WOLFERT, Carl; NEUMANN, Irene & SOMMERHOFF, Daniel
Kiel

Randomisierte Aufgaben für E-Assessments - Empirische Prüfung eines Schwierigkeitsmodells

Elektronische Assessments (E-Assessments) gehören zu den zentralen Themen, die, insbesondere im Hochschulwesen, im Zeitalter der Digitalisierung adressiert werden müssen (Laaziz et al., 2024). Das Bereitstellen vieler ähnlicher oder zumindest in ihrer Schwierigkeit vergleichbarer mathematischer Aufgaben ist dabei für E-Assessments aus mehreren Gründen von Bedeutung: So können entsprechende Aufgaben eingesetzt werden, um parallele Tests zu generieren und so Plagiate bzw. Abschreiben zu erschweren oder eine mehrfache Durchführung eines Tests durch die gleiche Person zu ermöglichen (Ashton et al., 2005; Rowlett, 2014; Sangwin, 2013). Gleichzeitig ermöglicht die Bereitstellung ähnlicher Aufgaben individuelle Lerntempi, da sie Lernenden die Möglichkeit bieten, Inhalte ggfs. mehrfach an ähnlichen Aufgaben einzuüben (Jordan, 2014). Aufgrund der beispielhaft genannten Vorteile ist es bereits gängige Praxis, zufällige Versionen von Aufgaben samt zugehöriger Musterlösungen zu generieren (Sangwin & Köcher, 2016). Dem liegt die Annahme zugrunde, dass diese von vergleichbarer Schwierigkeit sind (Gallaun et al., 2022). Bislang fehlt es jedoch an systematischer Forschung zu Randomisierungen. Daher befasst sich der folgende Beitrag mit der Entwicklung und Validierung eines Modells, welches der Systematisierung der Schwierigkeit von Randomisierungen in E-Assessments dient.

Theoretisches Framework

Als randomisierbare Elemente einer Aufgabe können unterschiedliche Aufgabenspezifika herangezogen werden, wie beispielsweise ihre sprachliche Formulierung und mathematisch variierbare Aspekte wie Zahlenwerte und Bezeichner (Sangwin, 2013; Schlager, 2020). Im Folgenden werden im Einklang mit etwa Sangwin (2013) ausschließlich zentrale mathematische Elemente betrachtet und elaboriert, während alle anderen Elemente konstant gehalten werden. Diese Entscheidung beruht u. a. darauf, dass Zahlenwerte den zur Lösung notwendigen Rechenaufwand maßgeblich beeinflussen, der wiederum eine mögliche Grundlage zur Systematisierung von Randomisierungen bietet (Sangwin, 2013).

Zur Beschreibung von Randomisierungen werden folgende Begriffe genutzt: Aus einer vorgegebenen Aufgabe ("Löse $x^2 - 3x + 2 = 0$") können Sammlungen möglicher Instanzen generiert werden, die zentrale Elemente der Aufgabe mittels Parametern und zugehöriger Parameterräume ("Löse $ax^2 + bx + c = d$" mit $a, b, c, d \in \mathbb{R}$) verallgemeinern. Konkrete Aufgaben, die

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

aus einer solchen Sammlung möglicher Instanzen generiert werden könnten, heißen Instanzen ("Löse $3x^2 - 2x + 3 = 5$").

Bei der Betrachtung von Randomisierungen und der Ähnlichkeit verschiedener Instanzen ist stets eine beliebige, jedoch feste Zielgruppe zu berücksichtigen. Andernfalls kann bereits eine einzelne Aufgabe bspw. unterschiedliche Lösungsquoten in verschiedenen Zielgruppen aufweisen. Dies ist insbesondere relevant, da die empirische Schwierigkeit bzw. Lösungsquote das meistverwendete Kriterium zur Operationalisierung der Ähnlichkeit von Instanzen darstellt. Aufbauend auf diesen Überlegungen und Vorarbeiten basiert das hier entwickelte Modell auf zwei zentralen Definitionen:

Konzeptäquivalenz (KÄ): Zwei mathematische Aufgaben werden als konzeptäquivalent bezeichnet, wenn sie einen bis auf Variationen in Zahlenwerten bzw. Bezeichnern identischen Lösungsweg besitzen.

