R&E-SOURCE https://journal.ph-noe.ac.at
Online Journal for Research and Education
Ausgabe 15, April 2021, ISSN: 2313-1640

Empirische Bildungsforschung mittels kognitiver Diagnosemodelle

Ziele, State of the Art & Entwicklungsperspektiven

Ann Cathrice George[1]

Zusammenfassung

Der vorliegende Artikel gibt einen Überblick über Studien in der empirischen Bildungsforschung, welche das Instrument der kognitiv-diagnostischen Modelle einsetzt (Cognitive Diagnosis Models; CDMs). Dazu wird zunächst motiviert, bei welchen Fragestellungen CDMs zum Einsatz kommen können und welche Ziele durch die Modelle erreicht werden können. Dann erfolgt ein kurzer Abriss der wesentlichen aktuellen Forschungsrichtungen unter Einsatz der Modelle. Schließlich werden Entwicklungsperspektiven für bildungswissenschaftliche Forschung unter dem Einsatz von CDMs aufgezeigt.

Schlüsselwörter:	*Keywords:*
Empirische Bildungsforschung	Empirical Educational Research
Kognitive Diagnosemodelle (CDMs)	Cognitive Diagnosis Models
Diagnostik von Subkompetenzen	Diagnosis of Sub-Competencies

1 Motivation, Ziele und Ergebnisse

Bei dem sogenannten „PISA-Schock“ Anfang der 2000er Jahre zeigte sich anhand von Ergebnissen aus der international vergleichenden Schulleistungsstudie PISA (Programme for International Student Assessment; OECD, 2019), dass das Abschneiden deutschsprachiger Schüler/innen im internationalen Vergleich nicht den (hohen) Erwartungen der jeweiligen Bildungssysteme entsprach. So zeigten sich in Large-Scale-Studien wie u. a. PISA zunächst Defizite der Schüler/innen in Deutschland und der Schweiz (PISA 2000) und wenig später auch in Österreich (PISA 2003). Als eine Konsequenz aus dem PISA-Schock entstand eine neue Form von Steuerungsorientierung in der Bildungspolitik, die sogenannte Outputorientierung (Fend, 2014). Bei dieser Steuerung liegt das Hauptaugenmerk auf dem schulischen Output, also den Leistungen der Schüler/innen. Neben den Ergebnissen nationaler und internationaler Schulleistungsmessungen nahmen auch aus ihnen abgeleitete bildungspolitische Maßnahmen einen höheren Stellenwert ein.

Mit dem zunehmenden Interesse an der Messung von Schülerkompetenzen stieg auch die Forschung zu statistisch-methodischen Grundlagen für die Analyse von Antwortdaten aus Large-Scale-Studien. Zur Auswertung und anschließenden Berichterstattung der Schülerantworten werden in Large-Scale-Studien hauptsächlich (eindimensionale) Item-Response-Modelle (IRT-Modelle; Baker & Kim, 2004) eingesetzt (Martin, Mullis & Hooper, 2017; OECD, 2017). Bei der Berichterstattung der Schülerkompetenzen werden Mittelwerte und Standardabweichungen für einzelne Kompetenzen wie beispielsweise „Mathematik“, „Lesen“ oder „Naturwissenschaften“ auf einer stetigen Skala geschätzt. Zusätzlich werden einige Teilkompetenzen berichtet. So werden beispielsweise für die Kompetenz Mathematik in TIMSS zusätzlich Werte in den (wiederum eindimensional gesehenen) Teilkompetenzen „Zahlen“, „Geometrische Formen und Maße“, „Darstellen von Daten“, „Wissen“, „Anwenden“ und „Begründen“ berichtet (Mullis & Martin, 2017). Um die empirischen Testwerte auf den eindimensionalen Skalen (z. B. „548“) mit fachdidaktischer Bedeutung zu füllen, werden mithilfe von Standard-Setting-Methoden Post-hoc-Kompetenzstufen für die Skalen bestimmt (Cizek, Bunch & Konns, 2004). Dabei wird die stetige 500er-Skala in Stufen unterteilt, welche dann durch eine Verallgemeinerung

[1] Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen, Alpenstraße 121, 5020 Salzburg.
E-Mail: anncathrice.george@iqs.gv.at

der auf diesen Stufen liegenden Itemanforderungen eine Bedeutung bekommen. Die Beschreibung wird zusammen mit dem Bereich als Kompetenzstufe bezeichnet. Laut TIMSS können Schüler/innen auf Kompetenzstufe II der Mathematikskala „ganze Zahlen und Brüche verstehen, dreidimensionale Formen anhand von zweidimensionalen Abbildungen visualisieren und Informationen in Balkendiagrammen, Piktogrammen und Tabellen interpretieren, um einfache Aufgaben zu lösen“ (für die Übersetzung siehe Suchan, Wallner-Paschon, Bergmüller & Schreiner, 2012). Die gemessenen Kompetenzen und Kompetenzstufen helfen, auf Bildungssystemebene bundesweite Ausbildungs- oder Fördermaßnahmen für Schülergruppen (z. B. Buben im Lesen) oder Lehrpersonen (z. B. Vertiefung der Lehrmethoden im Bereich Statistik) zu entwerfen.

