

Dorothea Krampen

Zur Bedeutung des Testformats für die Testauswertung

Aufgabenstamm- und
Antwortabhängigkeiten im C-Test



PETER LANG
EDITION

„Es w___ ihm unmö___ die Wör___ nicht i___ dem Bes___ ihrer Bedeu___ zu stö___.“¹
Georg Christoph Lichtenberg (1742–1799) [Tilgungen D. K.]

1 Einleitung

Der Einsatz von Tests bei Schülerinnen und Schülern ist in Deutschland, wie auch in anderen Staaten und Bildungssystemen, weit verbreitet und gehört mittlerweile zum Alltag. Kinder und Jugendliche werden im Verlauf ihrer Schulzeit mit einer Vielzahl an Tests konfrontiert, zum Beispiel in Form von Klassenarbeiten, Klausuren und abschließenden Lernkontrollen; aber auch in Form von sportbezogenen Tests und Wettbewerben oder freiwilligen Tests wie etwa der Mathematik-Olympiade. Es existiert demzufolge eine große Anzahl ganz unterschiedlicher Tests mit unterschiedlichen Zielsetzungen. Gemein ist den Tests jedoch, dass sie als repräsentative Verhaltensstichprobe konzipiert sind, mit der ein interessierendes Merkmal (sein Vorhandensein, seine Qualität, seine Authentizität) erfasst werden soll. Im schulischen Kontext ist das interessierende Merkmal häufig die (fachspezifische) Leistung von Schülerinnen und Schülern.

Von den oben angeführten Tests sind wissenschaftliche Tests aufgrund ihrer engeren Definition sowie ihrer einheitlichen und obligaten Qualitätsvorgaben abzugrenzen. Ein wissenschaftlicher Test ist ein Datenerhebungsverfahren, das gemeinhin aus mehreren Testaufgaben sowie festgelegten Regeln zu dessen Anwendung und Auswertung besteht. An einen wissenschaftlichen Test werden bestimmte Qualitätsanforderungen (z. B. die Erfüllung der klassischen Hauptgütekriterien Objektivität, Reliabilität und Validität; s. Amelang & Schmidt-Atzert, 2006, S. 135 ff.) gestellt. Standards für wissenschaftliches Testen liegen von mehreren internationalen und nationalen Organisationen erarbeitet, mit vergleichbarer Zielsetzung, vor (für einen Überblick s. Moosbrugger & Höfling, 2012). Beispiele dafür sind die Standards for Educational and Psychological Testing der American Educational Research Association (AERA), der American Psychological Association (APA) und dem National Council on Measurement in Education (NCME) (AERA, APA & NCME, 1999) oder die Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen (DIN

1 Vervollständigt lautet der Aphorismus: „Es war ihm unmöglich die Wörter nicht in dem Besitz ihrer Bedeutung zu stören.“

33430; DIN, 2002). Die Standards dienen vor allem der Bewertung und Optimierung der Schritte der Testkonstruktion und der Qualitätsvorgaben der Testanwendung. Ziel eines wissenschaftlichen Tests ist es, ein Merkmal – oftmals ein psychisches Merkmal wie eine Fähigkeit oder ein Persönlichkeitsmerkmal – in seiner absoluten oder relativen Ausprägung in der Regel quantitativ zu erfassen.

