pedocs
Open Access Erziehungswissenschaften

DIPF

Wilbert, Jürgen; Linnemann, Markus

Kriterien zur Analyse eines Tests zur Lernverlaufsdiagnostik

Empirische Sonderpädagogik 3 (2011) 3, S. 225-242

Empfohlene Zitierung/ Suggested Citation:
Wilbert, Jürgen; Linnemann, Markus: Kriterien zur Analyse eines Tests zur Lernverlaufsdiagnostik - In: Empirische Sonderpädagogik 3 (2011) 3, S. 225-242 - URN: urn:nbn:de:0111-opus-93256
http://nbn-resolving.de/urn:nbn:de:0111-opus-93256

Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

We grant a non-exclusive, non-transferable, individual and limited right to using this document.
This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Mitglied der Leibniz-Gemeinschaft

Empirische Sonderpädagogik

ISSN 1869-4845 3. Jahrgang 2011 | Heft 3

Schwerpunktthema:
Trends in der sonderpädagogischen Diagnostik
Gastherausgeberin: Gabi Ricken

Pabst Science Publishers

Empirische Sonderpädagogik, 2011, Nr. 3, S. 225-242

Kriterien zur Analyse eines Tests zur Lernverlaufsdiagnostik

Jürgen Wilbert, Markus Linnemann

Universität zu Köln

Lernverlaufsdiagnostik spielt in pädagogischen Kontexten eine weit wichtigere Rolle als bloße Statusdiagnostik. Tests, mit denen die Veränderung eines Merkmals über einen bestimmten Zeitraum hinweg reliabel und valide gemessen werden kann, stellen hohe testtheoretische Anforderungen an das Material. Ziel des vorliegenden Beitrags ist es, Kriterien, die ein Test zur Lernverlaufsdiagnose erfüllen muss, vorzustellen und darüber hinaus eine Verfahrensweise der Testanalyse zu erarbeiten, an der diese Kriterien geprüft werden können. Im ersten Teil des Beitrages werden einige zentrale Voraussetzungen für ein solches Instrument diskutiert. Hierzu gehören die Reliabilität, Eindimensionalität, gleiche Schwierigkeiten der dargebotenen Materialien und hohe Testfairness. Im Anschluss daran wird eine Verfahrensweise vorgeschlagen, mittels deren diese Kriterien systematisch geprüft werden können. Die entwickelte Vorgehensweise wird am Beispiel eines neuen Instrumentes zur Erfassung des Sprachstandes von Schülerinnen und Schülern mit Beeinträchtigungen des Lernens demonstriert. Die auf der Klassischen Testtheorie sowie der Probabilistischen Testtheorie beruhenden Analysen zeigen, dass das Instrument (C-Test) eine hohe Reliabilität, Eindimensionalität und hohe Testfairness bezüglich des Geschlechts und der Muttersprache der Testteilnehmer besitzt. Außerdem kann mit Hilfe von erweiterten Raschmodellen aufgezeigt werden, dass sich die Einzelitems schwierigkeitskorrigiert zur Lernverlaufsdiagnostik einsetzen lassen.

Schlüsselwörter: Lernverlaufsdiagnostik, erweiterte Rasch-Modelle, DIF, Sprachstand, Lernbeeinträchtigung

Criteria for Analyzing a Test Measuring Learning Progress

Evaluating learning progress is a vital element of educational interventions for students with learning disabilities. Measuring change imposes considerably different requirements on test construction compared to traditional psychometric diagnostic instruments. The present paper discusses four theoretical challenges for test construction, namely a high reliability, unidimensionality, constant item difficulty, and high test fairness. A procedure is proposed for analyzing tests to fit these four criteria making use of item analyses, confirmatory factor analyses, Rasch modeling, and analyses of differential item functioning. The suggested procedure is exemplified on a newly developed test for measuring language proficiency of students with learning difficulties on the basis of c-tests. The results disclose c-tests as highly suitable for measuring differences in general language development.

Key words: learning progress, polytomous Rasch model, DIF, language proficiency, learning difficulties

Diagnostik ist im engeren Sinn die Messung eines Merkmales und die folgende Bewertung des Ergebnisses dieser Messung unter Verwendung eines spezifischen Vergleichsstandards. Die Funktionen, die einer solchen Bewertung zukommen, sind dabei vielfältig und reichen von der reinen Feststellung eines Zustandes über die Ableitung einer geeigneten Behandlungsmaßnahme bis zur Evaluation eines Veränderungsprozesses (vgl. Fisseni, 1997, S. 3f).

Letztere Funktion steht im Fokus des vorliegenden Beitrages. Es werden die Voraussetzungen diskutiert, die ein diagnostisches Instrument erfüllen muss, damit es dazu geeignet ist, die Veränderung eines Merkmals über einen bestimmten Zeitraum hinweg zu erfassen. Präziser soll es um die Entwicklung einer Verfahrensweise zur Testanalyse gehen, mit dem Instrumente auf ihre Tauglichkeit geprüft werden können, das im Kontext pädagogischer Fragestellungen besonders interessierende Merkmal des Verlaufs von Lernprozessen beim Erwerb unterschiedlicher Fertigkeiten zu erfassen. Die entwickelte Vorgehensweise wird am Beispiel eines neuen Instrumentes zur Erfassung des Sprachstandes von Schülerinnen und Schülern mit Beeinträchtigungen des Lernens demonstriert.

Anforderungen an einen Test zur Lernverlaufsdiagnostik

Die Erfassung individueller Lernverläufe hat in den sonderpädagogischen Modellen zur Förderung der letzten Jahrzehnte zunehmend an Bedeutung gewonnen (vgl. Klauer, 2006; Müller & Hartmann, 2009; Walter, 2010). Unter den Begriffen *Lernfortschrittsmessung, lernprozessbegleitende Diagnostik, Dynamic Testing, Response-to-Intervention* oder *Curriculum-basierte Messung* werden im Detail zwar unterschiedliche, aber eng verwandte Konzepte diskutiert. Ihnen ist gemein, dass sie individuell unterschiedliche Verläufe im Prozess des Lernens sichtbar machen wollen. Dieser gemeinsame Kern soll im Folgenden durch den Begriff *Lernverlaufsdiagnostik* referenziert werden.

Das zentrale Merkmal einer Lernverlaufsdiagnostik ist, dass hierbei der Prozess der Veränderung betrachtet wird. Lernen beschreibt die Veränderung des Verhaltenspotenzials einer Person (Miller & Grace, 2003, S. 358). Das Konstrukt eines immanenten Potenzials lässt sich nur indirekt über das Verhalten einer Person erschließen. Demnach lässt sich Lernen durch die Veränderung im Verhalten in einer gleichbleibenden Situation beschreiben. Um den Lernverlauf zu diagnostizieren, bedarf es daher einer wiederholten Messung innerhalb bestimmter Zeitintervalle unter Bedingungen gleichen Aufforderungscharakters. Dies schließt ein, dass zwischen den einzelnen Messungen bestimmte Interventionen (allgemeiner: Veränderungen) stattgefunden haben, die wiederum die Veränderung im Verhalten bedingen. Im Fokus eines solchen diagnostischen Prozesses steht also die Veränderung selbst, die einen Hinweis darauf geben soll, in welchem Maß eine Person von einer bestimmten Intervention profitiert und sich ihre Fähigkeit bzw. Expertise in einem bestimmten Aufgabenbereich erhöht. Sternberg und Grigorenko (2002) bezeichnen Verfahren, die dieser Grundidee folgen, als dynamische Testungen (Dynamic Testing).

Die besonderen Bedingungen einer Veränderungsmessung stellen korrespondierende Anforderungen an die Messinstrumente, die sich zu einer Lernverlaufsdiagnostik eignen. Dabei sind diese Anforderungen nicht grundsätzlich verschieden zu denen, die an Tests zur Statusdiagnostik gestellt werden, erfahren aber eine abweichende Akzentuierung, und es entstehen besondere Probleme bei deren Umsetzung. Vier zentrale testtheoretische Kriterien, die *Reliabilität*, die *Eindimensionalität*, die *Itemschwierigkeit* und die *Testfairness* werden im Folgenden vor diesem Hintergrund besprochen.