Neben der Ableitung von Maßnahmen auf der Bildungssystemebene existieren auch andere Ziele empirischer Kompetenzmessungen: So dienten beispielsweise die zwischen 2012 und 2019 in Österreich durchgeführten Bildungsstandardüberprüfungen neben der Systementwicklung auch der Schul- und Unterrichtsentwicklung (Schreiner & Wiesner, 2019). Ein noch stärker formatives Konzept verfolgen die Vergleichsarbeiten (VERA) in Deutschland (Helmke und Hosenfeld, 2003) oder die Informelle Kompetenzmessung (IKM) in Österreich. Diese Art von Assessments richtet sich an Schulen, an Lehrpersonen mit ihren Klassen oder sogar an individuelle Schüler/innen. Damit steigt allerdings das Verlangen nach tiefer gehenden diagnostischen Informationen, die über die beschriebene, weitestgehend deskriptive Darstellung aus eindimensionalen IRT-Modellen hinausgehen (de la Torre & Minchen, 2014). Vielleicht aufgrund der bereits ausgereiften statistischen Methodik und/oder zur Gewährleistung der Vergleichbarkeit zwischen den Ebenen des Schulsystems (vgl. George, Robitzsch & Schreiner, 2019) wird allerdings auch in Assessments mit formativem Charakter zumeist das statistische Verfahren aus Large-Scale-Studien übernommen. Allerdings zeigen Lehrkräfte vermehrt Schwierigkeiten bei der Ableitung individueller oder gruppenspezifischer Fördermaßnahmen aus den in dieser Form entwickelten Rückmeldungen formativer Assessments (Bögeholz & Eggert, 2013).

Um durch detaillierte Rückmeldungen aus formativen Assessments deren Nutzbarkeit für die Schul- und Unterrichtspraxis zu erhöhen, helfen aus methodischer Sicht mehrdimensionale psychometrische Modelle. Als mehrdimensionale psychometrische Modelle eignen sich beispielsweise mehrdimensionale IRT-Modelle (vgl. Reckase, 2009) oder kognitive Diagnosemodelle (CDMs, Cognitive Diagnosis Models; DiBello, Roussos & Stout, 2007; Rupp, Templin & Henson, 2010). Letzteren Modellen wird besonders aufgrund von zwei Merkmalen ein großes Potenzial zugeschrieben: Erstens basieren die empirischen CDMs auf theoretisch-fachdidaktisch motivierten Kompetenzmodellen. Die sogenannte Q-Matrix (Tatsuoka, 1984) erlaubt, die (Test-)Items mit Teilkompetenzen (im Folgenden „Skills“ genannt) zu verbinden, die direkt aus einem theoretischen Kompetenzmodell übernommen werden. Somit kann das empirische CDM die theoretische mehrdimensionale Kompetenzstruktur widerspiegeln. Zweitens liefern CDM-Analysen sehr direkt interpretierbare Ergebnisse: (1) den Prozentsatz der Schüler/innen, die die vorab definierten Skills beherrschen (sogenannte „Skillverteilung“), (2) den Prozentsatz der Schüler/innen, die bestimmte Kombinationen von Skills beherrschen (sogenannte „Skillklassenverteilung“), und (3) jene Skills, die eine individuelle Schülerin bzw. ein individueller Schüler beherrscht („individuelles Skillmuster“). Alle drei Ergebnisse sind durch die Vorabdefinition der Skills direkt mit dem mehrdimensionalen Kompetenzmodell verbunden. Weitere lernpsychologische Annahmen, beispielsweise über Annahmen des Zusammenwirkens der Skills zur Lösung der Items, lassen sich durch die Wahl eines spezifischen CDMs einbringen. Ebenso wie bei IRT-Modellen handelt es sich bei kognitiven Diagnosemodellen um eine Klasse von Modellen. Diese besteht aus spezifischen Modelltypen wie etwa dem DINA (Haertel, 1989) oder dem RUM (Hartz, 2002) und verallgemeinerten Ansätzen wie dem G-DINA (de la Torre, 2011), dem LCDM (Henson, Templin & Willse, 2009) oder dem GDM (von Davier, 2008).

2 Ein Beispiel zu illustrativen Zwecken

Ein populäres Beispiel bei der Anwendung von CDMs ist der sogenannte Fraction-Subtraction-Test, in dem es dem Namen nach um das Subtrahieren von Brüchen geht. Der Test wurde ursprünglich von Tatsuoka (1984) eingesetzt. Drei der insgesamt 20 Items aus dem Test lauten wie folgt

(a) $\frac{5}{3} - \frac{3}{4}$ (b) $\frac{2}{3} - \frac{2}{3}$ (c) $4 - 1\frac{4}{3}$.

Expertinnen und Experten aus der Mathematik leiteten 8 Skills ab, die der Kompetenz des Subtrahierens von Brüchen zugrunde liegen. Es wird davon ausgegangen, dass Schüler/innen, wenn sie alle diese 8 Skills beherrschen, alle Items des Tests lösen können (abgesehen von Slipping-Fehlern). Die Skills lauten:

α_1 eine ganze Zahl in einen Bruch umwandeln

α_2 einem Bruch eine ganze Zahl entnehmen
α_3 vor dem Subtrahieren vereinfachen
α_4 einen gemeinsamen Nenner finden
α_5 einen ganzzahligen Anteil borgen
α_6 einen Zehner beim Subtrahieren der Zähler übertragen
α_7 Zähler subtrahieren
α_8 Antwort vereinfachen