Rechenaufwandsäquivalenz (RAÄ): Zwei mathematische Aufgaben werden als äquivalent in Hinblick auf den zu ihrer Lösung notwendigen Rechenaufwand bezeichnet, wenn sie von einer Zielgruppe dasselbe approximative Level an rechnerischem Aufwand erfordern.

Da es wenig sinnvoll erscheint, den Rechenaufwand beliebiger Aufgaben miteinander vergleichen zu wollen, wird die Definition von RAÄ ausschließlich im Falle von konzeptäquivalenten Instanzen verwendet. Zwei Instanzen heißen vergleichbar, falls sie konzeptäquivalent sind und eine vergleichbare empirische Schwierigkeit, d. h. Lösungsquote, besitzen.

Fragestellung

Das skizzierte Framework beschreibt Ansätze zur strukturierten Randomisierung von Aufgaben und definiert unterschiedliche Arten von Randomisierungen, die aus theoretischer Perspektive Aufgaben vergleichbarer empirischer Schwierigkeit hervorbringen sollten. Zur Validierung des Modells werden die folgenden Forschungsfragen fokussiert: 1. Inwiefern führen (i) konzeptäquivalente Instanzen mit größeren Zahlenwerten bzw. (ii) Instanzen mit einem zusätzlichen zur Lösung notwendigen Konzept zu geringeren Lösungsquoten? 2. Besitzen Instanzen, die konzept- und rechenaufwandsäquivalent sind, vergleichbare Lösungsquoten?

Methodik

In einer Querschnittstudie bearbeiteten $N = 105$ Mathematikstudierende zu Beginn des ersten Semesters je eine von fünf Testversionen, die gemäß des obigen Frameworks generierte Instanzen mathematischer Aufgaben enthielten. Dabei wurde die RAÄ operationalisiert, indem vier in ihrer

Schwierigkeit ordinal steigende RAÄ-Subkategorien verwendet wurden, deren Abgrenzung durch konkrete Intervalle natürlicher Zahlen realisiert wurde. Die Papierfragebögen wurden randomisiert zugeordnet. Die Ergebnisse der Studierenden wurden dichotom codiert und die Interrater-Reliabilitäten (Fleiss' κ über drei Rater) je Aufgabe berechnet. Letztere waren meistens nahezu perfekt. Die Daten wurden anschließend ausgewertet, um die Aufgaben bzw. ihre Lösungsquoten hinsichtlich der Forschungsfragen zu untersuchen.

Erste Ergebnisse

Die Ergebnisse stützen die durch das Framework gestellten Hypothesen tendenziell. Sie deuten darauf hin, dass größere Zahlenwerte in Rechnungen die empirische Schwierigkeit von konzeptäquivalenten Instanzen steigern. Ebenso zeigt sich, dass die Hinzunahme eines weiteren zur Lösung notwendigen Konzepts in der Regel zu einer geringeren Lösungsquote führt. Darüber hinaus weisen die Ergebnisse darauf hin, dass konzept- und rechenaufwandsäquivalente Instanzen gemäß (der Operationalisierung) des Frameworks eine vergleichbare empirische Schwierigkeit besitzen.

Diskussion

Die Analysen belegen, dass die Eigenschaften RAÄ und KÄ einen nennenswerten Einfluss auf die empirische Schwierigkeit von möglichen Instanzen einer Randomisierung haben. Dass die empirische Schwierigkeit über die vier verwendeten RAÄ-Subkategorien durch größere Zahlenwerte nur tendenziell ansteigt und nicht jede Änderung sich perfekt in den empirischen Lösungsquoten widerspiegelt, war bereits aufgrund der gewählten Operationalisierung zu erwarten, da das Kontinuum des Rechenaufwands diskretisiert wurde. Ferner weisen einige Aufgaben Abweichungen zwischen den tatsächlichen und den gemäß des Frameworks erwarteten Lösungsquoten auf. Dies inkludiert etwa Abweichungen der beschriebenen Form zwischen den RAÄ-Subkategorien, aber auch Abweichungen in Lösungsquoten von Instanzen innerhalb einer RAÄ-Subkategorie. Diese Abweichungen deuten darauf hin, dass weitere Eigenschaften der verwendeten Instanzen, die nicht durch die Kategorien RAÄ und KÄ abgebildet werden, einen großen Einfluss auf die tatsächliche empirische Schwierigkeit haben können. Eine detaillierte Analyse dieser Abweichungen könnte aufschlussreich sein und steht noch aus.