Reliabilität

Das Ergebnis einer Messung muss in hohem Maß auf die zugrunde liegende Fähigkeit verweisen und darf nur in geringem Maß durch einen Messfehler gestört werden. Die Merkmalsvarianz ist also deutlich größer als die Fehlervarianz, was der Klassischen Testtheorie (KTT) folgend einer hohen Reliabilität entspricht. Hier ergibt sich allerdings das Problem, dass das Messinstrument zugleich hoch reliabel und veränderungssensitiv sein muss. Die Konzeptualisierung der Messgenauigkeit als Retest-Reliabilität $\rho_{tt'}$ wie sie in der KTT vorgenommen wird, setzt eine gleichmäßige Abbildung der Veränderung von Messwerten auf Veränderungen des zugrunde liegenden Merkmals auf allen Niveaus der Merkmalsausprägung voraus. Nehmen wir z.B. an, zwei Schüler verbessern sich in einem Test zum Sprachstand auf einer 20-Punkt-Skala um 5 Punkte. Schüler 1 verbessert sich von 0 auf 5 Punkte und Schüler 2 verbessert sich von 15 auf 20 Punkte. Beide Schüler sollten sich nun um das gleiche Maß in ihrem Sprachstand verbessert haben. Damit dies stimmt, müssen einige kritische Eigenschaften des Instrumentes erfüllt sein: a) Homoskedastizität der Skala, d.h., der Messfehler ist in allen Skalenbereichen gleich (Cronbach & Furby, 1970) und b) die gemessene Skala hat mindestens Intervallskalenniveau, denn nur dann lassen sich Differenzwerte sinnvoll interpretieren. Letzteres bedeutet bei Skalen, die sich durch die Aufsummierung einzelner Itemwerte ergeben, dass diese Items einen homogenen, unkorrelierten Messfehler aufweisen und in gleichem Maß von der zugrunde liegenden Merkmalsausprägung τ beeinflusst sind. Im Rahmen der KTT werden diese Bedingungen als Voraussetzungen formuliert und nicht geprüft. Während bei Instrumenten, die zum Zweck der Statusdiagnostik entwickelt wurden, eine Verletzung dieser Annahmen weniger bedeutsam ist, führen diese jedoch bei Instrumenten, die zur Messung von Veränderungen konstruiert wurden, dazu, dass diese nicht sinnvoll nutzbar sind (Bereiter, 1963; Grigorenko & Sternberg, 1998). Anders verhält es sich mit dem Rasch Modell und seinen Erweiterungen. Einerseits wird hier die τ-Äquivalenz geprüft, indem die Trennschärfen der Items (die Steigungsparameter der Itemfunktionen) auf Homogenität getestet werden. Andererseits werden die einzelnen beobachteten Werte eines Items (Itemkategorien) in eine Logitskala transformiert, die mindestens Intervallskalenniveau aufweist. Letzteres bedeutet, dass im Rahmen der Rasch-Modellierung den Beobachtungen selbst lediglich Ordinalskalenniveau zugrunde liegen muss. Die Ausprägung der Eigenschaft einer Person (Personenparameter) lässt sich daraus auf Intervallskalenniveau ableiten, und dies sowohl innerhalb der Antwortkategorien eines Items, also auch zwischen den Antwortkategorien verschiedener Items (Rost, 1999). Dies macht die Rasch-Modellierung zur Konstruktion eines Tests zur Lernverlaufsdiagnostik besonders nützlich (Sternberg & Grigorenko, 2002, S. 163f), wenn nicht gar, zumindest aus messtheoretischer Sicht, unumgänglich.

Eindimensionalität

Um die Veränderung in einer spezifischen Fähigkeit zu erfassen, muss das Ergebnis einer Messung auch nur auf diese eine Fähigkeit zurückzuführen sein. So sollte die Leistung in einem Test zur Messung mathematischer Basisfertigkeiten allein durch die mathematischen Fähigkeiten bestimmt sein und nicht durch andere Merkmale wie z.B. sprachliche Kompetenzen, Anstrengungsbereitschaft, Vertrautheit mit dem Aufgabentyp oder Testängstlichkeit. Dies entspricht der Forderung, dass die gemessenen Leistungen in den Items eines Testes auf genau einen gemeinsamen latenten Faktor zurückzuführen sind (Hattie, 1985). Eine solche Eindimensionalität schließt des Weiteren auch eine lokale stochastische Unabhängigkeit der einzelnen

Testitems ein. Praktisch bedeutet dies, dass die Kenntnis um die Lösung eines bestimmten Testitems nicht die Wahrscheinlichkeit verändert, die anderen Items eines Testes zu lösen.

Wichtig ist, dass die Forderung nach Eindimensionalität nicht bedeutet, dass die Intervention, die zu der Veränderung führte, selbst nur eine Fähigkeit fördern darf oder sollte. Sollte aber eine Intervention mehrere Merkmale fördern, so muss für jedes Merkmal, dessen Veränderung evaluiert werden soll, ein eigenes eindimensionales diagnostisches Instrument herangezogen werden.

Itemschwierigkeit

Im Idealfall besteht ein Test zur Lernverlaufsdiagnostik aus einer Reihe exakt gleichschwerer Tests mit exakt gleicher Trennschärfe, die dann den einzelnen jeweiligen Testzeitpunkten zugeordnet werden. Mögliche Veränderungen in den erreichten Testpunkten zwischen zwei Messungen lassen sich dann auf die Veränderung der Fähigkeit einer Person zurückführen (unter der Annahme, dass das Testergebnis durch die Faktoren Itemschwierigkeit und Personenleistung bestimmt wird, wie es in sog. starken Latent-Trait-Modellen angenommen wird (Bühner, 2006, S. 300)). Präziser ausgedrückt soll es sich bei den Testungen um *Paralleltestungen* handeln, also Testungen, die truescore- und fehlervarianzhomogen sind.

Hierbei ergibt sich das Problem, dass es nur unter sehr hohem Aufwand möglich ist, eine ganze Reihe von Tests zu erstellen, die das Kriterium der Paralleltestung erfüllen. Während die Fehlervarianzhomogenität vergleichsweise leicht durch Tests gleicher Reliabilität zu bestimmen ist, lässt sich die Konstanthaltung der Testschwierigkeit (der Äquivalenz der wahren Werte t) nur schwer umsetzen (Stelzl, 1993). Eine Möglichkeit, dieses Problem zu umgehen, wäre, die exakte Schwierigkeit der Tests zu bestimmen und bei Tests unterschiedlicher Schwierigkeit einen Korrekturparameter zu definieren, d.h., eine Gewichtung der Schwierigkeit der Items vorzunehmen[1]. Dazu ist es allerdings notwendig, die Schwierigkeit eines Tests oder Testitems unabhängig von einer spezifischen (Normierungs-) Stichprobe zu bestimmen. Möglich ist dies nur durch eine Trennung von Personen- und Itemparameter. In der KTT sind die Itemparameter jedoch abhängig von den Merkmalen der zugrunde liegenden Stichprobe, da diese sich aus dem Mittelwert (und gegebenenfalls der Varianz) der beobachteten Punkte in einem Test ableiten. In der Logik der Probabilistischen Testtheorie (PTT) lassen sich Itemparameter unabhängig von den Personenparametern ermitteln, da diese auf den geschätzten Wahrscheinlichkeiten beruhen, einen bestimmten Punktwert in einem Test zu erlangen.

Testfairness

Ein letzter wichtiger Aspekt ist die Testfairness. Fairness im umfassenden Sinne bedeutet, dass das ermittelte Testergebnis „zu keiner systematischen Benachteiligung bestimmter Personen aufgrund ihrer Zugehörigkeit zu ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppen“ führt (Moosbrugger & Kelava, 2008, S. 23). Zur Beschreibung, Kontrolle und Verbesserung der Testfairness wurden seit den 60er Jahren verschiedene Modelle ausgearbeitet, die sich

[1] Eine zweite Möglichkeit zur Erreichung gleicher Testschwierigkeit erwächst aus dem Binomialmodell (Klauer, 1972). Einzelne Tests gleicher Schwierigkeit werden dadurch gewonnen, dass zur Erstellung eines Tests mehrere Items zufällig aus einem Pool unterschiedlich schwerer, heterogener und kontentvalider Items gezogen werden. Da dieses Verfahren vor allem im Kontext Curriculum-basierter Tests diskutiert wird, die nicht im Fokus dieses Beitrags stehen, wird auf diese Möglichkeit nicht weiter eingegangen.

auf die Fairness einer Selektionsstrategie bzw. auf die Vorhersage von Kriteriumswerten (vgl. Cleary, 1968; Anastasi, 1968, Thorndike, 1971) oder auf unterschiedliche Abweichungsnormen für definierte Subgruppen, wie Altersklassen, Geschlecht oder Schulform (vgl. Wottawa & Amelang, 1980) beziehen.