Jedem Item werden dann die Skills zugeordnet, die zu seiner Lösung notwendig sind, und die Zuordnung wird in Form einer Q-Matrix dargestellt. Aufgrund der insgesamt 8 definierten Skills erhält die Q-Matrix 8 Spalten. Die erste Zeile der Q-Matrix gibt die Skills an, welche Schüler/innen zur Lösung des ersten Items (a) benötigen, nämlich die Skills α_4, α_6 und α_7. Die anderen Skills werden für Item (a) nicht benötigt. Jeder zur Lösung von Item (a) notwendige Skill wird in der ersten Zeile der Q-Matrix mit „1“ zugewiesen und die nicht notwendigen Skills werden mit „0“ abgelehnt. Die erste Zeile der Q-Matrix lautet also $q_1 = [0,0,0,1,0,1,1,0]$. Ebenso läuft der Prozess für die restlichen Items.

Die Antwortdaten der Schüler/innen werden zusammen mit der Q-Matrix und einem Modell aus der Klasse der CDMs ausgewertet. Die drei Hauptergebnisse für den Bruchrechentest mit 536 Schülerinnen und Schülern, der auch in Studien von de la Torre und Douglas (2004) sowie De Carlo (2011) eingesetzt wurde, sind unter Einsatz des DINA-Modells in Abbildung 1 gegeben. Im oberen Teil der Abbildung kann der Prozentsatz der Schüler/innen abgelesen werden, die die 8 Skills beherrschen (bzw. die Wahrscheinlichkeit der Schüler/innen, die 8 Skills zu beherrschen). Jeweils über 80 % der Schüler/innen beherrschen α_7 „Zähler subtrahieren“ und α_8 „Antwort vereinfachen“, welche somit die einfachsten Skills sind. Dagegen beherrschen unter 60 % der Schüler/innen α_1 „eine ganze Zahl in einen Bruch umwandeln“ und α_5 „einen ganzzahligen Anteil borgen“. In der Mitte von Abbildung 1 kann die Verteilung der Schüler/innen auf die Skillklassen entnommen werden. Neben der Skillklasse 11111111, in der die Schüler/innen alle Skills beherrschen, zeigen sich größere Gruppierungen von Schülerinnen und Schülern in den beschrifteten Skillklassen wie beispielsweise 01000011. In der letztgenannten Skillklasse beherrschen die Schüler/innen die Skills α_1, α_7 und α_8, den Rest der Skills hingegen nicht. Im unteren Teil von Abbildung 1 ist die Wahrscheinlichkeit einer individuellen Schülerin/eines individuellen Schülers gezeigt, die 8 Skills zu beherrschen. Die Schülerin/der Schüler ist durch ihr/sein Antwortmuster zu den Items im Test charakterisiert. Diese Schülerin/dieser Schüler beherrscht alle Skills mit Ausnahme von α_4 „einen gemeinsamen Nenner finden“.

3 Forschungsrichtungen

Seit den Anfängen der Modellklasse von CDMs in den 2000er Jahren (siehe z. B. Junker & Sijtsma, 2001) werden die Modelle mittlerweile vermehrt in der Forschung eingesetzt. Die Forschung erfolgte zunächst schwerpunktmäßig in der statistisch-methodischen Richtung mit Fokus auf Modellentwicklung. Erst später wurden die Modelle auch für Anwendungsstudien genutzt.

3.1 Statistisch-methodische Studien

In der statistisch-methodischen Richtung liegt ein Schwerpunkt der Forschung auf der Entwicklung von spezifischen Modellen wie beispielsweise dem bekannten DINA und zeitlich etwas später der Generalisierung der spezifischen Modelle in verallgemeinerte Modellansätze wie dem G-DINA. So werden beim DINA-Modell nur zwei Antwortwahrscheinlichkeiten pro Item modelliert: Besitzen Schüler/innen alle dem Item in der Q-Matrix zugewiesenen Skills, haben sie eine Wahrscheinlichkeit von 1 minus „Slipping-Fehler“, das Item zu lösen, d. h. eine Wahrscheinlichkeit, das Item trotz Kenntnis der Skills versehentlich nicht zu lösen. Besitzen die Schüler/innen mindestens einen der zugewiesenen Skills nicht, so sinkt die Wahrscheinlichkeit, das Item zu lösen, auf eine Ratewahrscheinlichkeit (sog. „guessing“). Dabei handelt es sich um einen nichtkompensatorischen Ansatz: Das Fehlen eines relevanten Skills kann nicht durch andere Skills ausgeglichen werden. Ein umfangreicher Vergleich vieler spezifischer CDMs findet sich beispielsweise in DiBello et al. (2006). Bei der Verallgemeinerung im G-DINA-Modell trägt jeder dem Item zugewiesene Skill zu einer Veränderung der Antwortwahrscheinlichkeit bei. Auch Wechselwirkungen zwischen den zugewiesenen Skills können die Antwortwahrscheinlichkeit beeinflussen. Aus den allgemeinen Modellansätzen können durch bestimmte Parameterdefinitionen fast alle spezifischen Modelle abgeleitet werden (bspw. indem alle Wechselwirkungen zwischen den Skills auf null gesetzt

werden). Neben dem G-DINA existieren weiterhin die verallgemeinerten Modellansätze des GDM und des LCDM. In einem weiteren Schritt der Verallgemeinerung können CDMs auch über den Ansatz der strukturierten latenten Klassenmodelle erklärt werden (Robitzsch und George, 2019).

