 additional parameters to be estimated for the translation parameters d_k (cf. definition in the previous chapter). Since the number of estimated parameters for the items' difficulties are equal as well, the generalized subdimension model and the multidimensional model therefore comprise an equal number of parameters.

6.3 Equivalence With the Multidimensional Model

The equivalence of the GSM and the multidimensional model is provided via the definition of $\theta_{nk} = d_k\left(\theta_n + \gamma_{nk}\right)$, where θ_{nk} equals the ability estimate for the corresponding (sub-)dimension k in the multidimensional model. In order to show the equivalence, it is demonstrated in the following that the estimation of the means, variances, and covariances for the distributions of the parameters θ_{nk} are equal to those of the multidimensional model.

Without loss of generality, it is again assumed that both models are estimated using constraints on the cases. That is, the distributions of the parameters θ_n and γ_{nk} are constrained to a mean of zero. Hence, the K distributions of the θ_{nk} are trivially constrained to a mean of zero as well, and their means correspond to those in the multidimensional model.

The independence of the variance estimation for the subdimensions within the GSM is not as straightforward. By constraining the subdimension specific parameters to a sum of zero for each person (Constraint I), the estimation of the ability parameters for the different subdimensions is dependent on each other; because of this, the standard subdimension model includes an implicit variance restriction for the estimation of the subdimensions by applying this constraint. In order to neutralize this implicit variance constraint, the additional introduction of the translation parameters d_k is necessary. They yield that each variance for the K subdimensions is estimated independently. However, since the main dimension variance is also estimated (resulting in a total number of K+1 estimated variances) the variance parameters need a further constraint in order to be identified, which is yielded by constraining the square of the parameters d_k to the sum of K (Constraint II).

The correspondence of the covariances in the GSM and the multidimensional model is shown by Equation 6.2. For any two distributions of θ_{n1} and θ_{n2}, it is true that

$$\begin{aligned}
\mathrm{Cov}\left(\theta_{n1},\theta_{n2}\right) &= \mathrm{Cov}\left(d_1\left(\theta_n+\gamma_{n1}\right), d_2\left(\theta_n+\gamma_{n2}\right)\right) \\
&= d_1 d_2 \mathrm{Cov}\left(\theta_n+\gamma_{n1},\theta_n+\gamma_{n2}\right) \\
&= d_1 d_2 \left(\mathrm{Cov}\left(\theta_n,\theta_n\right)+\mathrm{Cov}\left(\theta_n,\gamma_{n2}\right)+\mathrm{Cov}\left(\gamma_{n1},\theta_n\right)+\mathrm{Cov}\left(\gamma_{n1},\gamma_{n2}\right)\right) \\
&= d_1 d_2 \left(\mathrm{Var}\left(\theta_n\right)+\mathrm{Cov}\left(\gamma_{n1},\gamma_{n2}\right)\right) \\
&\qquad\qquad \left(\text{since } \mathrm{Cov}\left(\gamma_{n1},\theta_n\right)=\mathrm{Cov}\left(\gamma_{n2},\theta_n\right)=0\right)
\end{aligned} \tag{6.2}$$

That is, the existing covariance structure between the dimensions in the multidimensional model can be fully recovered by the generalized subdimension model even though the underlying parameters θ_{nk} are split into the parameters γ_{nk}, θ_n, and d_k, and only the covariance structure of the parameters γ_{nk} is estimated.

6.4 Correspondence With the Unidimensional Model

If the assumed multidimensional structure does not exist but the subdimensions in fact measure the same construct, the variances of the subdimension specific parameters γ_{nk} reduce to zero, and, hence, it is true that all subdimension variances are equal (namely zero). In order to yield Constraint II of the GSM, the parameters d_k then are all equal to 1, and the variance of the main dimension equals the variance of the ability estimates in the unidimensional model. Due to the definition of Constraint II, it is generally yielded that the variance of the main dimension is the mean of the unidimensional variance components of the subdimensions, that is, the mean of the total subdimension variances less the mean of the subdimension specific variances (see Equation 6.3).