Auf testtheoretischer Ebene liegt ein möglicher Grund für die Unfairness eines Tests im Itembias. Itembias bedeutet, dass ein Item aufgrund seiner Charakteristika, die für den Gegenstandsbereich des Tests irrelevant sind, systematisch für verschiedene Personengruppen unterschiedlich schwierig ist. Liegt zudem *Differential Item Functioning* (DIF) vor, bedeutet das weiter, dass für Personen *mit gleicher Fähigkeit* das Item unterschiedlich schwierig ist. Wenn aber zwei Personen die gleiche Fähigkeit besitzen, so sollten sie auch die gleiche Wahrscheinlichkeit haben, ein bestimmtes Item zu lösen. Ist dies nicht der Fall, ist das Item bezüglich der zu messenden Eigenschaft unfair. Zeigt sich dieser Effekt auf der Ebene von sozialen Gruppen (Nationalität, Geschlecht, Alter, etc.), erreichen die Mitglieder einer bestimmten Gruppe bei eigentlich gleicher Kompetenz geringere Testwerte als die Mitglieder einer anderen Gruppe. Dies kann z.B. der Fall sein, wenn Schüler mit Migrationshintergrund bei einem Mathematiktest bei gleicher Mathematikfähigkeit schlechter abschneiden als Muttersprachler, da erstere aufgrund sprachlicher Probleme die Aufgabenstellung bei einigen Items nicht verstehen.

Bei erwarteter Eindimensionalität der Skala (z.B. Mathematikfähigkeit) kann also bei einem Itembias, insbesondere bei Vorliegen von DIF, zugleich die Validität des Items in Frage gestellt werden, denn offensichtlich wird eine weitere Eigenschaft (z.B. sprachliche Kompetenz) gemessen. Die Methode des Differential Item Functioning zur Messung der Fairness ist also auch in der Lage, Aussagen über die Konstruktvalidität eines Tests zu machen.

DIF lässt sich im Rahmen des Item-Response-Modells z.B. mit Hilfe binär logistischer bzw. ordinal logistischer Regression berechnen.

Über die vier bis hierhin vorgestellten Gütekriterien eines Testinstrumentes zur Erfassung von Veränderungen hinaus lassen sich weitere bedeutende Kriterien postulieren. Von zentraler Wichtigkeit ist hierbei die Fähigkeit eines Instrumentes, bei einer erfassten Veränderung zu differenzieren, inwiefern diese das Resultat der Veränderung der (latenten) Fähigkeit einer Person ist, oder aber auf eine Vertrautheit mit dem Aufgabenmaterial bei wiederholter Bearbeitung zurückgeht (Übungseffekte). Letzteres betrifft also die Fähigkeit eines Instrumentes, tatsächlich lernbezogene Leistungsveränderungen zu erfassen.

Der vorliegende Beitrag befasst sich hingegen mit Merkmalen von Tests und Aufgaben, die erfüllt sein müssen, damit weitergehende Testanalysen zur Differenzierung der Ursachen von Veränderungen möglich sind. Erst wenn Reliabilität, Homogenität der Schwierigkeiten, Eindimensionalität und Testfairness gewährleistet sind, lässt sich durch Anwendung probabilistischer Modelle, die den Verlauf des latenten Merkmals über die Zeit hinweg erfassen (z.B. Latente Wachstumsmodelle, Lineare Partial Credit Modelle) und Übungseffekte berücksichtigen, die Veränderung der latenten Fähigkeit bestimmen. Eine solche weitergehende Untersuchung verlangt ein Design mit Messwiederholungen und sollte daher im Anschluss an die in diesem Beitrag vorgestellten Analysen erfolgen.

Vorschlag einer Standardprozedur zur Prüfung der Güte eines Tests zur Lernverlaufsdiagnostik

Aus den vorgestellten Überlegungen heraus schlagen wir vor, dass bei der Erstellung eines Instrumentes, das sich zur Lernverlaufsdiag-

nostik nutzen lässt, im ersten Schritt folgende Analysen durchgeführt werden:

1. Rost (1999) argumentiert, dass die KTT als Messfehlertheorie komplementär zur PTT ist, da letztere eine Analyse des Messfehlers nicht durchführt. Diese Sichtweise soll auch hier geteilt werden. Zwar lassen sich auch in der PTT Konfidenzintervalle der Ausprägung eines Personenmerkmals berechnen, diese betreffen aber die Varianz des zu messenden (latenten) Merkmals selbst, und nicht des Messinstrumentes, werden also als stochastische Fluktuation des Merkmals und nicht als Messfehler verstanden. Daher soll zunächst eine Itemanalyse auf Basis der KTT, bei der die Trennschärfen, Schwierigkeiten, Interne Konsistenz, Homogenität und, soweit möglich, Retest-Reliabilität, durchgeführt werden.
2. Des Weiteren wird die Eindimensionalität des Konstruktes durch eine konfirmatorische Faktorenanalyse geprüft.
3. Zur genauen Bestimmung der Itemschwierigkeiten wird eine erweiterte polytome Raschmodellierung durchgeführt. Die Item- und Schwellenparameter lassen bei hinreichendem Modellfit eine t-äquivalente Messung auf Itemebene und Gesamttestebene zu. Wichtig ist hierbei, auch die Homogenität der Trennschärfen zu überprüfen (z.B. durch inferenzstatistische Testung der Q-Indizes (Bühner, 2011, S. 544 f))
4. Abschließend wird die Testfairness für relevante Subgruppen, die im Fokus der Testanwendung stehen, durch Berechnung von DIF-Werten geprüft. Dies geschieht mit Hilfe von ordinal logistischen Regressionsanalysen, die auf der Item-Response-Theorie beruhen. Neben der inferenzstatischen Absicherung des DIF werden zudem der Typ (uniform vs. nicht-uniform), die Richtung (welche Gruppe benachteiligt wird) sowie die Effektstärke des Bias angegeben.

Methode

Die vorgeschlagene Verfahrensweise soll im Folgenden am Beispiel eines neuen Instrumentes zur Erfassung des Sprachstandes von Kindern mit einer Lernbeeinträchtigung demonstriert werden. Das Instrument besteht insgesamt aus zehn Texten im Aufgabenformat eines C-Tests. C-Tests messen „globale Sprachkompetenz“ (Grotjahn, 2002; Grotjahn, Klein-Braley & Raatz, 2002), wenn auch die pragmatische und die illokutionäre Dimension u.E. weniger gut erfasst werden. C-Tests dienen meist der Feststellung der allgemeinen Kompetenz in Fremd-, Zweit- und Erstsprachen und werden dabei z.B. als Einstufungstests an Universitäten, Studienkollegs und Sprachschulen eingesetzt. Die Validität von C-Tests für die hier angestrebte Zielgruppe konnte von Linnemann und Wilbert (Linnemann & Wilbert, 2010; Wilbert & Linnemann, in Vorbereitung) bereits nachgewiesen werden. Ziel soll es nun sein, die zehn Texte dahingehend zu überprüfen, ob diese sich aus messtheoretischer Sicht eignen, zur Lernverlaufsdiagnostik oder formativen Evaluation eingesetzt zu werden.

Stichprobe

An der Erhebung nahmen 268 Schüler und Schülerinnen der Förderschule Lernen aus 32 Schulen in Nordrhein-Westfalen teil. Die Daten von 15 Schülern wurden nicht in der Auswertung berücksichtigt, da diese die Aufgaben gar nicht oder nur unvollständig bearbeiteten. Das Alter der verbleibenden 253 Schüler und Schülerinnen lag zwischen 11 und 17 Jahren (M = 14.2; SD = 1.6). 57 % (145) der Schüler waren männlich und 66% (167) gaben an, dass ihre Muttersprache Deutsch sei. Neben Deutsch wurden 18 weitere Sprachen als Muttersprache angegeben. Die größte Gruppe hiervon bildeten 26 Kinder (10.3%) mit Türkisch als Muttersprache. Tabelle 1

Tab. 1: Verteilung der Klassenstufen in der Stichprobe

Klassenstufe	Häufigkeit	Prozent
5	8	3.2
6	8	3.2
7	74	29.2
8	56	22.1
9	55	21.7
10	52	20.6

zeigt die Verteilung der Schüler auf die Klassenstufen.

Material und Durchführung

Es wurden zehn Texte im Format eines C-Tests verwendet. Die Texte stammen aus vorherigen unveröffentlichten Untersuchungen und sind in zwei Parallelformen mit jeweils fünf Texten aufgeteilt (Texte A1 bis A5 und B1 bis B5). In der vorliegenden Untersuchung wurden sie als ein einziges Instrument mit zehn Aufgaben eingesetzt. Die Texte haben einen Umfang von jeweils 60 bis 80 Wörtern und sind nach dem sogenannten „kanonischen Prinzip" aufgebaut. Dies bedeutet, dass bei jedem Text ab dem zweiten Wort des zweiten Satzes die Hälfte jedes zweiten Wortes gelöscht wurde. Die Überschrift und der letzte Satz bleiben jeweils unangetastet. Jeder der Texte enthält 20 Lücken, so dass zwischen 0 und 20 Punkte pro Text erreicht werden können. Abbildung 1 zeigt ein Beispiel eines so konstruierten Textes.