Obige Resultate stehen im Einklang mit theoretischen Überlegungen von Sangwin (2013), der konstatierte, dass eine allgemeine Erfassung von Aspekten wie Rechenaufwand für beliebige Instanzen einer Randomisierung problematisch sein kann, selbst wenn sie auf gewisse Weise konzeptäquivalent sind. Ferner wurden die Annahmen von Sangwin (2013) durch das

Framework weiter konkretisiert und im Feld evaluiert. Dies könnte dazu beitragen, offene Fragen zum Thema E-Assessment zu beantworten, wie sie etwa Kinnear et al. (2022) formuliert haben.

Insgesamt liefern die Daten empirische Hinweise darauf, welche Aspekte bei der Systematisierung von Randomisierungen entscheidend sein könnten und daher bei der Implementierung solcher, etwa in E-Assessments, berücksichtigt werden sollten. Dabei ist zu beachten, dass die dargelegte Untersuchung auf Papier durchgeführt wurde. Daher stehen Untersuchungen einer praxisnahen Generalisierbarkeit auf E-Assessments aus, wenngleich die Ergebnisse dieses Beitrags einen relevanten Ansatzpunkt für weitere Forschung bieten. Zudem könnten eine detailliertere Analyse von KÄ und insbesondere RAÄ und weitere Studien hierzu, sowie zu weiteren in diesem Beitrag nicht fokussierten zentralen randomisierbaren Aufgabencharakteristika, wichtige zusätzliche Erkenntnisse liefern.

Literatur

Ashton, H. S., Beevers, C. E., Korabinski, A. A., & Youngson, M. A. (2005). Incorporating partial credit in computer-aided assessment of Mathematics in secondary education. *British Journal of Educational Technology, 37*(1), 93-119.

Gallaun, D., Kruse, K., & Seifert, C. (2022). High-quality tasks for e-assessment in mathematics. *Teaching Mathematics and its Applications: An International Journal of the IMA, 41*(4), 270-279.

Jordan, S. E. (2014). *E-assessment for learning? Exploring the potential of computer-marked assessment and computer-generated feedback, from short-answer questions to assessment analytics* (PhD thesis). The Open University. https://oro.open.ac.uk/41115/

Kinnear, G., Jones, I., Sangwin, C., Alarfaj, M., Davies, B., Fearn, S., Foster, C., Heck, A., Henderson, K., Hunt, T., Iannone, P., Kontorovich, I., Larson, N., Lowe, T., Meyer, J. C., O'Shea, A., Rowlett, P., Sikurajapathi, I., & Wong, T. (2022). A Collaboratively-Derived Research Agenda for E-assessment in Undergraduate Mathematics. *International Journal of Research in Undergraduate Mathematics Education, 10*(1), 201-231.

Laaziz, Y., Chemsi, G., & Radid, M. (2024). The Influence of E-Assessment on Students' Cognitive Engagement in Higher Education. *International Journal of Engineering Pedagogy, 14*(4), 54-67.

Rowlett, P. (2014). Development and evaluation of a partially-automated approach to the assessment of undergraduate mathematics. *British Congress of Mathematics Education, 8*, 295-302.

Sangwin, C. J. (2013). *Computer Aided Assessment of Mathematics*. Oxford University Press.

Sangwin, C. J., & Köcher, N. (2016). Automation of mathematics examinations. *Computers & Education, 94*, 215-227.

Schlager, S. (2020). *Zur Erforschung des Zusammenhangs zwischen Sprachkompetenz und Mathematikleistung*. Springer Nature.