$$
\begin{aligned}
\mathrm{Var}\left(\theta_n\right) &= \mathrm{Var}\left(\theta_n\right)\frac{\sum_k d_k^2}{K} \qquad \text{(due to Constraint II)} \\
&= \frac{\sum_k d_k^2 \mathrm{Var}\left(\theta_n\right)}{K} + \frac{\sum_k \mathrm{Var}_k\left(d_k\gamma_{nk}\right)}{K} - \frac{\sum_k \mathrm{Var}_k\left(d_k\gamma_{nk}\right)}{K} \\
&= \frac{\sum_k \mathrm{Var}_k\left(d_k\left(\theta_n+\gamma_{nk}\right)\right)}{K} - \frac{\sum_k \mathrm{Var}_k\left(d_k\gamma_{nk}\right)}{K} \\
&= \frac{\sum_k \mathrm{Var}_k\left(\theta_{nk}\right)}{K} - \frac{\sum_k \mathrm{Var}_k\left(d_k\gamma_{nk}\right)}{K} \\
&= M\left(\mathrm{Var}_k\left(\theta_{nk}\right)\right) - M\left(\mathrm{Var}_k\left(d_k\gamma_{nk}\right)\right)
\end{aligned}
\tag{6.3}
$$

Furthermore, as shown by Equation 6.4, each person's ability in the main dimension is the mean of his/her abilities in the subdimensions considered on the common scale they are translated to. As noted before, this translation is yielded by the parameters d_k. A multiplication using the reciprocal of d_k therefore translates the subdimension ability

estimates onto the scale in which their variances are of equal extent. That is, while θ_{nk} is equivalent to the multidimensional ability estimate, $\frac{1}{d_k}\theta_{nk}$ is the corresponding ability estimate translated to the scale where the variances are of equal extents.

$$
\begin{aligned}
\theta_n &= \theta_n + \frac{\sum_k \gamma_{nk}}{K} && \text{(due to Constraint I)} \\
&= \frac{K \cdot \theta_n}{K} + \frac{\sum_k \gamma_{nk}}{K} \\
&= \frac{\sum_k (\theta_n + \gamma_{nk})}{K} \\
&= \mathrm{M}\left((\theta_{n1} + \gamma_{n1}), ..., (\theta_n + \gamma_{nK})\right) \\
&= \mathrm{M}\left(\frac{1}{d_1}\theta_{n1}, ..., \frac{1}{d_K}\theta_{nK}\right)
\end{aligned}
\tag{6.4}
$$

6.5 Relationship to the Hierarchical Model

The restriction of the translation parameters d_k via Constraint II in the GSM makes clear that these cannot be interpreted as regression coefficients. This is in contrast to the hierarchical models, in which the corresponding parameters are equivalent to the loadings or regression coefficients of the subdimensions on the main dimension (cf. de la Torre & Song, 2009). However, the estimation of these coefficients in the hierarchical models relies on the assumption that the subtest specific abilities are independent and will therefore correspond to the correlation of the subtest ability and the overall test ability only if this independence assumption holds. The GSM on the other hand allows for a correlation of the subdimension specific abilities. Following the definition of Holzinger and Swineford (1937), the GSM, thereby, corresponds to a modified hierarchical model, which denotes a hierarchical model with overlapping specific factors.

Besides the differences in the assumed covariance structure, the difference between the hierarchical model and the GSM is also depicted by the different levels on which the constraints of the models are applied. While the main constraint of the hierarchical model

yields a characteristic on the level of the test, or the latent trait (the independence of the distributions for the specific factors), the main constraint of the GSM yields a characteristic on the level of the individual person, namely, that the sum of the (translated) specific ability estimates for each person is zero.

6.6 Conclusion

Considering the multidimensional model, it has been shown above that the GSM yields equivalent parameter estimates and that the unidimensional parameter estimates are constructed as means of the translated multidimensional parameters. The application of the GSM thereby yields important advantages for the calculation of unidimensional achievement scores for multidimensional tests. First, the resulting unidimensional estimate is free of any negative impact of LID due to the subdimensions; second, the estimation is conducted within the common framework of IRT, allowing the calculation of reliable individual estimates for the comprehensive scores; and third, being defined as a mean score the interpretation of the estimate is clear and transparent, which is particularly important in high-stakes testing.

A further advantage of the GSM bases on its characteristic to directly yield standardized estimates and posterior distributions for the difference scores, that is, the differences in the achievements in the subdimensions. Often researchers are not interested in the absolute abilities in the subdimensions but in the differences between these. This might be in order to investigate if students differ in their subdimension abilities, to differentiate types of students via their ability profiles, or to consider trends in longitudinal studies (where the subdimensions represent measurements at different points in time). Brandt, Duckor, and Wilson (2014), for example, presented an approach based on the GSM's difference scores in order to investigate the dimensionality of a given test.

6.7 Prospect of Current and Future Research

The given definition of the GSM relates to the Rasch Model (Rasch, 1980). However, its extension to a corresponding 2-PL model (Birnbaum, 1968) is straightforward, and an application of a 2-PL variant of the GSM to NAEP data in order to compare the results of the unidimensional scores from the GSM with those based on the mean scores from the multidimensional plausible value estimates will be of great interest. Hitherto, the analysis of large-scale data using the GSM was difficult, though, since a calibration of the model was only possible using WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000; cf. chapter 5)[1].

