Z Erziehungswiss (2013) 16:41–45
DOI 10.1007/s11618-013-0383-3

ZfE

Wie robust sind Struktur- und Niveaumodelle? Wie zeitlich stabil und über Situationen hinweg konstant sind Kompetenzen?

Alexander Robitzsch

Zusammenfassung: Für viele Kompetenzdomänen ist die zeitliche und situationale Stabilität Untersuchungsgegenstand der empirischen Forschung. Es wird argumentiert, dass die Analyse der Robustheit von Kompetenzstruktur- oder Kompetenzniveaumodellen im Hinblick auf verschiedene Modellbestandteile (statistische Modellparameter oder Kompetenzstufenbeschreibungen) und verschiedene Analyseeinheiten der statistischen Inferenz (z. B. Schüler, Klassen oder Populationen) vorgenommen werden kann.

Schlüsselwörter: Kompetenzstrukturmodelle · Kompetenzniveaumodelle · Robustheit von Kompetenzmodellen

How robust are models of competence? How stable are competencies across time and across different situations?

Abstract: For many domains of competence the temporal and situational stability is in the focus of empirical research. It is argued that the analysis of the robustness of competence structure models or competence level models can be investigated in terms of different model components (statistical model parameters or descriptions of proficiency levels) and different units of analysis of statistical inference (e.g., students, classes or populations).

Keywords: Competence level models · Competence structure models · Robustness of competence models

Dipl.-Math. A. Robitzsch (✉)
Zentrum für Bildungsmonitoring & Bildungsstandards, Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens (BIFIE),
Alpenstraße 121, 5020 Salzburg, Österreich
E-Mail: a.robitzsch@bifie.at

1 Wie robust sind Kompetenzstruktur- und Kompetenzniveaumodelle?

Die Robustheit von Struktur- und Niveaumodellen hängt vom Determinationsgrad des Domäneninhalts ab. Unter Robustheit eines Modells soll dabei das Ausmaß der Invarianz von aus dem Modell abgeleiteten Aussagen und Parametern in verschiedenen Subpopulationen (Personen), Variationen von Operationalisierungen (Items) und Situationen (z. B. Zeitpunkte oder Änderungen von Testkontexten, Designs u. ä.) verstanden werden. Modelle gelten als robust, wenn sie unter den genannten Bedingungen approximativ invariant sind (d. h. nur „wenige Abweichungen" von absoluter Invarianz besitzen oder diese bei einer großen Anzahl von Items wenig Bedeutung erlangen). Die Ansätze der Sensitivität (resp. Sensitivitätsanalysen) und der Generalisierbarkeit sind mit der Frage der Robustheit eng verbunden.

Dabei scheint die Robustheit mit der Feinkörnigkeit (*grain size*) der Domäne zusammenzuhängen: Kompetenzstrukturmodelle mit niedriger Feinkörnigkeit (grober Struktur) haben eine eher höhere Robustheit und erscheinen weniger abhängig von kleineren Veränderungen der Domäne, ihre Nützlichkeit ist jedoch eingeschränkt (vgl. Neumann 2013 in diesem Heft).

Niveaumodelle sind mit dem Ziel eingeführt worden, die Kommunikation über die zugrunde liegende Kompetenz zu erleichtern. Inzwischen ist unter anderem durch die fachdidaktische Beschreibung von Anforderungsmerkmalen in vielen Domänen eine rationale Grundlage für Niveaubeschreibungen geschaffen worden. Empirisch kann dies durch die Erklärung von Itemschwierigkeiten belegt werden (vgl. Fleischer et al. 2013 in diesem Heft). Allerdings sind entsprechende Niveaumodelle noch zu wenig auf Robustheit geprüft. Die Abgrenzung und das Labeling für Niveaustufen (etwa in Verfahren des Standard Setting) haben derzeit allerdings noch einen gewissen Grad an Beliebigkeit. Die Etablierung theoretisch fundierter und robuster Niveaumodelle ist also nach wie vor in der Kompetenzdiagnostik ein Desiderat. Im angloamerikanischen Raum finden Verfahren der vertikalen Verlinkung (vgl. Dorans et al. 2007) Verwendung, bei der eine Domäne – wie beispielsweise das Leseverstehen – über verschiedene Klassenstufen eindimensional unter Verwendung von Item-Response-Modellen abgebildet wird. Die in diesen Modellen involvierten latenten Variablen scheinen allerdings eher den Status eines Ordinalskalenniveaus zu besitzen (vgl. z. B. Ballou 2009; Lord 1980), sodass beliebige monotone Transformationen Aussagen über längsschnittliche Veränderung oder querschnittliche Differenzen verschiedener Klassenstufen ändern können (vgl, Robitzsch et al. 2011).[1]