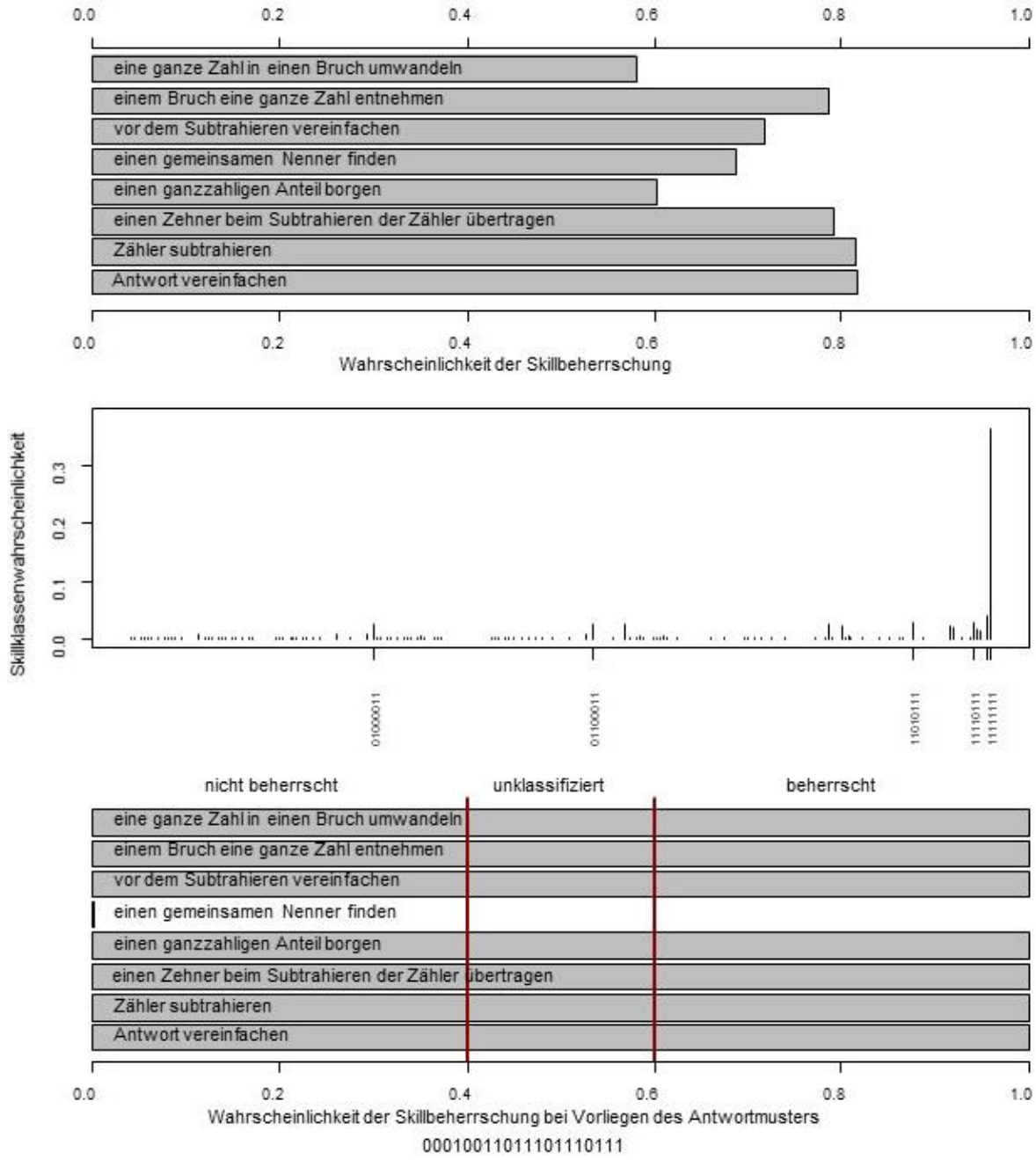

Abbildung 1: Darstellung von den 3 Hauptergebnissen aus einer CDM-Analyse am Beispiel Subtrahieren von Brüchen: Prozentuale Beherrschung der Skills in der Population (oben), Verteilung der Population in den Skillklassen (Mitte) und Wahrscheinlichkeit, die Skills zu beherrschen, für eine Schülerin/einen Schüler entsprechend ihrer/seiner Antwort auf die Items (unten; Grafik erstellt mit dem R-Paket „CDM"; Robitzsch, Kiefer, George & Ünlü, 2020).

Neben der Definition neuer Modellansätze wurden bestehende Modelle auf neue Anforderungen ausgeweitet. Beispielsweise entsteht aus der Verallgemeinerung von spezifischen Modellen auch die Möglichkeit, verschiedene spezifische Modelle für verschiedene Testitems zu definieren (Rupp et al., 2009). Auch sind inzwischen Ausweitungen der zunächst auf dichotom kodierte Items und dichotome Skills spezialisierten

Modelle auf polytome Antwortdaten und Skills (Chen & de la Torre, 2013; von Davier, 2008) entwickelt worden. Ein anderer komplexer Modelltyp, das sogenannte Multiple-Strategies DINA Model (Chen & de la Torre, 2013; Huo & de la Torre, 2014), bietet die Modellierung verschiedener Lösungsstrategien pro Item an.

