Z Erziehungswiss (2013) 16:47–51
DOI 10.1007/s11618-013-0386-0

ZfE

Sind Modelle der Item-Response-Theorie (IRT) das „Mittel der Wahl“ für die Modellierung von Kompetenzen?

Johannes Hartig · Andreas Frey

Zusammenfassung: Modelle der Item-Response-Theorie (IRT) gehören zur großen Gruppe von statistischen Analysemodellen mit latenten Variablen. Sie kommen bei der Auswertung standardisierter Tests zur Messung von Kompetenzen zunehmend zum Einsatz. Der vorliegende Beitrag fasst die spezifischen Vorteile von IRT-basierten Auswertungen gegenüber traditionellen Methoden sowie gegenüber anderen Modellen mit latenten Variablen (z. B. Strukturgleichungsmodellen) zusammen.

Schlüsselwörter: Item-Response-Theorie (IRT) · Skalierung · Messen · Testen

Benefits and limitations of modeling competencies by means of Item Response Theory (IRT)

Abstract: Item response theory (IRT) models can be subsumed under the larger class of statistical models with latent variables. IRT models are increasingly used for the scaling of the responses derived from standardized assessments of competencies. The paper summarizes the strengths of IRT in contrast to more traditional techniques as well as in contrast to alternative models with latent variables (e. g. structural equation modeling). Subsequently, specific limitations of IRT and cases where other methods might be preferable are lined out.

Keywords: Item response theory (IRT) · Measurement · Scaling models · Testing

Prof. Dr. J. Hartig (✉)
Bildungsqualität und Evaluation, Deutsches Institut für Internationale Pädagogische Forschung,
Schloßstr. 29, 60486 Frankfurt am Main, Deutschland
E-Mail: hartig@dipf.de

Prof. Dr. A. Frey (✉)
Institut für Erziehungswissenschaft, Friedrich-Schiller-Universität Jena,
Am Planetarium 4, 07737 Jena, Deutschland
E-Mail: andreas.frey@uni-jena.de

1 Welche Vorteile haben Methoden der Item-Response-Theorie (IRT) für die Modellierung von Kompetenzen?

Zur Messung und psychometrischen Modellierung von Kompetenzen (vgl. z. B. Hartig 2008) sollten vorzugsweise Methoden mit latenten Variablen (vgl. z. B. Skrondal und Rabe-Hesketh 2004) zum Einsatz kommen. Hierzu zählen Modelle der Item-Response-Theorie (IRT), je nach Datenlage und Zielsetzung aber auch andere Modellklassen wie beispielsweise Strukturgleichungsmodelle. Mit diesen Modellen kann systematische Varianz, die auf das zu messende Merkmal zurückgeht, von unsystematischer Fehlervarianz separiert und die Passung angenommener Modellstrukturen an empirische Daten statistisch getestet werden. Andere ältere Methoden (z. B. Methoden der klassischen Testtheorie [KTT] oder exploratorische Faktorenanalysen [EFA]), bei denen dies nicht der Fall ist, können flankierend eingesetzt werden, sollten aber nicht als alleiniger methodischer Zugang Verwendung finden.

Eine spezifische Stärke von IRT-Modellen liegt darin, dass sie die Lokalisation von sowohl Aufgabenschwierigkeiten als auch Personenmerkmalen auf einer gemeinsamen Skala ermöglichen. Dies ist eine wesentliche Grundlage für die kriteriumsorientierte Definition und Beschreibung von Kompetenzniveaus. Die gemeinsame Skala ist in der diagnostischen Praxis Voraussetzung für niveaubezogene Rückmeldungen von Testergebnissen und ermöglicht darüber hinaus eine effiziente Optimierung der Aufgabenauswahl beim computerisierten adaptiven Testen (CAT, z. B. Frey 2012). Eine weitere Stärke haben IRT-Modelle, wenn dieselben Aufgaben in verschiedenen Studien zum Einsatz kommen. Unter bestimmten Voraussetzungen können die Skalen der Studien, auch wenn sie nur eine gemeinsame Teilmenge von Aufgaben haben, im Rahmen eines sogenannten *Linkings* (vgl. z. B. Kolen und Brennan 2004) auf eine gemeinsame Metrik gebracht und vergleichend interpretiert werden. Ein weiterer Vorteil von IRT-Modellen besteht darin, dass mit ihnen auch Daten analysiert werden können, die auf Basis von Multi-Matrix-Designs (vgl. z. B. Frey et al. 2009) erhoben wurden. Die bei der Verwendung von Multi-Matrix-Designs resultierenden unvollständigen Datenstrukturen können mit auf Kovarianzstrukturen basierenden Modellen, wie beispielsweise Strukturgleichungsmodellen, nur analysiert werden, wenn das Matrix-Design dergestalt balanciert ist, dass jede mögliche Kombination von Items realisiert wurde. Eine weitere für die Modellierung von Kompetenzen häufig nützliche Eigenschaft von IRT-Modellen stellt die Möglichkeit dar, Eigenschaften der Testaufgaben explizit zu parametrisieren. Dies ist beispielsweise mit dem Linear-Logistischen Testmodell (LLTM; vgl. Fischer 1996) sowie verwandten Modellen (vgl. z. B. Janssen et al. 2004) möglich. Hierdurch können beispielsweise kognitive Anforderungen modelliert werden, was ein tieferes Verständnis der untersuchten Kompetenz ermöglicht (vgl. z. B. Hartig et al. 2012). Dieser Einbezug ist prinzipiell auch in anderen Modellen mit latenten Variablen möglich (vgl. z. B. Hartig et al. 2007), im Zusammenhang mit der durch IRT-Modelle gegebenen gemeinsamen Skala für Aufgabenschwierigkeiten und Personenmerkmale aber besonders nützlich. Bei Modellen mit Aufgabeneigenschaften muss beachtet werden, dass Modelle zum Zusammenhang von Aufgabeneigenschaften und Aufgabenschwierigkeiten nicht zu einfach formuliert werden und auch mögliche Interaktionen zwischen Aufgabenmerkmalen, aber auch zwischen Aufgaben- und Personenmerkmalen in Betracht gezogen werden.