C-Tests beruhen auf dem Prinzip der *reduzierten Redundanz*, da durch die getilgten Zeichen mehrfach kodierte (und somit redundante) Phänomene (z.B. Pluralformen) seltener vorkommen als im Original. Derart reduzierte Informationen machen es schwieriger, aber nicht unmöglich, Information aus einem Text zu entnehmen. Eine höhere sprachliche Fähigkeit sollte also mit einer größeren Lösungshäufigkeit einhergehen. Grotjahn (1992) argumentiert, dass Lerner mit steigender Kompetenz in einer Fremdsprache zunehmend besser C-Tests lösen können, da sie im Zuge einer konstruktiven und antizipatorischen Verarbeitung von Sprache

Die Popkomm - die größte Party der Welt

Im Sommer 1992 feierten die Menschen in Köln eine große Musikveranstaltung.

Sie w_____ sehr erfolg_____. Deshalb wol_____ man e_____ solches Fe_____

jedes Ja_____ im Som_____ machen. M_____ nannte e_____ Popkomm.

B_____ zum Ja_____ 2003 si_____ jährlich zw_____ Millionen Besu_____ zur

Popkomm geko_____. Sehr schn_____ wurde s_____ die grö_____

Musikveranstaltung d_____ Welt. Se_____ 2004 findet die Popkomm in Berlin

statt: mit noch mehr Bands, Bühnen und Besuchern.

Abb. 1: Der Text A1 als Beispiel eines C-Test-Textes (aus Linnemann & Wilbert, 2010).

Gebrauch von der Redundanz der Sprache machen.

Zur Erhebung der Daten wurden die teilnehmenden Schüler und Schülerinnen in kleine Gruppen (ca. fünf Kinder) eingeteilt. Da die komplette Bearbeitung in einer einzigen Sitzung die Belastbarkeit der Schülerinnen und Schüler überschritt, wurden zwei Sitzungen durchgeführt, die maximal eine Woche auseinander lagen. Es lässt sich aufgrund des sehr kurzen Intervalls zwischen den beiden Testungen ausschließen, dass die Sprachkompetenz der Schülerinnen und Schüler sich systematisch und bedeutsam verbessert hat. In der ersten Sitzung wurden die fünf Texte A1 bis A5 erhoben, in der zweiten Sitzung die fünf Texte B1 bis B5. Für die Bearbeitung jedes Textes standen maximal fünf Minuten zur Verfügung. Danach wurde zum nächsten Text weitergeblättert.

Ergebnisse

Itemanalyse

Tabelle 2 ist zu entnehmen, dass die Items eine nah beieinanderliegende mittlere Punktzahl (zwischen 8.4 und 11.6) sowie Standardabweichung der Punkte (zwischen 4.6 und 5.3) haben. Die daraus ermittelten Schwierigkeitsindizes weisen die Aufgaben als mittelschwer bis schwer mit recht homogenen Werten aus (*P* zwischen .23 und .38).

Betrachtet man die 10 Items als eine Gesamtskala, weist diese eine sehr hohe interne Konsistenz und Homogenität auf (Cronbachs α = .967; mittlere inter-item-Korrelation = .746). Die korrelierte Split-half-Reliabilität bei Verwendung aller 10 Items beträgt .967

Konfirmatorische Faktorenanalyse

Zur Prüfung der Annahme, dass durch die C-Tests eine einzige Fähigkeitsdimension erfasst wird, wird eine konfirmatorische Faktorenanalyse durchgeführt. Ein χ^2-Test weist ei-

Tab. 2: Kennwerte der Itemanalyse, konfirmatorischen Faktorenanalyse sowie Raschskalierung

Text	M	SD	P[1)]	$r_{i(t-i)}$[2)]	R^2	β[3)]	σ[4)]	Q[5)]
A1	10.6	5.3	.35	.84	.73	.85	-0.38	.058
A2	8.4	4.8	.23	.79	.63	.80	0.56	.073
A3	9.8	4.7	.30	.85	.74	.87	-0.02	.055
A4	9.8	5.5	.31	.86	.75	.87	-0.05	.048
A5	9.2	5.3	.28	.84	.72	.86	0.19	.055
B1	11.2	5.2	.38	.88	.77	.89	-0.56	.043
B2	10.0	5.1	.31	.87	.79	.89	-0.03	.047
B3	10.7	5.1	.35	.86	.76	.88	-0.33	.048
B4	9.6	5.1	.30	.87	.78	.89	-0.01	.045
B5	8.5	4.6	.23	.82	.69	.84	0.62	.073

Anmerkung: [1] Schwierigkeitsindex für mehrstufige Items nach Fisseni (1997, S. 45); [2] korrigierte Trennschärfe; [3] standardisierte Pfadkoeffizienten der konfirmatorischen Faktorenanalyse bei Annahme eines latenten Faktors; [4] Item locations (Schwierigkeitsparameter) des Partial Credit Models; [5] Indizes des Steigungsparameters (Trennschärfeparameter) des Partial Credit Models.

ne signifikante Abweichung der Beobachtungsdaten vom erwarteten Datenmuster bei Annahme eines einfaktoriellen Modells auf ($\chi^2(35) = 67$; $a < .001$). Der zu den Freiheitsgraden ins Verhältnis gesetzte χ^2-Wert bleibt allerdings unter zwei und weist auf Modellpassung hin ($\chi^2/df = 1.91$). Insgesamt kann der χ^2-Test nur als erster Indikator der Modellpassung gesehen werden, da er in hohem Maß von der Stichprobengröße abhängig ist. Bei einer Stichprobengröße von $n > 200$, wie sie in dieser Studie vorliegt, empfehlen sich daher Modellpassungsindizes, die diese Einschränkung nicht mit sich bringen. Dazu gehören der Goodness of fit *(GFI)* und der normierte Fit *(NFI)*, die bei der vorliegenden Analyse eine gute Modellpassung nachweisen (*GFI* = .943; *NFI* = .975). Des Weiteren beträgt der *RMSEA* .060. Mit einem close-fit Wert *PCLOSE* = .206 verwerfen wir die Hypothese, dass *RMSEA* signifikant über .05 liegt, und bestätigen die im Modell angenommene einfaktorielle Struktur. Ergänzend sind in Tabelle 2 (Spalte 7) die standardisierten Regressionsgewichte des Modells aufgezeigt.

Erweiterte Rasch-Modellierung

Da die erhobenen Daten ordinalskaliert sind (entsprechend den zu erreichenden Punkten pro Text auf einer Skala von 0 bis 20), wurden zur Analyse erweiterte Raschmodelle herangezogen. Hier bieten sich zwei alternative Modelle an. Das *Partial Credit Model* (PCM; Masters, 1982) und das *Rating Scale Model* (RSM; Andrich, 1978). Beide Modelle setzen eine ordinale Skalierung der Items voraus. Ebenso wird in beiden Modellen angenommen, dass die Wahrscheinlichkeit dafür, dass eine Person bei einem Item eine bestimmte Kategorie auswählt (oder erreicht), sich aus der Differenz der Fähigkeit einer Person und der Schwierigkeit des Erreichens der entsprechenden Kategorie ergibt. Dabei sind die Schwierigkeiten der einzelnen Kategorien innerhalb eines Items geordnet. Der zentrale Unterschied zwischen den beiden ist, dass im RSM das Verhältnis der Intervalle zwischen den einzelnen Kategorien eines Items für alle Items gleich ist, während im PCM diese Intervalle von Item zu Item variieren können.

Das PCM beschreibt die Wahrscheinlichkeit, dass eine Person n bei einem Item i den Wert x erreicht. Es stellt sich mathematisch folgend dar:

$$(1)\quad P\{X_{ni} = x\} = \frac{exp\sum_{k=0}^{x}(\theta_n - \tau_{ki})}{\sum_{x=0}^{m} exp\sum_{k=0}^{x}(\theta_n - \tau_{ki})}$$

wobei Θ_n die Fähigkeit einer Person *n* ist und τ_{ki} die Schwierigkeit des Erreichens der Kategorie *k* des Items *i* (Schwellenparameter).

