Elektronisches Supplement 2. Exkurs: Modellwahl in der IRT.

"Many researchers have made the choice of a particular item response model into the focus of a debate that can only be termed religious. However, there is no theoretical or practical reason that we should confine ourselves to a single item response model, and a number of reasons why we should not." (Wise & Kingsbury, 2000, S. 139).

In der Literatur zur Testentwicklung und der Item Response Theorie (IRT) hinsichtlich der Auswahl eines Item Response Modells werden zum einen eher datengetriebene und zum anderen stärker theoretisch motivierte Perspektiven vertreten. Auch deshalb lässt sich das Thema sehr facettenreich diskutieren. Es gibt unseres Wissens nach bisher keine Literatur, die das Thema der Modellwahl explizit und umfassend für den Kontext von CATs mit Multiple-Choice (MC)-Textverständnisaufgaben diskutiert. Wir werden uns bemühen verschieden Perspektiven zu berücksichtigen. Dabei beschränken wir uns auf die Aspekte der theoretischen Plausibilität und die Passung zwischen dem Modell und den Antwortdaten. Es sollte aber vorab betont werden, dass kein Modell alternativlos ist und sicherlich verschiedene Modelle zu einem zufriedenstellenden Ergebnis führen würden.

In der IRT wird häufig im Zusammenhang mit dichotomen Antwortformaten zwischen 1PL-, 2PL- und 3PL-Modellen unterschieden. Die Zahl des Modells bezeichnet die Parametrisierung der Item Characteristic Curve (ICC). Die ICC beschreibt die Wahrscheinlichkeit einer korrekten Antwort für ein Item in Abhängigkeit zur Personenfähigkeit mit einer Exponentialfunktion. Ein 1PL-Model parametrisiert nur die Item-spezifische Lage der Exponentialfunktion, die auch als Schwierigkeit bezeichnet wird. Ein 2PL-Model parametrisiert zusätzlich ein Item-spezifische Steigung, die auch als Diskriminationsparameter bezeichnet wird. Ein 2PL-Modell geht davon aus, dass die Wahrscheinlichkeit einer korrekten Antwort für Personen mit einer relativ zur Itemschwierigkeit niedrigen Fähigkeit immer irgendwann Null ist. Ein 3PL-Model parametrisiert zusätzlich eine Asymptote im unteren Bereich der Exponentialfunktion. Die Asymptote definiert die Wahrscheinlichkeit einer korrekten Antwort, für SuS mit einer relativ zur Itemschwierigkeit geringen Fähigkeit. Es kann deshalb mit einbeziehen, dass die Wahrscheinlichkeit einer korrekten Antwort nie unter eine bestimmte Ratewahrscheinlichkeit sinkt. Die Ratewahrscheinlichkeit kann Item-spezifisch, für bestimmte Itemgruppen (z. B. eines Antwortformats) oder für alle Items gleich geschätzt werden. Wir nennen ein 3PL-Modell das für alle Items eine gemeinsame Ratewahrscheinlichkeit schätzt 3PLc.

<u>Theoretische Plausibilität der Parametrisierung</u>

Van der Linden (2009) begründet die Modellauswahl für eine adaptives Testverfahren wie folgt: „The choice of model was made to get compatibility with the response format of the items" (p. 197). Bei der Wahl eines Item Response Modells wird häufig das Antwortformat berücksichtigt. Grundsätzlich lässt sich die Parametrisierung mit Schwierigkeit, Diskriminationsfähigkeit und Ratewahrscheinlichkeit bei Textverständnisitems im MC-Format theoretisch plausibel begründen.

- Textverständnisitems unterscheiden sich anhand ihrer Schwierigkeit, weil unter anderem Texte unterschiedliche sprachliche Komplexität aufweisen oder unterschiedliche Aufgabenanforderung haben.
- Textverständnisitems unterscheiden sich anhand ihrer Diskriminationsfähigkeit, weil sie zum Beispiel unterschiedliche Aspekte eines Konstrukts erfassen (z. B. Hartig, 2008).
- Textverständnisitems im Multiple-Choice (MC)-Format bieten die Möglichkeit richtige Antwort zu Raten. In diesem Zusammenhang sollte zwischen der theoretischen (tR) und der empirischen Ratewahrscheinlichkeit (eR) unterschieden werden. Die eR weicht häufig von der tR ab (Hambleton & Swaminathan, 2013).
    - Die tR bei einem MC-Item ergibt sich aus der Anzahl der Distraktoren und korrekten Antworten. Die Textverständnisitems haben vier Antwortmöglichkeiten und eine korrekte Antwort, damit ist die tR ¼ = 25 %. Die tR sollte in einem 3PL-Modell nicht als Ratewahrscheinlichkeit festgelegt werden (Waller, 1989).
    - Die eR ergibt sich aus den Anforderungen (bzw. den fehlenden Anforderungen) die mit der Attraktivität von Distraktoren in Verbindung stehen.
        - Die eR kann über der tR liegen, wenn SuS mit relativ zur Itemschwierigkeit geringen Fähigkeiten Distraktoren ausschließen können und unter den verbleibenden Antwortoptionen zufällig wählen. Bei Textverständnisitems könnte dies zum Beispiel der Fall sein, wenn Distraktoren augenscheinlich unsinnig sind

und ohne jegliches Textverständnis ausgeschlossen werden. Um dies zu vermeiden, sollten Distraktoren möglichst plausibel gestaltet werden (Gierl et al., 2017).