Der Einsatz von wissenschaftlichen Tests bei Schülerinnen und Schülern ist in Deutschland mittlerweile ebenfalls verbreitet. Die Tests werden zur Diagnose des Potentials und der Performanz der Kognitionen und der Emotionen von Schülerinnen und Schülern eingesetzt. Sie kommen im Rahmen der Diagnose und Behandlung von Verhaltensproblemen zur Anwendung, um grundlegende Informationen über die besonderen Stärken und Schwächen bestimmter Schülerinnen und Schüler zu gewinnen und therapeutische Strategien zu entwickeln. Die Tests dienen der Findung des geeignetsten Bildungsgangs im Zuge einer schulpsychologischen Laufbahnberatung oder sie unterstützen die Berufsfindung. Eine andere im schulischen Kontext eingesetzte Gruppe wissenschaftlicher Tests, deren Verbreitung in Deutschland in den letzten beiden Jahrzehnten stark zugenommen hat (Stanat, 2008), sind groß angelegte Schulleistungsstudien. Diese repräsentativen internationalen oder nationalen Studien beschränken sich zumeist auf Schülerinnen und Schüler einer bestimmten Jahrgangsstufe oder eines bestimmten Alters und haben das Ziel, Auskunft über den Leistungsstand zu geben. Neben diesem wichtigen Ergebnis von Bildungsprozessen werden auch Prozess- und Kontextfaktoren berücksichtigt, die Aussagen über die Wirkungsweise eines Bildungssystems erlauben. Schulleistungsstudien legen unterschiedliche Perspektiven an: Sie informieren über den Zustand eines Bildungssystems im internationalen oder nationalen Vergleich, sie lassen sich anhand von inhaltlichen Kriterien (z. B. Kompetenzstufen) verorten und geben – bei wiederholten Erhebungsrunden – Auskunft über Veränderungen über die Zeit. Obgleich bereits in den 1960er Jahren erste internationale Schulleistungsstudien initiiert wurden, nahm Deutschland bis in die 1990er Jahre nur mit kleinen Stichproben an wenigen Studien teil (für eine Übersicht s. Radisch, 2008; van Ackeren & Klemm, 2000). Seit seiner Teilnahme an der Trends in International Mathematics and Science Study (TIMSS; Baumert et al., 1997) Mitte der 1990er Jahre beteiligt sich Deutschland regelmäßig an groß angelegten internationalen Schulleistungsstudien. Im Jahr 2000 wurde im Auftrag der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) erstmals das Programme for International Student Assessment (PISA; OECD, 1999) durchgeführt, das seitdem alle drei Jahre international mit deutscher Beteiligung stattfindet. Als Reaktion auf den aufgrund des unterdurchschnittlichen Abschneidens von Schülerinnen und Schülern in Deutschland

sogenannten „PISA-Schock“ im Jahr 2001 (Baumert et al., 2001) führte die KMK flächendeckende Bildungsstandards in verschiedenen Schulfächern und für verschiedene Schulabschlüsse und -übergänge ein. Formuliertes Ziel ist dabei, die Qualität des Schulsystems in Deutschland kontinuierlich zu prüfen und zu entwickeln (Köller, 2010). Deutsch Englisch Schülerleistungen International (DESI) war die erste repräsentative nationale Schulleistungsstudie. Sie diente der Beschreibung der sprachlichen Kompetenzen von Neuntklässlerinnen und Neuntklässlern in den Fächern Deutsch und Englisch (Beck & Klieme, 2007; DESI-Konsortium, 2008). Der Einsatz wissenschaftlicher Tests ist damit auch in empirisch arbeitenden Disziplinen wie der Bildungsforschung sehr üblich und Gegenstand extensiver Forschung.