Das RSM nimmt zusätzlich einen Parameter σ_i an, der die Schwierigkeit eines Items bezeichnet. Im Unterschied zum PCM werden die Schwellenparameter τ_k als konstant für alle Items angenommen. Die jeweilige Schwierigkeit ergibt sich nun durch Subtraktion des Schwellenparameters von der jeweiligen Itemschwierigkeit. Die Modellformel lautet:

$$(2)\quad P\{X_{ni} = x\} = \frac{exp\sum_{k=0}^{x}(\theta_n - (\sigma_i - \tau_k))}{\sum_{x=0}^{m} exp\sum_{k=0}^{x}(\theta_n - (\sigma_i - \tau_k))}$$

Welches der beiden Modelle anzuwenden ist, soll hier empirisch bestimmt werden. Dem Gebot der Sparsamkeit folgend, ist das weniger komplexe RSM bei gleich guter Eignung dem PCM vorzuziehen.

Neben der Entscheidung für das passende Modell ergibt sich das Problem, dass aufgrund der relativ kleinen Stichprobe bei gleichzeitig vielen (21) Kategorien pro Item, die Datenbasis zu spärlich ist, um eine verlässliche Schätzung der Modellparameter zu erlangen. Daher wurden die 21 Kategorien (0 bis 20) pro Item auf eine Skala mit sieben Kategorien (0 bis 6) umskaliert (die Punkte pro Text wurden durch drei dividiert und der resultierende Wert auf die nächste natürliche Zahl abgerundet).

Die im Folgenden dargestellten Analysen wurden mit den Programmen Winmira (von Davier, 1997) sowie dem eRm Packetmodul (Mair, Hatzinger & Maier, 2010) des Programms R (R Development Core Team 2008) berechnet. Bei beiden Programmen basieren die Berechnungen auf Conditional-Maximum-Likelihood (CML) Schätzern.

Ein Vergleich der Modellierung im PCM und RSM (Tabelle 3) zeigt eine sehr hohe Reliabilität in beiden Modellen. Ebenso liegen die Q-Indizes (Indikatoren für die Homogenität der Trennschärfeparameter[2]) bei beiden Modellen im gleichen Wertebereich und nahe an der Form eines Guttman-Antwortmusters. Hinsichtlich der Homogenität der Q-Indizes werden im RSM ein Item mit Overfit und zwei Items mit Underfit identifiziert, während im PCM-Modell keine Abweichungen auftreten. Die geschätzten Schwellenparameter zeigen in beiden Modellen keine Verletzung der Ordnung auf. Da das PCM bei jedem Item unterschiedliche Verhältnisse der Schwellenparameter zulässt, müssen deutlich mehr Parameter geschätzt werden als im RSM (119 vs. 74). Ein Vergleich der Komplexität der Modelle bei der Passung auf die empirischen Daten (Informationskriterien) weist das PCM als leicht komplexer aus. Aufgrund der kleinen Stichprobe im Verhältnis zu den möglichen Antwortmustern in beiden Modellen wurden Bootstrapverfahren zur Schätzung der Modellgüte eingesetzt. Beide Modelle zeigen einen hinreichenden Fit der empirisch vorliegenden Daten (Cressie Read Test und Pearson χ^2 Test jeweils mit $p > .05$).

Aufgrund des fehlenden Fits bei drei der zehn Items im RSM wird das PCM zugrunde gelegt und für die weiteren Analysen verwendet. Unter Verwendung dieses Modells darf angenommen werden, dass die Items raschskaliert sind. Tabelle 2 (Spalte 8 und 9) zeigt die Item Locations (Schwierigkeitsparameter; Logitwert, der dem Mittelwert der Schwellenparameter des entsprechenden Items entspricht) sowie die Q-Indizes auf. Letztere können als homogen angesehen werden. Daher lässt sich für die einzelnen Items essentielle τ-Äquivalenz postulieren. Dabei werden im PCM nicht nur die Itemparameter bestimmt, sondern auch die Parameter der Übergänge von einer Antwortkategorie zur nächsten innerhalb eines jeden Items (die sog. Schwellenparameter). Auch für diese Schwellenparameter kann essentielle τ-Äquivalenz angenommen werden.

Abbildung 2 stellt die sechs Schwellenparameter für jedes Item graphisch dar. Auch die Schwellenparameter werden in einer Logitskala ausgedrückt und sind dadurch direkt mit den Personenparametern vergleichbar. Dies bedeutet, dass der Schwellenwert sowohl die Schwierigkeit des Erreichens der Ka-

[2] Der Q-Index gibt an, inwiefern das beobachtete Antwortmuster auf eine Reihe von Items einem Guttmann Muster folgt. Bei einem Q-Index von 0 liegt ein perfektes Guttmann Antwortmuster vor, bei einem Wert von 1 ein perfektes Anti-Guttmann Muster, ein Wert von 0.5 verweist auf fehlende Trennschärfe. Als geeignet gelten Werte < .30. Die Formel zur Berechnung des Q-Index lautet:

(3)

$$Q_i = \frac{ln\frac{P(\vec{X}_{beo})}{P(\vec{X}_{GP})}}{ln\frac{P(\vec{X}_{AGP})}{P(\vec{X}_{GP})}}$$

wobei $P(\vec{X}_{beo})$ die Wahrscheinlichkeit dafür ist, dass der Vektor des Antwortmusters den beobachteten Werten entspricht, $P(\vec{X}_{GP})$ die Wahrscheinlichkeit, dass der Vektor einem Guttmann-Muster folgt, und $P(\vec{X}_{AGP})$ die Wahrscheinlichkeit für ein Anti-Guttmann-Muster (ausführlich siehe Bühner, 2011, S. 545).

Tab. 3: Modellparameter im Vergleich zwischen Partial Credit Model (PCM) und Rating Scale Model (RSM)

	PCM	**RSM**
Andrichs Reiliabilität	.953	.952
Q-Index	.043 bis .073	.043 bis .073
Items mit Overfit	0	1
Items mit Underfit	0	2
Anzahl ungeordneter Schwellen	0	0
Anzahl Parameter des Modells	119	74
Informationskriterium		
AIC	7161	7156
BIC	7582	7417
CAIC	7701	7491
Modellfit[1)]		
p-Wert [1)] Cressie Read Test	.17	.08
p-Wert [2)] χ^2-Test	.32	.11

Anmerkungen: [1] ermittelt durch parametrisches Bootstrapverfahren mit 250 Samples; [2] beruhend auf der empirischen Dichtefunktion.

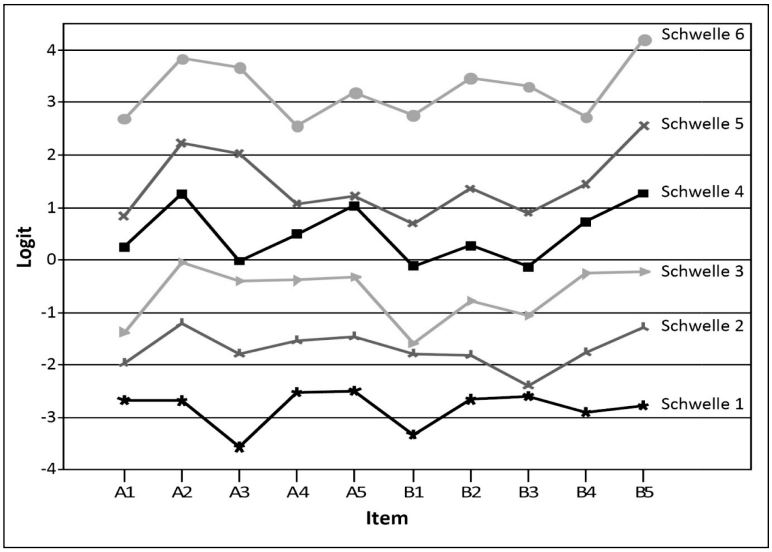

Abb. 2: Darstellung der Schwellenparameter der zehn verwendeten Items. Schwelle 1 beschreibt den Wert Θ, bei dem die Wahrscheinlichkeiten einer Person, in Kategorie 0 oder Kategorie 1 zu sein, gleich groß sind. Entsprechendes gilt für die Schwellen 2 bis 6.

tegorie als auch die Fähigkeit, die eine Person haben muss, um diese Kategorie zu erreichen, beschreibt. Exakt ausgedrückt, beschreibt der Schwellenparameter die Fähigkeitsausprägung einer Person, bei der die Wahrscheinlichkeit gleich groß ist, in die Kategorie über oder unter der Schwelle zu fallen.