Mit der Anzahl unterschiedlicher Modelle nahm auch die Notwendigkeit zu, Software zu entwickeln, mit deren Hilfe die Parameter der Modelle geschätzt werden konnten. Auch hier lässt sich eine Tendenz von Einzelsoftware für spezifische Modelle hin zu Softwarepaketen mit vielen verallgemeinerten Modelltypen beobachten. So startete die Entwicklung der Software mit einem Skript von de la Torre in der Matrixumgebung Ox (de la Torre, 2009), der kostenpflichtigen Arpeggio-Software zur Schätzung des RUM (DiBello und Stout, 2008), der Einzelsoftware mdltm zur Schätzung des GDM (von Davier, 2008) und etwas später der Bereitstellung einer Umgebung in Mplus (Templin & Hoffman, 2013), mit deren Hilfe Modelle aus dem LCDM-Framework geschätzt werden konnten. Aktuell bieten sich einige Pakete in der kostenfreien Programmierumgebung R an sowie das CDM- und das GDINA-Paket (Ma & de la Torre, 2020), in denen neben Modellschätzungen auch statistische Fitkriterien oder grafische Plot-Funktionen genutzt werden können. Einen ersten Vergleich unterschiedlicher Softwarepakete bieten Rupp und van der Rijn (2018) oder von Davier und Lee (2019). Neben R bietet sich auch der bayesianische Ansatz über JAGS an (Zhan, Jiao, Man & Wang, 2019).

Das dritte große Thema in der statistisch-methodischen Forschungsrichtung sind empirisch gesteuerte Ansätze zur Q-Matrix-Generierung. Ursprünglich wird in CDM-Modellen davon ausgegangen, dass die Einträge der Q-Matrix durch Expertinnen und Experten der jeweiligen Fachdomäne entwickelt werden. Es liegen einige Ansätze vor, wie Q-Matrizen auch in Teams von Expertinnen und Experten entwickelt und diskutiert werden können (Akbay, Terzi, Kaplan & Karaaslan, 2018; Bley, 2017). In den entwickelten empirischen Ansätzen können solche durch Expertinnen und Experten generierten Q-Matrizen entweder mithilfe der Antwortdaten verfeinert werden (de la Torre & Chiu, 2016; Wang et al., 2018) oder sogar rein datengetrieben ohne fachliche Expertinnen und Experten entwickelt werden (Wang, Song & Ding, 2018; Wang et al., 2020).

3.2 Anwendungsstudien

Bei den meisten der bestehenden Anwendungsstudien handelt es sich um sogenannte Retrofit-Studien (für ein Review von Anwendungsstudien siehe auch Sessoms & Henson, 2018). In diesen Studien werden Items und deren Schülerantworten aus bestehenden (häufig Large-Scale-)Settings mithilfe von CDMs neu ausgewertet, um diagnostische Informationen zu erhalten. Dabei wird durch Fachexpertinnen/-experten zunächst eine Liste von Skills erstellt, die der im Test überprüften Kompetenz zugrunde liegen. Diese Liste dieser Skills kann aus verschiedenen Quellen abgeleitet werden, wie beispielsweise Fachliteratur, Kompetenzmodelle, Abschlussarbeiten oder auch Think-Aloud-Protokolle von Schülerinnen und Schülern (z. B. Bley, 2017). Diese Skills werden dann (in häufig mehreren Diskussionsrunden) den Items zugeordnet und durch diese Zuordnungen wird die Q-Matrix gebildet. García, Olea und de la Torre (2014) nutzen die sogenannte Delphi-Methode über drei Runden: In der ersten Runde definiert jede Expertin/jeder Experte eine Q-Matrix, in der zweiten Runde erhält jede Expertin/jeder Experte die Q-Matrizen der anderen Expertinnen/Experten und kann in ihrer/seiner eigenen Q-Matrix Änderungen vornehmen und in der dritten Runde werden die noch bestehenden Differenzen in den Q-Matrizen zwischen den Expertinnen und Experten diskutiert. In einigen Ansätzen werden zur Bildung der Q-Matrix auch verschiedene Lösungsstrategien einbezogen, so untersucht Roberts et al. (2014) mithilfe von Think-Aloud-Studien, wie Lehrpersonen im Vergleich zu Schülerinnen und Schülern bei der Lösung von Items vorgehen. In einigen Studien wird nach der Definition der Q-Matrix durch die Expertinnen und Experten eine empirische Validierung angeschlossen, in der mittels verschiedener Verfahren die Q-Matrix leicht abgeändert werden kann (z. B. Kim, 2015; Li & Suen, 2013).

Die Domänen, in denen CDM-basierte Studien eingesetzt werden, lassen sich zuvorderst auf Mathematik (Akbay et al., 2018) und Lesen in der Muttersprache (Chen & Chen, 2016) sowie in der ersten Fremdsprache (Jang, 2009; Kim, 2015) eingrenzen. Eine Erweiterung bieten Studien, die naturwissenschaftliche Kompetenzen (Kabiri, Ghazi-Tabatabaei, Bazargan, Shokoohi-Yekta & Kharrazi, 2016) und Kompetenzen im Hören (Aryadoust, 2018) untersuchen. Außerhalb dieser Schwerpunktgebiete gibt es vereinzelt Studien mit speziellen Anwendungen: die Bewertung verschiedener Situationen im beruflichen Kontext (Garcia, Olea & de la Torre, 2014), die Diagnose psychischer Störungen am Beispiel der Glücksspielsucht (Templin & Henson, 2006) und die Analyse komplexer professioneller Kompetenzen am Beispiel Intrapreneurship (Abele & von Davier, 2019; George, Bley & Pellegrino, 2019). Je komplexer die Kompetenzen werden, desto schwieriger erscheint es, für die Kompetenzen Skills zu definieren und diese wiederum in Items abzubilden.

