KIRSTEN, Katharina; GREEFRATH Gilbert & EMMRICH, Rico
Münster, Berlin

Technologiebasiert vs. papierbasiert: Moduseffekte in VERA

In vielen Bereichen werden Tests zunehmend technologiegestützt durchgeführt ("technology based assessment", TBA). Dieser Trend ist nicht nur in internationalen Vergleichsstudien wie PISA und TIMSS erkennbar, sondern auch in den deutschlandweiten VERgleichsArbeiten VERA wird neben dem traditionellen, papierbasierten Test ("paper-pencil-assessment", PPA) seit einigen Jahren eine TBA-Version angeboten. Insbesondere dann, wenn zwei Testversionen simultan eingesetzt werden, stellt sich jedoch die Frage, inwieweit TBA in Bezug auf die Schwierigkeit, Validität und Konstruktäquivalenz mit PPA vergleichbar ist. Hat die Testdurchführung mit PPA und TBA einen kausalen Effekt auf das Testresultat, z. B. die geschätzte Kompetenz, spricht man von Moduseffekten (Kroehne & Martens, 2011).

Moduseffekte in Large-Scale-Assessments und ihre Ursachen

Empirische Ergebnisse zu Moduseffekten sind nicht konsistent. Während eine Metastudie von Wang et al. (2007) keine statistischen Unterschiede zwischen TBA- und PPA-Versionen von Mathematiktests nachweisen konnte, traten in den Feldversuchsdaten für PISA 2015 Moduseffekte zugunsten der PPA-Version auf (Jerrim et al., 2018). Ähnliche Studien zu Moduseffekten basierend auf TIMSS 2019 legen nahe, dass PPA und TBA dasselbe Konstrukt messen, die technologiebasierten Items im Durchschnitt jedoch etwas schwieriger sind als die papierbasierten Items (Fishbein et al., 2018).

Ursachen von Moduseffekten können vielfältig sein. Neben Personenmerkmalen wie der Vertrautheit mit der verwendeten Technologie können insbesondere auch Itemmerkmale Moduseffekte hervorrufen. Der Transfer von PPA-Items in eine technologiegestützte Testumgebung umfasst Designentscheidungen, welche die Wahrnehmung und Teststrategie der Lernenden beeinflussen können. Bei der Betrachtung des Testdesigns unterscheiden wir zwischen formal-technischen und inhaltlichen Itemmerkmalen. Formaltechnische Merkmale beziehen sich auf die technische Umsetzung eines Items und umfassen Merkmale wie das Layout, das Antwortformat oder die Navigation innerhalb der Items (Kroehne & Martens, 2011). Zumindest für Lesetests konnte bereits nachgewiesen werden, dass formal-technische Itemmerkmale Moduseffekte moderieren können. So unterscheiden sich TBA- und PPA-Items in ihrer Schwierigkeit in Abhängigkeit von ihrem Layout und dem geforderten Antwortformat (Buerger et al., 2019). Für die technische Umsetzung von geometrischen Konstruktionen ist in Mathematiktests zudem die Integration digitaler Werkzeuge wie GeoGebra relevant. Da digitale

In: L. Schick, M. Platz & A. Lambert (Hrsg.),
Beiträge zum Mathematikunterricht 2025.
58. Jahrestagung der Gesellschaft für Didaktik der Mathematik. WTM.
https://doi.org/10.37626/GA9783959873307.0

Werkzeuge beeinflussen können, wie Lernende mit visuellen mathematischen Darstellungen interagieren (Sedig & Sumner, 2006), ist der Einsatz von GeoGebra ebenfalls eine potenzielle Ursache für Moduseffekte.

Inhaltliche Itemmerkmale umfassen die mathematischen Anforderungen, die ein Item an die Testperson stellt. Analysen von PISA- und TIMSS-Daten legen nahe, dass Moduseffekte in bestimmen Sachgebieten oder im Zusammenhang mit bestimmten prozessbezogenen Kompetenzen auftreten könnten. So zeigten sich höhere Itemschwierigkeiten für TBA-Items im Vergleich zu PPA-Items bei geometrischen und stochastischen Items sowie bei Items, die den Umgang mit mathematischen Darstellungen oder das Problemlösen erforderten (Jentsch et al., 2024; Keng et al., 2008).