Die Niveaubeschreibungen in einem Niveaumodell der Domäne des Leseverstehens können dabei jedoch in Modellen für verschiedene Klassenstufen differieren. Anhand dieses Konstrukts wird deutlich, dass Robustheit eines Modells zunächst „Robustheit" des zugrunde liegenden Konstrukts (oder einer Domäne) als Voraussetzung besitzt und damit Fragen der Robustheit nicht von Fragen der Validität zu trennen sind. In Domänen wie der Mathematik scheint für bestimmte Teilkompetenzen aufgrund curricularer Rahmenbedingungen eine vertikale Verlinkung nicht möglich und auch nicht sinnvoll.

Zur Untersuchung von Robustheitseigenschaften von Niveau- und Strukturmodellen wäre die Anwendung von Multitrait-Multimethod-Ansätzen (vgl. Nussbeck et al. 2007) oder der Generalisierbarkeitstheorie (vgl. Brennan 2011) erstrebenswert, die explizit verschiedene Facetten der Robustheit von Modellen anhand empirischer Daten untersuchen.

In diesem Rahmen kann die Prüfung der Robustheit über verschiedene Zeitpunkte und Personengruppen hinweg erfolgen. Robustheit von Strukturmodellen wird dabei tendenziell mit der Prüfung auf gewisse invariante Modellbestandteile auf der Personenseite (also hinsichtlich der Dimensionalität) untersucht, während Robustheit von Niveaumodellen eher auf die Untersuchung der Invarianz von Itemparametern (Itemschwierigkeiten und Itemladungen) unter Variation von Personenpopulationen, Zeitpunkten und Situationen abzielt. Beobachtet man beispielsweise für einen Test zum Leseverstehen nichtinvariante Itemschwierigkeiten in zwei aufeinanderfolgenden Klassenstufen, so spricht dies im Hinblick auf die Modellinterpretation über verschiedene Klassenstufen hinweg gegen die Robustheit eines zugehörigen Niveaumodells.

Für die Beschreibung einer konkreten Domäne müssen zugehörige Niveau- und Strukturmodelle nicht zugleich robust sein. Die notwendige Robustheit von Modellen ist dabei im Hinblick auf den Verwertungszweck der Modellresultate zu interpretieren. Für viele Modelle in der Kompetenzdiagnostik gilt aufgrund noch ausstehender Analysen zur Robustheit, dass wir uns weitgehend noch in dem Stadium der Modellexploration und noch nicht der Modellprüfung befinden.

2 Wie zeitlich stabil und über Situationen hinweg konstant sind Kompetenzen?

Faktisch werden Kompetenzmodelle für spezifische Populationen in spezifischen (mehr oder weniger breit angelegten) Kontexten zu spezifischen Zeitpunkten entwickelt. Soll eine Generalisierung stattfinden, so ist zu prüfen, inwieweit Struktur- und Niveaumodelle in unterschiedlichen Populationen und zu unterschiedlichen Zeitpunkten im Entwicklungsverlauf Gültigkeit haben. Dabei müssen Niveaumodelle nicht notwendigerweise individuelle Entwicklungen vorhersagen. Querschnittlich angelegte Niveaumodelle stellen daher keine Kompetenzentwicklungsmodelle dar. Dies wird am Paradigma der Unterscheidung intraindividueller von interindividuellen Unterschieden deutlich (vgl. Molenaar und Campbell 2009).

In einer längsschnittlichen Betrachtung sind Kompetenzen intraindividuell veränderlich. Auch wenn Kompetenzmodelle invariant und robust sind, wie es häufig angenommen wird, sind die Kompetenzausprägungen veränderbar bzw. beeinflussbar. Kompetenzausprägungen verändern sich dabei über die Zeit, wobei die Entwicklung nicht notwendigerweise einer monoton wachsenden Funktion folgt. Beispielsweise können Skalenwerte curricular abhängiger(er) Kompetenzbereiche längsschnittlich stagnieren oder sogar absinken. Generell kann dann untersucht werden, inwiefern sich die stichprobenabhängige Dimensionalität von untersuchten Konstrukten über die Zeit ändert (sog. *construct shift*; vgl. Reckase 2009).