Der Zusammenhang zwischen dem in einem wissenschaftlichen Test gezeigten Verhalten und dem interessierenden Merkmal wird allgemein von einer Testtheorie beschrieben. Eine solche Testtheorie, die sich vor allem in den letzten zwei Jahrzehnten bei der Konstruktion und Evaluation wissenschaftlicher Tests als nützlich erwiesen hat (Hambleton, Swaminathan & Rogers, 1991), ist die Item-Response-Theorie (IRT; z. B. van der Linden & Hambleton, 1997). Grundidee dieser Testtheorie ist, dass die in einem Test beobachtbaren Antworten auf eine nicht beobachtbare (latente) Personenfähigkeit zurückführbar sind, wobei die individuelle Personenfähigkeit und das beobachtete Antwortverhalten in einem probabilistischen Zusammenhang stehen. Auch die Auswertung von Schulleistungsstudien erfolgt heute mit Modellen der IRT, da diese am besten geeignet sind, präzise Aussagen über die Ergebnisse für Personengruppen zu machen. Eine weitere Gemeinsamkeit der Schulleistungsstudien – neben der routinemäßigen Auswertung mittels Modellen der IRT – ist, dass sie häufig ein spezifisches Testformat verwenden. Bei den sogenannten *Testlets* (Wainer & Kiely, 1987) beziehen sich mehrere Testaufgaben eines Tests auf einen gemeinsamen Aufgabenstamm; es folgen beispielsweise in einem Leseverstehenstest auf einen Lesetext mehrere Fragen zu diesem Text. Testlets stellen eine ernstzunehmende Herausforderung für Modelle der IRT dar, da sie die Grundannahme der lokalen stochastischen Unabhängigkeit (vgl. Abschnitt 2.1.1.1) verletzen können, von der diese Modelle in der Regel ausgehen (Henning, 1989; Lord & Novick, 1968). Die Annahme der lokalen stochastischen Unabhängigkeit besagt, dass Antworten auf verschiedene Testaufgaben unter Kontrolle der gemessenen Fähigkeit voneinander unabhängig sind und damit die Wahrscheinlichkeit der Lösung einer Testaufgabe nur von der gemessenen Fähigkeit und nicht von den Antworten auf andere Testaufgaben abhängig sein darf. Das Testletformat kann nun dazu führen, dass Abhängigkeiten zwischen Antworten auf Testaufgaben innerhalb eines Testlets, sogenannte lokale Abhängigkeiten, resultieren. Die

Nichtberücksichtigung vorhandener lokaler Abhängigkeiten bei der Testauswertung kann unter anderem zu verzerrten Itemparametern und zu einer Überschätzung der Messgenauigkeit führen (z. B. Sireci, Thissen & Wainer, 1991; Smith, 2005; Yen & Fitzpatrick, 2006), was letztlich auch falsche Schlussfolgerungen nach sich ziehen kann.

Hinsichtlich der Formen lokaler Abhängigkeiten innerhalb von Testlets kann zwischen auf den gemeinsamen Aufgabenstamm zurückgehenden Abhängigkeiten und spezifischen Effekten des Lösens einzelner Testaufgaben auf das Lösen von darauf folgenden Testaufgaben unterschieden werden (Marais & Andrich, 2008a, b). Im ersten Fall führt der Bezug auf einen gemeinsamen Aufgabenstamm zu Zusammenhängen zwischen Testaufgaben innerhalb desselben Testlets, die nicht durch das im Gesamttest gemessene Merkmal erklärt werden können. Solche *Aufgabenstammabhängigkeiten* können durch zusätzliche Personenvariablen modelliert werden, die inhaltlich etwa aufgabenstammspezifisches Vorwissen oder interindividuelle Unterschiede im Gesamtverständnis des Stamms repräsentieren. Im zweiten Fall ergeben sich Zusammenhänge zwischen Testaufgaben innerhalb desselben Testlets dadurch, dass das Lösen einer Testaufgabe das Lösen der darauf folgenden Testaufgabe erleichtert (bzw. vice versa). Derartige *Antwortabhängigkeiten* können entstehen, wenn etwa eine Information, nach der in einer Testaufgabe gefragt wird, auch für die Lösung der nächsten Testaufgabe nützlich ist. Es ist darüber hinaus denkbar, dass beide Formen lokaler Abhängigkeiten innerhalb eines Testlets simultan vorliegen.