Forschungsinteresse und Studiendesign

Mathematikspezifische Untersuchungen von Moduseffekten, die potenzielle Ursachen berücksichtigen und so konkrete Hinweise für Designentscheidungen geben können, sind bislang kaum vorhanden. Da für VERA-Erhebungen ein vollständiger Umstieg von PPA auf TBA geplant ist, wurden in den letzten Jahren Anstrengungen unternommen, eine gleichwertige technologiegestützte Version der Vergleichsarbeiten bereitzustellen. In einigen Ländern wie Berlin wird VERA-8 entsprechend bereits in zwei Versionen realisiert. In dieser Studie untersuchen wir die Vergleichbarkeit der TBA- und PPA-Versionen, um Aufschluss über Itemmerkmale zu erlangen, die Moduseffekte im Fach Mathematik verursachen. Im Einzelnen fragen wir:

FF1: Inwieweit unterscheiden sich die VERA-8-Itemschwierigkeiten signifikant zwischen der technologiebasierten und der papierbasierten Version?

FF2: In welchen formal-technischen und inhaltlichen Itemmerkmalen unterscheiden sich verzerrte Items qualitativ?

Als Datenbasis stehen die Berliner VERA-8-Ergebnisse aus dem Jahr 2022 zur Verfügung. Der Mathematiktest umfasste in diesem Jahr 80 Items, die auf vier Testhefte verteilt waren. Jede Schule konnte sich eigenständig zwischen der PPA- und der TBA-Version entscheiden. Insgesamt bearbeiteten 6.785 Schülerinnen und Schüler die PPA-Version und 16.488 Schülerinnen und Schüler die TBA-Version des Tests. Die Testantworten wurden durch die jeweiligen Lehrkräfte dichotom als korrekt oder inkorrekt bewertet. Fehlende Antworten werden in dieser Studie als inkorrekte Antworten gewertet.

Für die Auswertung wurden die Daten mithilfe des R-Pakets eRm (Mair et al., 2006) in einem IRT-Raschmodell skaliert und hinsichtlich Differential Item Functioning (DIF) untersucht. Dabei wurden Itemschwierigkeiten (in logit) für die beiden Testversionen separat geschätzt und versionsabhängige

Unterschiede mithilfe des Likelihood-Ratio-Tests nach Andersen geprüft. Mithilfe des Wald-Tests wurden einzelne Items identifiziert, die Moduseffekte aufwiesen. Diese Items wurden anschließend im Rahmen einer qualitativen Inhaltsanalyse bezüglich ihrer formal-technischen sowie mathematischen Merkmale analysiert (Mayring, 2010). Entsprechend des Forschungsstandes umfasste die Analyse folgende deduktive Kategorien: das mathematische Sachgebiet, die erforderlichen prozessbezogenen Kompetenzen, das Antwortformat sowie die digitale Werkzeugnutzung. Ergänzend wurden Angaben zum Itemdesign (z. B. Eingabefeld, kariertes Papier) offen kodiert.

Ergebnisse

Der Likelihood-Ratio-Test bestätigte das Vorhandensein von verzerrten Items innerhalb des Tests und somit eine unterschiedliche Funktionalität der Items ($p<.001$). Der Wald-Test zeigte signifikante Unterschiede in der Itemschwierigkeit für 45 von 80 Items ($p<.05$). Da die Stichprobe sehr groß ist und der Wald-Test schnell signifikant wird, wurden die 16 Items für eine tiefergehende Analyse ausgewählt, die einen großen und damit substantiellen Unterschied in der Itemschwierigkeit ($\geq.05$ logit) aufwiesen. Die Hälfte dieser Items wies eine höhere Schwierigkeit in der PPA-Version auf.