Abbildung 3 soll dies am Beispiel des Items A1 verdeutlichen. Der Schwellenpara-

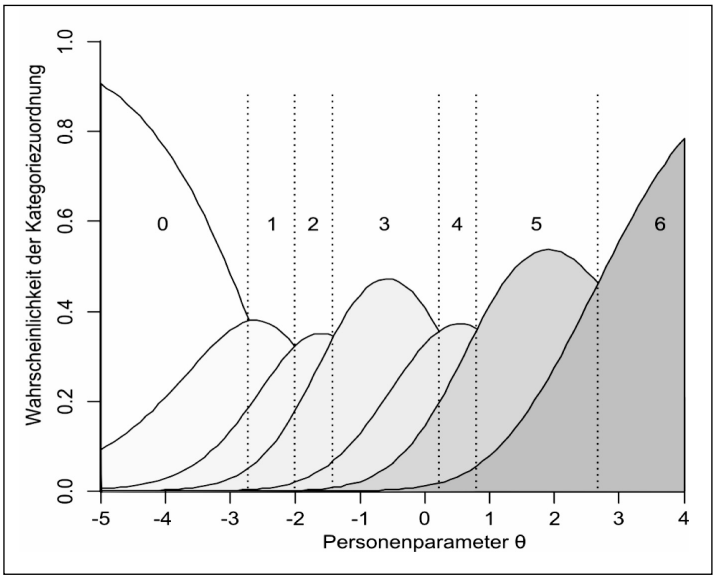

Abb. 3: Kategoriefunktionen des Items A1. Die gestrichelten Linien geben die Schwellenparameter an.

meter 1 (Übergang zwischen Kategorie 0 und 1) liegt bei einer Personenfähigkeit Θ = -2.67. Eine Person mit dieser Fähigkeitsausprägung wird mit gleich großer Wahrscheinlichkeit (ca. 38%) bei diesem Item die Kategorie 0 oder 1 erreichen (und mit einer Restwahrscheinlichkeit von ca. 24% eine der höheren Kategorien).

Umgekehrt lässt sich schlussfolgern, dass eine Person, die bei Item A1 die Kategorie 1 erreicht hat, mit der höchsten Wahrscheinlichkeit eine Personenfähigkeit zwischen Θ = -2.67 (Schwellenparameter 1) und Θ = -2.00 (Schwellenparameter 2) hat. Entsprechendes lässt sich von jeder Kategorie jedes Items bestimmen. In Tabelle 4 wurde eine solche Zuordnung realisiert. Zur größeren Anschaulichkeit wurden die Logitwerte linear transformiert (Fähigkeitswert = logit × 10 + 50). Jetzt lässt sich bei einem einzelnen Item anhand der erreichten Kategorie feststellen, welche Fähigkeit die getestete Person hat. Diese Personenfähigkeiten sind über alle Items hinweg miteinander vergleichbar.

Differential Item Functioning

Differential Item Functioning (DIF) lässt sich mit Hilfe der logistischen Regression (LR) ermitteln, die gegenüber anderen Verfahren für die vorliegende Studie drei Vorteile besitzt: (1) Das Modell kann von binären auf ordinale Antwortkategorien erweitert werden, (2) neben der inferenzstatistischen Absicherung des DIF liegt ein Maß für die Effektstärke vor und (3) das Modell kann sowohl uniformes als auch nicht-uniformes DIF gleichzeitig untersuchen (vgl. Zumbo 1999). Letztere Formen des DIF unterscheiden sich dahingehend, dass beim uniformen DIF lediglich die Schwierigkeit (Item Location) für verschiedene Gruppen unterschiedlich ist, was sich grafisch als parallel verschobene Item-Characteristic-Kurve (ICC) ausdrückt. Bei Vorliegen dieses Typs lässt sich eine Benachteiligung also direkt ablesen. Nicht-uniformes DIF bezieht den Trennschärfeparameter eines Items ein, d.h. nicht-uniformes DIF liegt vor, wenn für verschiedene Gruppen unterschiedliche Trennschärfeparameter eines Items vorliegen. Grafisch drückt sich dies als nicht parallele ICCs aus. Ob es sich hierbei um eine Benach-

Tab. 4: Umrechnung der erreichten Punkte pro Test (Rohwerte RW) in Fähigkeitswerte. Der Bereich der Rohwerte entspricht jeweils einer Kategorie (Kat.). Angegeben ist das Intervall der Fähigkeitswerte, innerhalb dessen die Wahrscheinlichkeit am höchsten ist, einen entsprechenden Rohwert zu erreichen (linear transformierte Logitwerte der angrenzenden Schwellenparameter).

		Item									
RW	Kat.	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5
0 - 2	0	< 23	< 23	< 14	< 24	< 25	< 16	< 23	< 24	< 20	< 22
3 - 5	1	23 - 29	23 - 36	14 - 31	24 - 33	25 - 34	16 - 31	23 - 30	24	20 - 31	22 - 36
6 - 8	2	30 - 35	37 - 48	32 - 45	34 - 45	35 - 45	32 - 33	31 - 41	25 - 38	32 - 46	37 - 46
9 - 11	3	36 - 51	49 - 61	46 - 48	46 - 53	46 - 59	34 - 47	42 - 51	39 - 47	47 - 56	47 - 61
12 - 14	4	52 - 57	62 - 71	49 - 69	54 - 59	60 - 61	48 - 55	52 - 62	48 - 57	57 - 63	62 - 74
15 - 17	5	58 - 76	72 - 88	70 - 86	60 - 75	62 - 81	56 - 77	63 - 84	58 - 83	64 - 77	75 - 91
18 - 20	6	> 76	> 88	> 86	> 75	> 81	> 77	> 84	> 83	> 77	> 91

teiligung einer Gruppe handelt, lässt sich nur durch weitergehende Analysen feststellen.

Eine Verfahrensweise, DIF aufzuklären, ist der Vergleich von drei logistischen Regressionsmodellen, bei denen sukzessiv die Parameter Fähigkeit (Modell 1), Gruppe (Modell 2) und die Interaktion Fähigkeit × Gruppe (Modell 3) hinzugefügt werden. Getestet wird hier der Beitrag jedes Terms. Mit Hilfe dieser Analyse wird gleichzeitig auf uniformes und nicht-uniformes DIF getestet, indem ein χ^2-Differenztest zwischen Modell 1 und Modell 3 durchgeführt wird. Die resultierende Statistik folgt einer χ^2-Verteilung mit 2 Freiheitsgraden (Swaminathan & Rogers, 1990). Die Effektstärke des DIF ergibt sich analog dazu aus der Subtraktion der entsprechenden Effektstärkemaße (R^2 als Pseudo-Bestimmtheitsmaß, vgl. McKelvey & Zavoina, 1975) der logistischen Regression ($\Delta R^2 = R^2_{\text{Modell 3}} - R^2_{\text{Modell 1}}$). Durch die sequentielle Vorgehensweise wird es zudem möglich, die spezifischen Anteile von uniformem und nicht-uniformem DIF zu berechnen. Die Effektstärke des uniformen DIF lässt sich als Differenz aus dem ersten und zweiten Modell begreifen, die Effektstärke des nicht-uniformen DIF aus der Differenz von Modell 2 und Modell 3. Als praktisch relevante Größe, um von DIF zu sprechen, geben Zumbo (1999) und Zumbo & Thomas (1997) neben der Signifikanz die Werte $.13 \leq \Delta R^2 \leq .26$ für ein moderates, $R^2 > .26$ für ein großes DIF an, Jodoin & Gierl (2001) die Werte $.035 \leq \Delta R^2 \leq .070$ für ein moderates, $R^2 > .070$ für ein großes DIF.

Tabelle 5 zeigt die Zusammenfassung der DIF-Analyse bezüglich der Variable Geschlecht. Signifikantes DIF auf dem (nicht α-korrigierten) 5%-Niveau zeigt ausschließlich der Text A2. Bei gleicher Kompetenz werden hier die weiblichen Probanden benachteiligt. Die Effektstärke ist jedoch mit 0.81% weitaus geringer, als von Jodoin & Gierl (2001) gefordert. Auf eine Analyse, um welchen Typ DIF es sich handelt, wird daher hier verzichtet.

Ein ähnliches Bild zeigt sich bei der Unterscheidung von Deutsch als Zweitsprachelernern (DaZ) und Schülerinnen und Schülern mit der Muttersprache Deutsch (DaM; siehe Tabelle 6). Signifikantes DIF zeigt sich auch hier lediglich bei Text A2 mit einer sehr geringen Effektstärke von 0.93%. Auf eine weitere Analyse wird aufgrund der geringen Effektstärke verzichtet.

Insgesamt zeigt sich der C-Test als fair bezogen auf die Geschlechter und die Gruppen Schülerinnen und Schüler mit der Erst- und Zweitsprache Deutsch. Das Item A2 zeigt zwar hinsichtlich beider Gruppen DIF, jedoch ist die Höhe des Effektes sehr gering.