Die Modellierung von Zeitverläufen ist innerhalb von IRT-Modellen auf verschiedene Weisen möglich (vgl. z. B. Fischer und Seliger 1996; Hartig und Kühnbach 2006), allerdings bisher mit Strukturgleichungsmodellen (vgl. z. B. Bollen und Curran 2006) oder Mehrebenen-Modellen (vgl. z. B. Singer und Willett 2003) besser etabliert.

2 Wo liegen Grenzen der Modellierung von Kompetenzen mit IRT-Modellen?

An Grenzen stoßen IRT-Modelle bei der Berücksichtigung von Verletzungen der lokalen stochastischen Unabhängigkeit. Die Annahme drückt aus, dass die Art der Beantwortung einer Aufgabe unabhängig davon ist, wie andere Aufgaben im selben Test beantwortet werden. Lokale Abhängigkeiten zwischen den Antworten auf verschiedene Aufgaben können bei Kompetenzmessungen beispielsweise bei Aufgaben mit gemeinsamen Aufgabenstämmen (Testlets) oder in Messwiederholungsdesigns auftreten. Prinzipiell ist eine Berücksichtigung derartiger Abhängigkeiten möglich (Überblick in Chen 2010 sowie Wainer et al. 2007). Für empirische Datensätze mit vielen Aufgabenstämmen können diese Modelle jedoch sehr große Stichproben benötigen, um zuverlässige Schätzungen zu gewährleisten, oder sogar zu komplex werden, um mit momentan verfügbaren Computern geschätzt werden zu können.

Eine Schwäche vieler IRT-Modelle ist derzeit das Fehlen etablierter globaler Kriterien für die Beurteilung der Modellgüte. Während konkurrierende Modelle für dieselbe Datenlage anhand der Informationskriterien (AIC, BIC) verglichen werden können, fehlen für viele Modelle etablierte Maße und Tests zur Einschätzung der absoluten Passung der Daten auf das Modell. Bei Strukturgleichungsmodellen ist die Situation günstiger, da hier eine größere Anzahl von etablierten globalen Fitindizes berechnet werden kann. Insofern kann die Anwendung von Strukturgleichungsmodellen für ordinale Daten von Vorteil sein, wenn etwa mehrdimensionale Strukturen überprüft werden sollen.

Eine Herausforderung – aber kein prinzipielles Problem – besteht darin, den Kontext, in dem die Messungen durchgeführt werden, seitens der verwendeten IRT-Modelle zu berücksichtigen (z. B. high-stakes vs. low-stakes, unterschiedliche Bearbeitungsstrategien etc.). Denkbar ist der Einbezug derartiger Kontexte durch die Verwendung von Mischverteilungsmodellen (vgl. z. B. Rost und von Davier 1995). Mit diesen speziellen IRT-Modellen können Gruppen in unterschiedlichen Kontexten oder mit unterschiedlichen Bearbeitungsstrategien durch latente Klassen repräsentiert werden. Die hiermit verbundenen möglichen Probleme sind allerdings nicht IRT-spezifisch, sondern bei anderen Auswertungsmethoden (z. B. Strukturgleichungsmodelle, KTT) mindestens in gleichem Umfang gegeben. Generell sollte eine Vergleichbarkeit verschiedener Datenerhebungen (Durchführungsobjektivität) schon durch geeignete Erhebungsstrategien sichergestellt werden.