In vielen der vorliegenden Anwendungsstudien werden unterschiedliche CDMs an die Daten angepasst und anhand statistischer Fitkriterien wird ein Modell ausgewählt, welches die Daten am besten beschreibt.

Beispielsweise wählen Ravand & Robitzsch (2018) im Lesen aus 6 Modellen, darunter G-DINA und DINA, oder Aryadoust (2018) in Mathematik aus 5 Modellen anhand von relativen und absoluten statistischen Fitkriterien. Anschließend wird in einigen der Studien mithilfe des gewählten Modells ein Feedback auf Ebene der Testpopulation erstellt. Es wird also beschrieben, wie viel Prozent der Schüler/innen jeweils die vorab definierten Skills beherrschen (vgl. die exemplarische Darstellung in Abbildung 1 oben und Mitte). In selteneren Fällen wird ein Individualfeedback erzeugt (Jang, 2009) oder sogar validiert, inwiefern Schüler/innen ein solches Individualfeedback als hilfreich empfinden (Kim, 2015).

4 Entwicklungsperspektiven

Ursprünglich wurden CDMs entwickelt, um mehr diagnostische Informationen aus Assessments zu generieren. Diesem Ziel wurde bisher hauptsächlich nachgegangen, indem CDMs als Retrofit von z. B. Large-Scale-Studien eingesetzt wurden, um auf Ebene der Testpopulation Skills zu analysieren, die der getesteten Kompetenz zugrunde liegen. Im Fokus dieser Analysen stand bisher aber weniger, wie die erlangten Ergebnisse interpretiert werden könnten und was daraus für die Bildungspolitik, Lehrerfortbildung oder Schulpraxis abzuleiten wäre. Stattdessen stand mehr die methodisch-statistische Verfeinerung der Modelle im Vordergrund. Diese lässt nun komplexe Modellierungen der Antwortdaten zu. Unter der Komplexität kann aber möglicherweise die Interpretation leiden: Wird beispielsweise bei jedem Item ein anderes spezifisches Modell eingesetzt, so liegen jedem Item andere Regeln bei der Lösung zugrunde. D. h. beispielsweise, dass bei einer Teilmenge von Items angenommen werden kann, dass alle zugewiesenen Skills zur Lösung beherrscht werden müssen und bei einer anderen Teilmenge von Items ein Fehlen von zugewiesenen Skills kompensiert werden kann. Dies ist möglicherweise aus theoretisch-fachdidaktischer Sicht schwer erklärbar. In bisherigen Studien wird allerdings der häufig bessere statistische Fit komplexerer Modelle den möglichen Auswirkungen auf die Interpretation vorgezogen (für eine Diskussion siehe auch George et al., 2019). Aus der rein statistisch-methodischen Richtung scheint allerdings noch die Entwicklung einiger Werkzeuge für den Umgang mit CDM-basierten Large-Scale-Studien zu fehlen, wie beispielsweise der Einsatz von Plausible Values oder der Umgang mit Längsschnitt- und Trendanalysen.

Neben der Analyse von Skills auf Ebene der Testpopulation könnte eine weitere, bisher wenig eingesetzte Richtung auch sein, konfirmatorische Ansätze von CDMs zur Validierung fachdidaktischer Annahmen über die untersuchten Skills zu nutzen. So könnten theoretische Annahmen über die Struktur zwischen den Skills (z. B. hierarchisch vs. nichthierarchisch), die Anzahl der Skills und die Art des Zusammenwirkens (z. B. kompensatorisch vs. nichtkompensatorisch) jeweils konfirmatorisch modelliert und die Modelle unter Einsatz der Antwortdaten statistisch verglichen bzw. evaluiert werden (siehe auch George und Robitzsch, 2018). Basierend auf den ausgewählten (theoretischen und empirischen) Modellen könnten des Weiteren Ableitungen über Lernpfade erstellt werden (Wu, Wu, Chang, Kong & Zhang, 2020).

Ein großes, bisher weitgehend ungenutztes Potenzial von CDMs könnte in deren Einsatz für die Individualdiagnose und zur Unterstützung von Lehrpersonen bei der Diagnose von Schülerkompetenzen liegen. Erste Schritte für individuelle Feedbacksysteme, deren Ergebnisse auf CDMs beruhen, finden sich in Rupp et al. (Kapitel 3) und im Speziellen bei Kim (2015). Die Nutzung vieler psychometrischer Modelle, so auch CDMs, für den Unterricht gestaltet sich wegen der kleinen Stichproben herausfordernd. Um dieses Problem zu umgehen, beschreiben You, Li, Zhang & Liu (2018) den Aufbau eines Lernsystems, in dem Items und die zugehörigen Modellparameter zentral verwaltet werden. Somit kann eine Schätzung der Skills innerhalb der Klasse und für jede einzelne Schülerin/jeden einzelnen Schüler vorgenommen werden.