[1] In contrast to the Subdimension Model, a calibration of the GSM using ConQuest is not possible.

Besides the complexity of correctly defining models, estimations in WinBUGS are even for small data samples extremely time consuming. Since a recent update in the R software package TAM (Kiefer, Robitzsch, & Wu, 2015), however, the GSM can be calibrated using TAM and can be estimated as efficient as any standard multidimensional IRT model. Furthermore, TAM also allows the calibration of the GSM including regression models. Future applications of the GSM to large scale assessment data are therefore now unproblematic.

Besides a broader application of the GSM in order to further explore the statistical differences of the unbiased and weighted unidimensional GSM estimates and the standard unidimensional IRT estimates, it is hoped that the GSM might also add to the discussion and determination of the dimensionality of data sets. The opportunity of selecting an IRT model that provides a unidimensional score but does not assume unidimensionality might help test developers in accepting more easily given multidimensionality and allowing for the construction of more multidimensional tests. Furthermore, the reliable difference scores yielded by the GSM might help in guiding the discussion of dimensionality from a decision based on model fit towards a decision based on utility. Typically, the dimensionality of a test is defined via model fit comparisons, even though often enough the results are contradicting and depend on the chosen fit criterion. In a utility based approach a very clear question is asked: Is there a relevant number of persons that actually differ on the given dimensions, so a separate interpretation of the dimensions is useful? Being based on the utility, the approach thereby also yields a direct relation to the validity of the assumed dimensionality (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; cf. Brandt et al., 2014).

6.8 References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Brandt, S., Duckor, B., & Wilson, M. (2014). *A utility-based validation study for the dimensionality of the performance assessment for california teachers*. Presented at the 2014 annual conference of the American Educational Research Association (AERA), Philadelphia, PA.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*, 620–639.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*, 41–54.

Kiefer, T., Robitzsch, A., & Wu, M. (2015). TAM: test analysis modules (Version 1.3) [R]. Retrieved from http://cran.r-project.org/package=TAM

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*, 3–21.

Lunn, D. J., Thomas, A., Best, N. G., & Spiegelhalter, D. J. (2000). WinBugs - A Bayesian modelling framework: Concepts, structure, and extensibilty. *Statistics and Computing*, *10*, 325–337.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*, 361–372.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion [Textbook test theory, test construction]*. Bern; Göttingen; Toronto; Seattle: Verlag Hans Huber.

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, *68*, 413–430.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*, 126–149.

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model, *64*, 113–128.

Lebenslauf

Steffen Jan Brandt

Geburt: 21. April 1974, Hamburg, Deutschland
Nationalität: Deutsch

Werdegang

Februar 2016	Abschluss der Promotion im Fach Pädagogik an der Christian-Albrechts-Universität in Kiel
seit Sept. 2008	Selbständig als freier Mitarbeiter und Berater in der empirischen Bildungsforschung
Juni 2004 – Aug. 2007	Wissenschaftlicher Mitarbeiter am Institut für die Pädagogik der Naturwissenschaften (IPN) an der Christian-Albrechts-Universität in Kiel im Projekt PISA 2006; Beginn der Promotion
November 2003	Vertretungsstelle als Mathematiklehrer am Gymnasium Kirchdorf-Wilhelmsburg in Hamburg
April 2002 – Sept. 2003	Studium der Mathematik und Wirtschaft/Politik für das Lehramt an Gymnasien an der Christian-Albrechts-Universität in Kiel
Sept. 1999 – Dez. 2001	Zunächst Projektmitarbeiter, später Teilprojektleiter und dann Projektleiter bei einem Tochterunternehmen der Hamburgischen Landesbank
Sept. 2000	Abschluss der Diplom-Informatik mit Note „gut“
Sept. 1998 – Aug. 1999	Einjähriges Auslandsstudium „International Business“ an der Hawaii Pacific University in den USA
Juni 1997	Vordiplom in Volkswirtschaftslehre
Okt. 1996	Vordiplom in Informatik; Zweitstudium Volkswirtschaftslehre, ebenfalls an der Christian-Albrechts-Universität in Kiel
1994-2000	Studium der Informatik mit Nebenfach Betriebswirtschaftslehre an der Christian-Albrechts-Universität in Kiel
1993-94	Wehrdienst
1984-93	Allgemeine Hochschulreife, Gymnasium Altenholz