Grundsätzlich reflektiert werden sollte die den meisten IRT-Modellen zugrunde liegende Annahme kontinuierlicher Merkmalsdimensionen. Kategoriale Eigenschaften, zum Beispiel qualitative Übergänge wie in der Conceptual-Change-Forschung, sind mit Modellen mit latenten Klassen angemessener modellierbar. Hier ist unter dem Schlagwort „Kognitive Diagnosemodelle“ (vgl. z. B. Rupp et al. 2010) aktuell zwar eine rege Methodenentwicklung zu verzeichnen, aber es fehlen erfolgreiche Anwendungsbeispiele

(vgl. auch Kunina-Habenicht et al. 2009). Ferner sind die für komplexere kognitive Diagnosemodelle benötigten Stichproben teilweise extrem groß und in vielen Studien nicht zu erreichen.

Die Notwendigkeit vergleichsweise großer Stichproben stellt beim Einsatz von IRT-Modellen ein generelles forschungspraktisches Problem dar. Die großen Stichproben werden benötigt, um hinreichend präzise Parameterschätzungen zu erhalten. Während bei Analysen mit dem eindimensionalen Rasch-Modell häufig schon 100 Antworten pro Item ausreichend sind, benötigen beispielsweise komplexe kognitive Diagnosemodelle sechsstellige Stichprobengrößen.

Als ein praktisches Problem kann die schwierige Vermittelbarkeit von Ergebnissen aus komplexen Modellen an ein breiteres Publikum (z. B. psychometrisch nicht vorgebildete Lehrkräfte oder Politiker/innen) betrachtet werden.

3 Sind neue Verfahren notwendig, um komplexe Kompetenzen adäquat modellieren zu können?

IRT-Modelle sind unter anderem deshalb attraktiv, weil viele Faktoren integriert werden können und sie (bei Einbezug latenter Klassenmodelle) nicht notwendigerweise an kontinuierliche Fähigkeitsdimensionen gebunden sind. Die in der psychometrischen Fachliteratur beschriebenen Modelle sind sehr umfangreich, sodass für das DFG-Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ nicht die Notwendigkeit besteht, neue Verfahren zu entwickeln. Die Herausforderung für die kommenden Jahre besteht vielmehr darin, die mathematisch formulierten Modelle für die Kompetenzdiagnostik zugänglich zu machen und deren Nützlichkeit anhand empirischer Daten zu belegen.

Danksagung: Diese Veröffentlichung wurde ermöglicht durch Sachbeihilfen der Deutschen Forschungsgemeinschaft (Kennz.: HA 5050/2-2 und FR 2552/2-2) im Schwerpunktprogramm „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPP 1293).

Literatur

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. New York: Wiley.

Chen, T. A. (2010). *Random or fixed testlet effects: A comparison of two multilevel testlet models* [Unveröffentlichte Dissertation]. Austin: University of Texas.

Fischer, G. H. (1996). Unidimensional linear logistic Rasch models. In W. J. van der Linden & R. K. Hambleton (Hrsg.), *Handbook of modern item response theory* (S. 225–243). New York: Springer.

Fischer, G. H., & Seliger, E. (1996). Multidimensional linear logistic models for change. In W. J. van der Linden & R. K. Hambleton (Hrsg.), *Handbook of modern item response theory* (S. 323–346). New York: Springer.

Frey, A. (2012). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2., akt. und überarb. Aufl., S. 275–293). Berlin: Springer.

Frey, A., Hartig, J., & Rupp, A. (2009). Booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28,* 39–53.

Hartig, J. (2008). Psychometric models for the assessment of competencies. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of competencies in educational contexts* (S. 69–90). Göttingen: Hogrefe.

Hartig, J., & Kühnbach, O. (2006). Schätzung von Veränderungswerten mit Plausible Values in mehrdimensionalen Rasch-Modellen. In A. Ittel & H. Merkens (Hrsg.), *Veränderungsmessung und Längsschnittstudien in der Erziehungswissenschaft* (S. 27–44). Wiesbaden: VS Verlag für Sozialwissenschaften.

Hartig, J., Hölzel, B., & Moosbrugger, H. (2007). A confirmatory analysis of item reliability trends (CAIRT): Differentiating true score and error variance in the analysis of item context effects. *Multivariate Behavioral Research Methods, 42,* 157–183.

Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement, 72,* 665–686.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Hrsg.), *Explanatory item response models: A generalized linear and nonlinear approach* (S. 189–212). New York: Springer.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation, 35,* 64–70.

Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In G. Fischer & I. Molenaar (Hrsg.), *Rasch Models: Foundations, recent developments, and applications* (S. 257–268). New York: Springer.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guildford Press.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: University Press.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton: Chapman & Hall.

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