Ein weiterer, bisher in der Literatur wenig beachteter Aspekt scheint die Entwicklung diagnostischer Items zu sein. Häufig werden derzeit Retrofitanalysen gerechnet, die dementsprechend auf Items basieren, die für den Einsatz eindimensionaler Modelle ausgewählt wurden. Wie Items allerdings konstruiert werden müssten, wenn sie für diagnostische Zwecke genutzt werden, scheint eher implizites Wissen zu sein. Einige grundlegende Erfahrungen beschreibt Nagele (2020) in seiner Dissertation.

Literatur

Abele, S. & von Davier, M. (2019). CDMs in vocational education: Assessment and usage of diagnostic problem-solving strategies in car mechatronics. In M. von Davier & Y.-S. Lee (Hrsg.), *Handbook of Diagnostic Classification Models – State of the Art in Modeling, Estimation, and Applications* (S. 462–488). New York: Springer.

Akbay, L., Terzi, R., Kaplan, M. & Gizem Karaaslan, K. (2018). Expert-based attribute identification and validation on fraction subtraction: A cognitively diagnostic assessment application. *Journal on Mathematics Education*, 9(1), 103–120.

Aryadoust, V. (2018). A cognitive diagnostic assessment study of the listening test of the Singapore-Cambridge general certificate of education O-level: Application of DINA, DINO, G-DINA, HO-DINA, and RRUM. *International Journal of Listening* (online first). DOI: 10.1080/10904018.2018.1500915.

Baker, F. B. & Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques*. New York: Marcel Dekker.

Bley, S. (2017). Developing and validating a technology-based diagnostic assessment using the evidence-centered game design approach –An example of intrapreneurship competence. *Empirical Research in Vocational Education and Training*, *9*(6), 1–32.

Bögeholz, S. & Eggert, S. (2013). Welche Rolle spielt Kompetenzdiagnostik im Rahmen von Lehr-Lernprozessen? *Zeitschrift für Erziehungswissenschaft*, 16, 59–64.

Chen, H. & Chen, J. (2016). Exploring reading comprehension skill relationships through the G-DINA model. *Educational Psychology*, *36*(6), 1049–1064.

Chen, J. & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, *37*, 419–437.

Cizek, G. J., Bunch, M. B. & Konns, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, *23*(4), 31–50.

de la Torre, J. (2009). DINA model parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199.

de la Torre, J. & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*(2), 253–273.

de la Torre, J. & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

DiBello, L. V., Roussos, L. A. & Stout, W. (2006). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of Statistics, Volume 26, Psychometrics* (S. 979–1030). Elsevier, Amsterdam.

DiBello, L. & Stout, W. (2008). *Arpeggio Documentation and Analyst Manual* [Software-Handbuch].

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA Model, classification, latent class sizes, and the Q-Matrix. *Applied Psychological Measurement*, 35, 8–26.

de la Torre, J. & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253–273.

de la Torre, J. & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*(2), 89–97.

Fend, H. (2014). Die Wirksamkeit der neuen Steuerung. Theoretische und methodische Probleme ihrer Evaluation. In K. Maag-Merki, R. Langer & H. Altrichter (Hrsg.), *Educational Governance als Forschungsperspektive. Strategien. Methoden. Ansätze* (S. 27–50). Wiesbaden: Springer VS.

García, P. E., Olea, J. & de la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgement tests. *Psicothema*, *26*(3), 372–377.

George, A., Bley, S. & Pellegrino, J. (2019). Characterizing and diagnosing complex professional competencies – an example of intrapreneurship. *Educational Measurement: Issues and Practice*, *38*(2), 89–100.

George, A. & Robitzsch, A. (2018). Focusing on interactions between content and cognition: A new perspective on gender differences in mathematical sub-competencies. *Applied Measurement in Education*, *31*(1), 79–97.

George, A. C., Robitzsch, A. & Schreiner, C. (2019). Eine Diskussionsgrundlage zur Weiterentwicklung von Rückmeldungen aus standardisierten Kompetenzmessungen am Beispiel Mathematik. In A. C. George, C. Schreiner, C. Wiesner, M. Pointinger & K. Pacher (Hrsg.), *Fünf Jahre flächendeckende Bildungsstandard-Überprüfungen in Österreich* (S. 225–238). Münster: Waxmann.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26, 301–323.

Hartz, S. M. (2002). *A Baysian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Ph.D. thesis University of Illinois Urbana Champaign, IL.

Helmke, A. & Hosenfeld, I. (2003). Vergleichsarbeiten (VERA). Eine Standortbestimmung zur Sicherung schulischer Kompetenzen; Teil 1 & 2. *SchulVerwaltung. Ausgabe Nordrhein-Westfalen*, 4, 107–110.

Henson, R., Templin, J. & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191–210.

Huo, Y. & de la Torre, J. (2014). Estimating a cognitive diagnostic model for multiple strategies via the EM algorithm. *Applied Psychological Measurement*, *38*, 464–485.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion model application to LanguEdge assessment. *Language Testing*, *26*(1), 31–73.

Junker, B. & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272.

Kabiri, M., Ghazi-Tabatabaei, M., Bazargan, A., Shokoohi-Yekta, M. & Kharrazi, K. (2016). Diagnosing competency mastery in science: An application of GDM to TIMSS 2011 data. *Applied Measurement in Education*, *30*(1), 27–38.