Tab. 5: Ergebnisse der DIF-Analyse für die Gruppierungsvariable Geschlecht

Text	**Modell 1[1)]**	**Modell 2**	**Modell 3**	**Test auf uniformes und nicht-uniformes DIF**					
	χ^2	R^2	χ^2	R^2	χ^2	R^2	$\bullet\chi^2$	p[2)]	$\bullet R^2$
A1	284.97	.690	285.06	.690	285.56	.691	.59	.75	.0009
A2	291.55	.707	295.28	.712	298.17	.715	6.62	.04*	.0081
A3	339.46	.765	34.53	.768	345.09	.772	5.63	.06	.0067
A4	379.16	.802	379.27	.802	379.54	.803	.38	.83	.0005
A5	342.42	.761	342.42	.761	344.22	.763	1.80	.41	.0017
B1	394.74	.818	398.68	.822	398.92	.822	4.18	.12	.0037
B2	383.69	.804	387.62	.808	388.51	.809	4.82	.09	.0045
B3	326.33	.747	326.51	.747	328.49	.748	2.15	.34	.0013
B4	381.02	.803	381.02	.803	381.03	.803	.01	1.00	.0000
B5	299.94	.718	301.64	.721	303.26	.722	3.33	.19	.0048

Anmerkungen: * = p < .05; [1] Modell 1 bis 3: Modelle der logistischen Regression ;[2] df = 2.

Tab. 6: Zusammenfassung der DIF-Analyse für die Gruppen Schülerinnen und Schüler mit DaZ und DaM

Text	**Modell 1[1)]**	**Modell 2**	**Modell 3**	**Test auf uniformes und nicht-uniformes DIF**					
	χ^2	R^2	χ^2	R^2	χ^2	R^2	$\Delta\chi^2$	p[2)]	ΔR^2
A1	274.25	.701	274.70	.702	275.77	.703	1.52	.47	.0025
A2	268.15	.701	271.88	.706	274.90	.710	6.75	.03*	.0093
A3	316.63	.765	317.43	.767	32.86	.770	4.23	.12	.0051
A4	355.25	.804	355.33	.804	355.53	.804	.28	.87	.0004
A5	32.69	.761	32.69	.761	323.34	.764	2.65	.27	.0025
B1	372.76	.823	376.41	.826	376.63	.826	3.87	.14	.0037
B2	365.89	.812	369.82	.816	37.99	.817	5.10	.08	.0052
B3	312.11	.755	312.35	.755	314.05	.757	1.94	.38	.0016
B4	356.65	.804	356.74	.804	356.82	.804	.17	.92	.0003
B5	281.52	.719	282.29	.722	283.29	.723	1.77	.41	.0038

Anmerkungen: * = p < .05; [1] Modell 1 bis 3: Modelle der logistischen Regression ;[2] df = 2.

Diskussion

Ziel des vorliegenden Beitrags ist es zum einen, Kriterien, die ein Test zur Lernverlaufsdiagnose erfüllen muss, vorzustellen, zum anderen aber auch eine Verfahrensweise der Testanalyse zu erarbeiten, an der diese Kriterien geprüft werden können. Exemplifiziert wurde diese Verfahrensweise an einem neu entwickelten Testinstrument zur Bestimmung des Sprachstandes lernbeeinträchtigter Schüler im Format eines C-Tests.

Die Analyse nach der vorgeschlagenen Prozedur ergab zunächst eine sehr gute Passung des Testes nach Maßen der KTT: Interne Konsistenz, Trennschärfen, Split-half- und Paralleltest-Reliabilität können als ausgezeichnet bezeichnet werden. Es ist also davon auszugehen, dass das Instrument als Ganzes einen geringen Messfehler aufweist.

Des Weiteren weist das Instrument eine eindimensionale Struktur auf: Die Items haben eine hohe Homogenität. Die Struktur eines latenten Faktors kann in einer konfirmatorischen Faktorenanalyse bestätigt werden.

Zur Kontrolle der Itemschwierigkeit wurde eine Analyse nach dem erweiterten Rasch-Modell vorgenommen. Durch die Anwendung des PCM ließ sich erfolgreich eine Trennung von Personen- und Itemparameter umsetzen. Dadurch sind nicht nur alle Items zueinander, sondern auch auf der Ebene der Kategorien innerhalb jedes Items τ-äquivalent. Anders ausgedrückt lassen sich die Schwierigkeiten der Kategorien innerhalb jedes Items genau bestimmen. Dies ermöglicht selbst bei Items unterschiedlicher Schwierigkeit, anhand der Testleistung einer Person exakt deren Merkmalsausprägung zu bestimmen.

Zur Analyse der Testfairness wurden Analysen zum DIF durchgeführt. Im Hinblick auf mögliche kritische Effekte des Instrumentes sind die beiden Gruppierungen *Geschlecht* und *Muttersprache* (DaZ und DaM) untersucht worden. Dabei zeigt sich nur bei wenigen Items ein sehr schwaches DIF. Aus dieser Sicht kann der Test als fair für die untersuchten Gruppen bezeichnet werden. Dies schließt aber grundsätzlich nicht aus, dass für andere, hier nicht berücksichtigte Gruppen, DIF besteht. Welche Gruppen dies allerdings sein sollen, müsste entweder theoretisch abgeleitet oder explorativ durch eine Latente-Klassen-Analyse (Rost, 1985) aufgezeigt werden.

Nach diesen Analysen können die schwierigkeitskorrigierten Testleistungen (siehe Tabelle 4) als reliable und (konstrukt)valide Messungen der Fähigkeit einer Person angesehen werden. Da diese Fähigkeit durch jedes Items äquivalent bestimmt werden kann, können die einzelnen C-Test-Texte beliebig zur wiederholten Lernverlaufsmessung herangezogen werden. Die Differenzen der einzelnen Messergebnisse lassen sich direkt in eine Veränderung der Fähigkeit der Person übersetzen.

Kritikpunkte

Die hier vorgeschlagene Verfahrensweise weist eine Reihe unklarer Punkte auf:

1. Die Analyse nach der KTT wurde zu Beginn auf der Basis der von den Schülerinnen und Schülern erreichten Punkte im C-Test durchgeführt. Rost (1999) argumentiert, dass bei der Konfundierung von Personenfähigkeit und Itemschwierigkeit lediglich Beobachtungen und keine Messungen im engeren Sinne vorliegen. Hingegen ist das Ergebnis einer Raschmodellierung die Transformation der Beobachtungswerte in Messwerte, die erst dann von der KTT aufgegriffen und hinsichtlich ihrer Reliabilität untersucht werden können. Würde man diesem Argument hier folgen, so müsste zunächst über die Modellierung im erweiterten Rasch-Modell für jede Person bei jedem Item die Personenfähigkeit geschätzt werden. Erst anschließend sollten dann die Analysen im Rahmen der klassischen Testtheorie mit der resultierenden

Matrix von Messwerten durchgeführt werden.

2. Als nächstes stellt sich die Frage, ob eine Testung auf Eindimensionalität mittels Faktorenanalysen überhaupt nötig ist, da diese auch im Rasch-Modell und seinen Erweiterungen geprüft wird. Diesem lässt sich entgegenhalten, dass im Rahmen der Testung des erweiterten Rasch-Modells nicht nur die Eindimensionalität, sondern zugleich auch weitere Modellmerkmale geprüft werden (z.B. die Homogenität der Trennschärfen bzw. Diskriminanzparameter). Sollte also das Rasch-Modell nicht passen, so bleibt die Ursache dafür ohne weitere Analysen unklar. Des Weiteren können in der konfirmatorischen Faktorenanalyse problemlos Items unterschiedlicher Trennschärfe berücksichtigt werden (Ewing, Salzberger & Sinkovics, 2009). Ebenso lässt sich der Fit mehrfaktorieller Strukturen (evtl. auf mehreren Ebenen) prüfen, die zudem auch korreliert sein können. Hier zumindest ist die konfirmatorische Faktorenanalyse deutlich flexibler als das einfache oder erweiterte Rasch-Modell.
3. Im Hinblick auf die Umrechnung der erreichten Antwortkategorie in einem Item in eine Personenfähigkeit (Tabelle 4) bleibt offen, ob die Angabe eines Intervalls auf der Grundlage der anliegenden Schwellenparameter die optimale Information ist. Alternativ ließe sich auch das Maximum der entsprechenden Kategoriefunktion berechnen (siehe die Gipfel der Berge in Abbildung 3) und die entsprechend zugeordnete Personenfähigkeit angeben.
4. Bei den Analysen zur Testfairness wurde sowohl auf uniformes als auch auf nicht-uniformes DIF getestet. Es stellt sich die Frage, ob eine unterschiedliche Ausprägung der Trennschärfen eines Items für zwei Gruppen, wie sie durch ein nicht-uniformes DIF ausgedrückt wird, tatsächlich die Testfairness betrifft, da nicht direkt bestimmt werden kann, welche Gruppe hierdurch benachteiligt sein soll. Evtl. verweist das nicht-uniforme DIF nicht auf fehlende Fairness, sondern auf schlechtere Reliabilität eines Items für eine Untergruppe. Entsprechend könnte diese Komponente im Kontext der Bestimmung der Testfairness entfallen.