Ein Messinstrument, das in querschnittlicher Perspektive nicht diskriminant valide ist, bzw. ein Konstrukt, das sich empirisch (noch) nicht von anderen Konstrukten unterscheiden lässt, könnte auch dann sinnvoll sein, wenn sich in längsschnittlicher Sicht zum Beispiel in Abhängigkeit von Instruktion differenzielle Entwicklung vollzieht (vgl. Fleischer et al. 2013 in diesem Heft zur Definition des Kompetenzbegriffs; vgl. auch Briggs 2011). Maße der instruktionalen Sensitivität (vgl. Polikoff 2010) sind dabei häufig auf Populations- oder Subpopulationsebene definiert und fokussieren dabei nicht auf die

intraindividuelle Veränderung, sondern gegebenenfalls auf höhere Organisationseinheiten (z. B. Schulklassen, Schulen oder Bundesstaaten mit verschiedenen Curricula).

Zusammenfassend deuten empirische Befunde auf Populationsebene darauf hin, dass es ebenso zeitliche Instabilität gibt. Die Veränderungen verlaufen jedoch deutlich langsamer als auf individueller Ebene. Auch situationale Stabilität scheint nur schwer zu erreichen zu sein, was allerdings dem Konzept von Kontextspezifität bei Kompetenzen entspricht. Selbst Kompetenzmodelle (zumindest Kompetenzniveaumodelle) sind im Allgemeinen nicht situationsinvariant, da sie dem Einfluss der Volition bei verschiedenen Gegenständen unterworfen sind.

Danksagung: Diese Veröffentlichung wurde ermöglicht durch Sachbeihilfen der Deutschen Forschungsgemeinschaft (Kennz.: WI 2667/7-1) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

Anmerkung

1 Für die Frage der Definition einer geeigneten Skalierung sind Analysen mit der Item-Response-Theorie gegenüber jenen mit klassischer Testtheorie oder der Generalisierbarkeitstheorie nicht zwingend vorzuziehen (vgl. Brennan 2011).

Literatur

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy, 4,* 351–383.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24,* 1–21.

Briggs, D. C. (2011). Cause or effect? Validating the use of tests for high-stakes inferences in education. In N. Dorans & S. Sinharay (Hrsg.), *Looking back: Proceedings of a conference in honor of the career of Paul Holland* (S. 131–147). New York: Springer.

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales*. New York: Springer.

Fleischer, J., Koeppen, K., Kenk, M., Klieme, E., & Leutner, D. (2013). Kompetenzmodellierung. Struktur, Konzepte und Forschungszugänge des DFG-Schwerpunktprogramms. In D. Leutner, E. Klieme, J. Fleischer & H. Kuper (Hrsg.), *Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen: aktuelle Diskurse im DFG-Schwerpunktprogramm* (18. Sonderheft der Zeitschrift für Erziehungswissenschaft, DOI: 10.1007/s11618-013-0379-z). Wiesbaden: VS Verlag für Sozialwissenschaften.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Erlbaum.

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science, 18,* 112–117.

Neumann, K. (2013). Mit welchem Auflösungsgrad können Kompetenzen modelliert werden? In welcher Beziehung stehen Modelle zueinander, die Kompetenz in einer Domäne mit unterschiedlichem Auflösungsgrad beschreiben? In D. Leutner, E. Klieme, J. Fleischer & H. Kuper (Hrsg.), *Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen: aktuelle Diskurse im DFG-Schwerpunktprogramm* (18. Sonderheft der Zeitschrift für Erziehungswissenschaft, DOI: 10.1007/s11618-013-0382-4). Wiesbaden: VS Verlag für Sozialwissenschaften.

Nussbeck, F. W., Eid, M., Geiser, C., Courvoisier, D. S., & Cole, D. A. (2007). Multitrait-Multimethod-Analysen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 361–388). Berlin: Springer.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice, 29,* 3–14.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Robitzsch, A., Dörfler, T., Pfost, M., & Artelt, C. (2011). Die Bedeutung der Itemauswahl und der Modellwahl für die längsschnittliche Erfassung von Kompetenzen: Lesekompetenzentwicklung in der Primarstufe. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie, 43,* 213–227.