Von den acht Items, die papierbasiert schwieriger waren, wurden sechs dem Sachgebiet Geometrie zugeordnet und sieben forderten die Kompetenz, mathematische Darstellungen zu verwenden. Vier der geometrischen Items nutzten GeoGebra, um eine Konstruktion durchführen zu lassen. Die offenen Kodierungen zum Itemdesign zeigten, dass zwei Items in der TBA-Version Eingabehilfen enthielten, und drei Items in der TBA-Version insofern stärker vorstrukturiert waren, als Achsen vollständig beschriftet oder karierte anstatt leere Hintergründe implementiert waren. Die Itemmerkmale der Items, die technologiebasiert schwieriger waren, waren sehr heterogen, sodass fast alle Sachgebiete, prozessbezogenen Kompetenzen und Antwortformate kodiert wurden. Bei zwei der acht Items wurde das Antwortformat zumindest in Teilen verändert. Ein weiteres Item forderte die Eingabe von Sonderzeichen. Eine Gemeinsamkeit der Items war jedoch, dass sie mehrschrittige Lösungen in Form von mehreren Problemlöse- oder Rechenschritten voraussetzten.

Diskussion

Unsere Ergebnisse zeigen, dass in der VERA-8-Erhebung 2022 Moduseffekte im Sinne divergierender Itemschwierigkeiten auftreten. Anders als in früheren Studien ist dabei keine generelle Verschiebung der Itemschwierigkeiten zu erkennen (Fishbein et al., 2018; Jerrim et al., 2018). Vielmehr sind einige Items schwieriger in PPA und andere in TBA. Die qualitative Analyse der auffälligen Items deutet darauf hin, dass das spezifische Design und das

Antwortformat von TBA- und PPA-Items sowie die erforderlichen Kompetenzen Moduseffekte verursachen können. Während geometrische Konstruktionsitems, die technologiebasiert mithilfe von GeoGebra umgesetzt werden, leichter in TBA waren, waren mehrschrittige Items, die das Operieren und Problemlösen fordern, tendenziell leichter in PPA. Inwiefern sich die Unterschiede durch Lösungsstrategien der Lernenden wie z. B. der Nutzung von Schmierpapier oder der dynamischen Manipulation in GeoGebra erklären lassen, soll in Folgestudien untersucht werden. Zusammenfassend unterstreichen die Ergebnisse die Bedeutung einer sorgfältigen Gestaltung technologiegestützter Mathematiktests, um die Vergleichbarkeit zu gewährleisten.

Literatur

Buerger, S., Kroehne, U., Koehler, C. & Goldhammer, F. (2019). What makes the difference? The impact of item properties on mode effects in reading assessments. *Studies in Educational Evaluation, 62*, 1–9. https://doi.org/10.1016/j.stueduc.2019.04.005

Fishbein, B., Martin, M. O., Mullis, I. V. S. & Foy, P. (2018). The TIMSS 2019 Item Equivalence Study: examining mode effects for computer-based assessment and implications for measuring trends. *Large-scale Assessments in Education, 6*(1), 11. https://doi.org/10.1186/s40536-018-0064-z

Jentsch, A., Beese, C. & Schwippert, K. (2024). Papier- oder computerbasierte Kompetenztests? Eine Generalisierbarkeitsstudie zu Moduseffekten in Deutschland im Rahmen von TIMSS 2019. *Journal for Educational Research Online, 2023(*1), 30–50. https://doi.org/10.31244/jero.2023.01.02

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C. & McKeown, C. (2018). PISA 2015: how big is the 'mode effect' and what has been done about it? *Oxford Review of Education, 44*(4), 476–493.

Keng, L., McClarty, K. L. & Davis, L. L. (2008). Item-Level Comparative Analysis of Online and Paper Administrations of the Texas Assessment of Knowledge and Skills. *Applied Measurement in Education, 21*(3), 207–226. https://doi.org/10.1080/08957340802161774

Kroehne, U. & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft, 14*(S2), 169–186. https://doi.org/10.1007/s11618-011-0185-4

Mair, P., Rusch, T., Hatzinger, R. & Maier, M. J. (2006). CRAN: Contributed Packages. https://doi.org/10.32614/CRAN.package.eRm

Mayring, P. (2010). *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. Beltz.

Sedig, K. & Sumner, M. (2006). Characterizing Interaction with Visual Mathematical Representations. *International Journal of Computers for Mathematical Learning, 11*(1), 1–55. https://doi.org/10.1007/s10758-006-0001-z

Wang, S., Jiao, H., Young, M. J., Brooks, T. & Olson, J. (2007). A Meta-Analysis of Testing Mode Effects in Grade K-12 Mathematics Tests. *Educational and Psychological Measurement, 67*(2), 219–238. https://doi.org/10.1177/0013164406288166