Ausblick

Die hier vorgestellte Verfahrensweise ist der erste Schritt hin zu einem Messinstrument, das für den Einsatz von Lernverlaufsmessungen tauglich ist. Für die vorgestellten Analysen ist eine Datenerhebung an einem einzigen Zeitpunkt ausreichend, der Aufwand der Datenerhebung lässt sich somit zunächst begrenzen. Sollte sich das Testinstrument nämlich bereits zu diesem Zeitpunkt aufgrund schlechter statistischer Kennwerte als gänzlich ungeeignet erweisen oder zumindest Veränderungen erfahren müssen, muss das Instrument in einer weiteren Pilotierung erneut getestet werden. Zeigt sich das Messinstrument jedoch als prinzipiell geeignet, ist es unabdingbar, es in einem zweiten Schritt dahingehend zu untersuchen, ob sich der Verlauf eines latenten Merkmals an einer konkreten Person, unter Berücksichtigung von Übungseffekten, über die Zeit hinweg modellieren lässt. Hierzu sind verschiedene Modelle entwickelt worden, die auf Strukturgleichungsmodellen (z.B. Latente Wachstumsmodelle) und probabilistischen Modellen (z.B. das Lineare Partial Credit Model, LPCM, das Lineare Rating Scale Model, LRST, oder das (Hybrid) Linear Logistic Model with Relaxed Assumptions (Hybrid-LLRA); vgl. Glück & Spiel 1997) beruhen. Notwendig für diesen zweiten Schritt ist eine Messung zu verschiedenen Zeitpunkten und ggf. an verschiedenen Gruppen mit Hilfe eines aufwändigeren Forschungsdesigns.

Über das vorgestellte Instrument zur Erfassung des globalen Sprachstandes (C-Test) hinaus bedarf es weiterer Prüfungen an ande-

ren Messinstrumenten, wodurch eine kritische Weiterentwicklung unserer angebotenen Verfahrensweise möglich ist. Im Hinblick auf den vorgestellten C-Test steht nun der Nachweis einer praktischen Nützlichkeit zur Messung von Veränderungen des Sprachstandes im Rahmen von Einzelfallstudien und einer Evaluationsuntersuchung aus.

Literatur

Anastasi, A. (1968). Psychological Testing. (3rd ed.). New York: Macmillan.

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

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), Problems in measuring change (pp. 3-20). Madison: University of Wisconsin Press.

Bühner, M. (2006). Einführung in die Test- und Fragebogenkonstruktion (2. Aufl.). München: Pearson.

Bühner, M. (2011). Einführung in die Test- und Fragebogenkonstruktion (3. Aufl.). München: Pearson.

Cleary, T.A. (1968). Testbias: Prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124.

Cronbach, L. J. & Furby, L. (1970). How we should measure „change“: Or should we? Psychological Bulletin, 74, 68-80.

Ewing, M. T., Salzberger, T. & Sinkovics, R. R. (2009). Confirmatory factor analysis vs. Rasch approaches: Differences and Measurement Implications. Rasch Measurement Transactions, 23, 1194-1195.

Fisseni, H.-J. (1997). Lehrbuch der psychologischen Diagnostik. Göttingen: Hogrefe.

Glück, J. & Spiel, C. (1997). Item Response-Modelle für Messwiederholungsdesigns: Anwendung und Grenzen verschiedener Ansätze. Methods of Psychological Research Online, 2, 1. Abgerufen von http://www.dgps.de/fachgruppen/methoden/mpr-online/

Grigorenko, E. L. & Sternberg, R. J. (1998). Dynamic testing. Psychological Bulletin, 124, 75-111.

Grotjahn, R. (1992). Der C-Test im Französischen. Quantitative Analysen. In R. Grotjahn (Hrsg.), Der C-Test. Theoretische und praktische Anwendungen. Band 1 (S. 205-252). Bochum: Universitätsverlag Dr. N. Brockmeyer.

Grotjahn, R. (2002). Konstruktion und Einsatz von C-Tests: Ein Leitfaden für die Praxis. In R. Grotjahn (Hrsg.), Der C-Test. Theoretische und praktische Anwendungen (Vol. 4, S. 211-221). Bochum: AKS-Verlag.

Grotjahn, R., Klein-Braley, C. & Raatz, U. (2002). C-Tests: An overview. In J. Coleman, R. Grotjahn & U. Raatz (Hrsg.), University language testing and the C-Test (S. 93-114). Bochum: AKS-Verlag.

Hattie, J. (1985). Methodology Review: Assessing Unidimensionality of Tests and Items. Applied Psychological Measurement, 9, 139-164.

Jodoin, M. G. & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. Applied Measurement in Education, 14, 329-349.

Klauer, K. J. (1972). Zur Theorie und Praxis des binomialen Modells lernzielorientierter Tests. In K. J. Klauer, R. Fricke, M. Herbig, H. Rupprecht & F. Schott (Hrsg.), Lehrzielorientierte Tests (S. 161-195). Düsseldorf: Schwann.

Klauer, K. J. (2006). Erfassung des Lernfortschritts durch curriculumbasierte Messung. Heilpädagogische Forschung, 32, 16-26.

Linnemann, M. & Wilbert, J. (2010). The C-test: A valid instrument for screening language skills and reading comprehension of children with learning problems? In R. Grotjahn (Hrsg.), Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from Current Research. Frankfurt a.M.: Lang.

Mair, P., Hatzinger, R. & Maier, M. (2010). eRm: Extended Rasch modeling. [Computer software manual]. Available from http://CRAN.R-project.org/package=eRm

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

McKelvey, R. D. & Zavoina, L (1975). A statistical model for the analysis of ordinal dependent variables. Journal of Mathematical Sociology, 4, 103-120.

Miller, R. R. & Grace, R. C. (2003). Conditioning and Learning. In A. F. Healy & R. W. Proctor (Hrsg.), Handbook of Psychology. Vol. 4: Experimental Psychology (S. 357-398). New York: John Wiley & Sons, Inc.

Moosbrugger, H. & Kelava, A. (2008). Testtheorie und Fragebogenkonstruktion. Berlin u.a.: Springer.

Müller, C. & Hartmann, E. (2009). Lernfortschritte im Unterricht erheben – Möglichkeiten und Grenzen des curriculumbasierten Messens. Schweizerische Zeitschrift für Heilpädagogik, 10, 36-42.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing (Version 2.13). Wien: R Foundation for Statistical Computing.

Rost, J. (1985). A latent class model for rating data. Psychometrika, 50, 37-49.

Rost, J. (1999). Was ist aus dem Rasch-Modell geworden? Psychologische Rundschau, 50, 140-156.

Stelzl, I. (1993). Testtheoretische Modelle. In L. Tent (Hrsg.), Pädagogisch-psychologische Diagnostik (Band 1) – Theoretische und methodische Grundlagen (S. 39-201). Göttingen: Hogrefe-Verlag.

Sternberg, R. J. & Grigorenko, E. L. (2002). Dynamic Testing: The Nature and Measurement of Learning Potential. Cambridge, UK: Cambridge University Press.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.

Thorndike, R.L. (1971). Concepts of culture-fairness. Journal of Educational Measurement, 8, 63-70.

von Davier, M. (1997). WINMIRA – Program description and recent enhancements. Methods of Psychological Research Online, 2, 25 - 28.

Walter, J. (2010). Lernfortschrittsdiagnostik am Beispiel der Lesekompetenz (LDL): Messtechnische Grundlagen sowie Befunde über zu erwartende Zuwachsraten während der Grundschulzeit. Heilpädagogische Forschung, 36, 162-176.

Wilbert, J. & Linnemann, M. (in Vorbereitung). Two studies on the validity of c-tests for measuring language comprehension of learning disabled students.

Wottawa, H. & Amelang, M. (1980). Einige Probleme der „Testfairness“ und ihre Implikationen für Hochschulzulassungsverfahren. Diagnostica, 26, 3, 199-121.

Zumbo, B. D. & Thomas, D. R. (1997). A measure of effect size for a model-based approach for studying DIF. Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science. University of Northern British Columbia: Prince George, B.C.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Anschrift des korrespondierenden Autors

Dr. Jürgen Wilbert
Sonderpädagogik und Rehabilitation bei Lernstörungen
Department Heilpädagogik und Rehabilitation
Humanwissenschaftliche Fakultät
Universität zu Köln
Klosterstr. 79b
50931 Köln
juergen.wilbert@uni-koeln.de