Kim, A.-Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, *32*(2), 227–258.

Li, H. & Suen, H. K. (2013). Constructing and validating a Q-Matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment*, *18*(1), 1–25.

Ma, W. & de la Torre, J. (2020). *GDINA: The generalized DINA model framework*. R package version 2.8. (Verfügbar unter https://CRAN.Rproject.org/package=GDINA).

Martin, M. O., Mullis, I. V. S. & Hooper, M. (2017). *Methods and Procedures in PIRLS 2016*. Boston College: TIMSS & PIRLS International Study Center.

Mullis, I. V. S. & Martin, M. O. (2017). *TIMSS 2019 Assessment Frameworks*. Boston College: TIMSS & PIRLS International Study Center.

Nagele, F. (2020). *Professionelle Entscheidungen angehender Lehrpersonen in der situationsspezifischen Förderung selbstregulierten Lernens: Entwicklung und Validierung eines vignettenbasierten Testinstruments anhand eines kognitiven Diagnosemodells*. Unveröffentlichte Dissertation, Paris-Lodron-Universität Salzburg.

OECD. (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing.

OECD. (2019). *PISA 2018 Assessment and Analytical Framework*. Paris: OECD Publishing.

Ravand, H. & Robitzsch, A. (2018). Cognitive diagnostic model of best choice: a study of reading comprehension. *Educational Psychology*, *38*(10), 1255–1277.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York: Springer.

Robitzsch, A. & George, A. C. (2019). The R package CDM for diagnostic modeling. In M. von Davier & Y.-S. Lee (Hrsg.), *Handbook of Diagnostic Classification Models – State of the Art in Modeling, Estimation, and Applications* (S. 549–572). New York: Springer.

Robitzsch, A., Kiefer, T., George, A. & Ünlü, A. (2020). The R package CDM. *A Software Package for Estimation and Simulation of CDMs*, Version 7.5. (Verfügbar unter http://cran.r-project.org/packages/cdm)

Roberts, R. M., Alves, C. B., Chu, M.-W., Thompson, M., Bahry, L. M. & Gotzmann, A. (2014). Testing expert-based versus student-based cognitive models for a grade 3 diagnostic mathematics assessment. *Applied Measurement in Education*, 27(3), 173–195.

Rupp, A., Templin, J. & Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York: The Guilford Press.

Rupp, A. A. & van Rijn, P. W. (2018). GDINA und CDM Packages in R. Measurement: *Interdisciplinary Research and Perspectives*. *16*(1), 71–77.

Schreiner, C. & Wiesner, C. (2019). Die Überprüfung der Bildungsstandards in Österreich: Der erste Zyklus als Meilenstein für die Schul- und Unterrichtsentwicklung – Eine gelungene Innovation im österreichischen Schulsystem. In A. C. George, C. Schreiner, C. Wiesner, M. Pointinger & K. Pacher (Hrsg.), *Fünf Jahre flächendeckende Bildungsstandard-Überprüfungen in Österreich* (S. 13–54). Münster: Waxmann.

Sessoms, J. & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 1–17.

Suchan, B., Wallner-Paschon, C., Bergmüller, S. & Schreiner, C. (2012). *PIRLS & TIMSS 2011. Schülerleistungen in Lesen, Mathematik und Naturwissenschaft in der Grundschule: Die Studie im Überblick*. Graz: Leykam.

Tatsuoka, K. (1984). Analysis of errors in fraction addition and subtraction problems. *Final report for NIE-G-81-0002*. University of Illinois, Urbana-Champaign.

Templin, J. & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287–305.

Templin, J. & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. Educational *Measurement: Issues and Practice, 32*(2), 37–50.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(4), 287–307.

von Davier, M. & Lee, Y.-S. (Hrsg.). (2019). *Handbook of Diagnostic Classification Models – State of the Art in Modeling, Estimation, and Applications*. New York: Springer.

Wang, W., Song, L. & Ding, S. (2018). An exploratory discrete factor loading method for Q-matrix specification in cognitive diagnostic models. In: M. Wiberg, S. Culpepper, R. Janssen, J. González & D. Molenaar (Hrsg.), *Quantitative Psychology. IMPS 2017. Springer Proceedings in Mathematics & Statistics, Vol. 233*. Springer, Cham.

Wang, W., Song, L., Ding, S., Meng, Y., Cao, C. & Jie, Y. (2018). An EM-based method for Q-matrix validation. *Applied Psychological Measurement*, *42*(6), 446–459.

Wang, W., Song, L., Ding, S., Wang, T., Gao, P. & Xiong, J. (2020). A semi-supervised learning method for Q-matrix specification under the DINA and DINO Model with independent structure. *Frontiers in Psychology*, *11*, 2120.

Wu, X., Wu, R., Chang, H.-H., Kong, Q. & Zhang, Y (2020). International comparative study on PISA mathematics achievement test based on cognitive diagnostic models. *Frontiers in Psychology*, *11*, 2230.

You, X., Li, M., Zhang, D. & Liu, H. (2018) Application of a Learning Diagnosis System in Chinese Classrooms. *Applied Psychological Measurement*, *42*(1), 89–94.

Zhan, P., Jiao, H., Man, K. & Wang, L. (2019). Using JAGS for Bayesian Cognitive Diagnosis Modeling: A Tutorial. *Journal of Educational and Behavioral Statistics*, *44*(4), 473–503.