Das Vorhandensein lokaler Abhängigkeiten in testletbasierten Tests konnte bereits an mehreren Stellen empirisch aufgezeigt werden (z. B. He, Li, Wolfe & Mao, 2012; Lee, 2004; Yen, 1993; Zenisky, Hambleton & Sireci, 2002), konkret auch in Schulleistungsstudien (z. B. Harsch & Hartig, 2010; Monseur, Baye, Lafontaine & Quittre, 2011). Im Rahmen der oben bereits erwähnten nationalen Schulleistungsstudie DESI (Beck & Klieme, 2007; DESI-Konsortium, 2008) wurde beispielsweise der lückentextähnliche C-Test (vgl. Abschnitt 2.3) zur Erfassung der allgemeinen sprachlichen Kompetenz in Englisch als Fremdsprache eingesetzt. Es wurden mehrere Kurztexte zu verschiedenen Inhaltsbereichen mit jeweils einer festen Anzahl Lücken verwendet, die es zu vervollständigen galt. Der C-Test weist damit ein charakteristisches Testletformat auf: Mehrere Testaufgaben (Lücken) beziehen sich auf einen gemeinsamen Aufgabenstamm (Text), was bekanntermaßen zu lokalen Abhängigkeiten zwischen den Lücken führen kann. Tatsächlich konnten Harsch und Hartig (2010) sowohl Aufgabenstamm- als auch Antwortabhängigkeiten im C-Test der DESI-Studie aufzeigen.

In der Zusammenschau lässt sich festhalten, dass gerade im Bildungsbereich häufig testletbasierte Tests eingesetzt werden, zu deren Auswertung

routinemäßig Modelle der IRT herangezogen werden. Das Testletformat kann zu Abhängigkeiten zwischen Antworten auf Testaufgaben und damit zu einer Verletzung der IRT-Grundannahme der lokalen stochastischen Unabhängigkeit führen. Zwei Formen lokaler Abhängigkeiten in testletbasierten Tests können dabei unterschieden werden: Aufgabenstammabhängigkeiten und Antwortabhängigkeiten (Marais & Andrich, 2008a, b). Werden vorhandene Abhängigkeiten – gleich welcher Form – bei der Testauswertung ignoriert, kann dies zu verzerrten Ergebnissen führen. Die vorliegende Arbeit setzt an dieser Problematik an und baut auf den Analysen von Harsch und Hartig (2010) auf, indem sowohl die Unterscheidung von Aufgabenstamm- und Antwortabhängigkeiten maßgebend ist als auch der testletbasierte C-Test der Schulleistungsstudie DESI als verwendete Untersuchungs- und Datengrundlage der von Harsch und Hartig (2010) analog ist. Der Fokus der vorliegenden Arbeit liegt speziell auf der empirischen Unterscheidbarkeit und angemessenen psychometrischen Modellierung der beiden Formen lokaler Abhängigkeiten. Dazu wird ein neuartiger Ansatz zur Modellierung von Aufgabenstamm- und Antwortabhängigkeiten entwickelt, erprobt und diskutiert.

Die Arbeit stellt zunächst die theoretischen Grundlagen in einer Literaturübersicht bereit, die für ein Verständnis des neuen Ansatzes zur Modellierung der unterschiedlichen Formen lokaler Abhängigkeiten sowie der Zielsetzung wichtig sind, und fasst den aktuellen Forschungsstand zusammen (Abschnitt 2). Der eigene Modellierungsansatz wird daran anknüpfend ausführlich dargestellt (Abschnitt 3), bevor die konkreten Fragestellungen ausformuliert werden (Abschnitt 4). Die Untersuchungs- und Datengrundlage sowie das detaillierte Vorgehen bei den durchgeführten Analysen werden im Methodenteil erläutert (Abschnitt 5). Den Ergebnissen aus deskriptiven Analysen und Modellvergleichen (Abschnitt 6) folgt eine Zusammenfassung der zentralen Ergebnisse der Arbeit mit Bezug auf die formulierten Fragestellungen sowie eine kritische Diskussion der Ergebnisse unter Berücksichtigung der theoretischen Annahmen und des bisherigen empirischen Forschungsstands mit Blick auf die (theoretische und praktische) Relevanz und mögliche zukünftige Forschungsfragen (Abschnitt 7).