- Die eR kann unter der tR liegen, wenn SuS mit relativ zur Itemschwierigkeit geringen Fähigkeiten Distraktoren attraktiver finden als die korrekte Antwortoption. Bei Textverständnisitems könnte dies zum Beispiel der Fall sein, wenn Distraktoren allgemeingültige Aussagen treffen oder oberflächlich mit dem Text übereinstimmen, aber nicht auf die Frage antworten (Gierl et al., 2017).

Bei dem eingesetzten MC-Format kann prinzipiell durch Raten die richtige Antwort ausgewählt werden. Die eR weicht dabei häufig von der tR ab. Wir haben daher 3PL-Modelle, die auch die eR modellieren, in Betracht gezogen. Für die Modellwahl sollten jedoch noch weitere Aspekte in Erwägung gezogen werden.

Passung zwischen Modell und Daten

Für eine Modellwahl sollte auch die Passung zwischen Modell und den Daten berücksichtig werden. Kang und Cohen (2007) betonten: „Choosing an appropriate IRT model is crucial to obtain the benefits of IRT applications such as test development, item banking, detection of differential item functioning, adaptive testing, and test equating." (p. 353). Dabei ist zu beachten, dass ein Modell, das eine höhere Parametrisierung annimmt als nötig, gegen das Prinzip der Parsimonie verstößt. Im Wesentlichen sollte man das einfachste Modell wählen, das die Daten noch gut erklärt (Embretson & Reise, 2000). Passungsindikatoren wie Araike Informationskritierum (AIC) und das Baysian Informationskriterium (BIC) erlauben einen Vergleich der Modellpassung zwischen verschieden Modellen der IRT und berücksichtigen dabei das Parsimonieprinzip indem sie die Parameteranzahl eines Modells berücksichtigen. Ein Modell, das nach diesem Prinzip ausgewählt wird, hat eine geringere Chance, Inkonsistenzen, Mehrdeutigkeiten und Redundanzen einzuführen (Embretson & Reise, 2000).

Zusätzlich sollte abgewogen werden, ob der gegebene Stichprobenumfang die Schätzung eines komplexen Modells zulässt. Die Schätzung eines 1PL-Modelles auf Basis der Antworten von 500 Probandinnen und Probanden führt häufig schon zu zuverlässigen Ergebnissen (Stone & Yumoto, 2004). Die Stichprobenanforderungen von Modellen mit höherer Parametrisierung können größer sein (Stone & Yumoto, 2004). Die Angemessenheit eines Stichprobenumfangs hängt von verschiedenen Faktoren ab, wie der Anzahl der Items, der Ausgestaltung eines Multi-Matrix-Design und der Korrelation der Itemparameter. Ein wichtiger Indikator für die Angemessenheit des Stichprobenumfangs ist nach Wright (1999) die Parameterkonvergenz.

Wir haben die Passung zwischen einem 2PL-Modell und einem 3PLc-Modell anhand von Passungsindikatoren verglichen und festgestellt, dass die Modellpassung eines 3PLc-Modells besser ist als, die des 2PL-Modells, obwohl das 3PLc-Modell nur einen weiteren Parameter geschätzt. Bei der Schätzung des deutlich komplexeren unrestringierten 3PL-Modelles sind hingegen Konvergenzprobleme aufgetreten. Ein unzureichender Stichprobenumfang könnte ein wichtiger Grund dafür sein. Wir haben ein 3PL-Modell nicht mehr in Betracht gezogen, weil wir mit den gegebenen Daten keine zuverlässigen Parameterschätzungen erzielen konnten.

Zusammenfassung Modellwahl

Wir haben uns für ein Modell mit allgemeinen pseudo-Rateparameter (3PLc) entschieden, weil das Modell theoretisch plausibel zu den Textverständnisaufgaben im MC-Format passt und, weil die Modellpassung unter Berücksichtigung des Parsimonieprinzips besser auf die Daten passt als ein 2PL-Modell. Die zugrundeliegende Stichprobe erlaubte uns keine zuverlässige Parameterschätzung eines 3PL-Modell mit Item-spezifischen Ratewahrscheinlichkeiten.

Referenzen

Gierl, M. J., Bulut, O., Guo, Q. & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: a comprehensive review. *Review of Educational Research, 87*(6), 1082–1116.

Hambleton, R. K. & Swaminathan, H. (2013). *Item response theory: Principles and applications.* New York, NY: Springer Science & Business Media.

Hartig, J. (2008). Psychometric models for the assessment of competencies. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of Competencies in Educational Contexts* (pp. 69–90). Cambridge, MA: Hogrefe & Huber.

Kang, T. & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*(4), 331–358.

Stone, M. & Yumoto, F. (2004). The effect of sample size for estimating Rasch / IRT parameters with dichotomous items. *Journal of applied measurement*. *5*(1), 48–61.

Van der Linden, W. J. & Pashley, P. J. (2009). Item selection and ability estimation in adaptive testing. In W. J. Van der Linden & P. J. Pashley (Eds.), *Elements of adaptive testing* (pp. 3–30). Springer, New York, NY.

Waller, M. I. (1989). Modeling guessing behavior: A comparison of two IRT models. *Applied Psychological Measurement, 13*(3), 233–243.

Wise, S. L. & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica, 21*(1), 135–155.

Wright, B. (1999). Fundamental Measurement for Psychology. In Embretson, S. E. & Hershberger L. H. (Eds.), *The New Rules of Measurement* (pp. 65–104.). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.