Computerisierte adaptive Klausuren im Psychologiestudium

Aron Fink, Patrick Naumann und Christoph König

Goethe-Universität Frankfurt am Main

In ihrem Positionspapier stellen Lindner, Sparfeldt, Köller, Lukas und Leutner verschiedene Maßnahmen zur Verbesserung von schriftlichen Prüfungen im Psychologiestudium dar. Wie die Autor_innen sehen auch wir die Klausurdurchführung am Computer als eine entscheidende Stellschraube an. Computerbasierte Prüfungen ermöglichen, neben dem Einsatz innovativer Klausurbau- steine und einer ökonomischen Klausurkonstruktion, auch die Nutzung aktueller Ansätze der Psychologischen Diagnostik. Ein besonderes Potential sehen wir im computerisierten adaptiven Testen (CAT; z.B. Frey, 2020). Der Einschätzung von Lindner et al., wonach CAT für Klausuren ungeeignet sei, möchten wir mit diesem Beitrag entgegenreten.

Bei adaptiven Klausuren werden nur solche Items vorgelegt, die möglichst viel diagnostische Information über das individuelle Kompetenzniveau liefern. Daraus resultiert, im Vergleich zu nicht-adaptiven Tests, eine Steigerung der Messpräzision (z.B. Segall, 2005). Diese ist für die Qualitätssteigerung von Klausuren in zweierlei Hinsicht von Bedeutung:

(1) CAT führt zu einer erheblichen Steigerung der Messeffizienz, wodurch die im Positionspapier genannte notwendige hohe Itemanzahl bei geschlossenen Antwortformaten deutlich reduziert werden kann. Darüber hinaus

wirkt die verkürzte Testlänge, ebenso wie eine kompetenzangemessene Itemvorgabe, auch den im Positionspapier genannten Ermüdungseffekten entgegen. Die individualisierte Itemvorgabe erschwert zudem das Bekanntwerden von Items und bietet insgesamt eine höhere Test-sicherheit.

(2) Mit CAT kann die Messpräzision über die gesamte Kompetenzverteilung angeglichen werden (z.B. Frey & Ehmke, 2007). Die Messpräzision üblicher Klausuren ist im mittleren Bereich der Kompetenzverteilung am höchsten, und fällt in den Randbereichen erheblich geringer aus (Dolan & Burling, 2017). Die Genauigkeit der Klausurbewertung ist daher bei Studierenden mit sehr geringer oder sehr hoher Kompetenz häufig deutlich niedriger. Gerade in diesen Kompetenzbereichen sind die Bewertungen jedoch von besonderer Relevanz, da von diesen beispielsweise die Weiterführung des Studiums (z.B. drohende Exmatrikulation durch zu viele Fehlversuche, Bewerbung um einen Platz in einem konsekutiven Master-Studiengang) oder eine Studienförderung abhängt.

Das Positionspapier führt folgende Implementations- hürden für adaptive Klausuren auf: (1) Die Erstellung hinreichend großer, kalibrierter Itempools, (2) die Vergleichbarkeit der Prüfungsvoraussetzungen für Studierende bei der Bearbeitung individualisierter Klausuren,

unter anderem durch das Vorhandensein von (3) Itempositionseffekten und schließlich (4) die Begrenzung auf automatisch auswertbare Antwortformate.

Zu (1). Die Kalibrierung von Itempools für adaptive Klausuren ist aufgrund der an Hochschulen häufig vorzufindenden kleinen Stichproben oder fehlenden Ressourcen zur Konstruktion eines großen Itempools vor der ersten Klausuranwendung durchaus eine Herausforderung. Ist die im Positionspapier angesprochene hochschulübergreifende Zusammenarbeit nicht möglich, können durch eine kontinuierliche Kalibrierungsstrategie (KKS; Fink, Born, Spoden & Frey, 2018) Itempools sukzessive über mehrere Klausuranwendungen aufgebaut werden, wobei die Berichtsmetrik beibehalten wird und defizitäre und bekanntgewordene Items statistisch identifiziert werden können. Dadurch wird der Itempool über Klausuranwendungen hinweg stetig größer und die Klausur immer adaptiver. Der Aufwand zum Aufbau einer adaptiven Klausur unter Nutzung eines solchen Vorgehens unterscheidet sich dann nur geringfügig von dem einer nicht-adaptiven computerbasierten Klausur.

Zu (2). Der Grundsatz der Chancengleichheit der Prüflinge ist nicht gleichbedeutend mit einer absoluten Gleichbehandlung (Niehues, Fischer & Jeremias, 2018). Auch bei mündlichen Prüfungen ist es üblich, dass Prüfer die Schwierigkeit ihrer Fragen an das Antwortverhalten der Prüflinge anpassen. Bei adaptiven Klausuren wird Gleichheit geschaffen, indem Items nach transparenten Regeln auf Basis der individuellen Kompetenzausprägung ausgewählt werden (für eine prüfungsrechtliche Einordnung adaptiver Klausuren siehe Frey, Spoden, Fink & Born, 2020).

Zu (3). Itempositionseffekte sind für adaptive Klausuren potentiell problematisch, da sie bei Nicht-Beachtung zu verzerrten Kompetenzschätzungen führen können (z.B. Frey, Bernhardt & Born, 2017). Mit balancierten Testdesigns kann jedoch für mögliche Itempositionseffekte kontrolliert werden. Die balancierte Erweiterung der KKS (Frey & Fink, im Druck) illustriert, wie dies im Rahmen von CAT realisiert werden kann.

Zu (4). Auch offene Antwortformate können in adaptive Klausuren eingebunden werden. Sie werden auf Grundlage der vorläufigen Kompetenzschätzung ausgewählt, und die gegebenen Antworten werden gespeichert. Diese Antworten liefern zwar keine Information für die Auswahl des nächsten Items, können aber nach der Klausurdurchführung ausgewertet werden.

CAT bietet also das Potenzial, schriftliche Prüfungen hinsichtlich ihrer Messgenauigkeit und Fairness deutlich zu optimieren. Aus den genannten Gründen werden sie in den USA beispielsweise seit vielen Jahren im high-stakes Bereich erfolgreich eingesetzt (z.B. NCLEX Prüfung für Krankenschwestern und -Pfleger; <https://www.ncsbn.org/>

nclex.htm). Die im Positionspapier dargestellten Implementationshürden sind aus unserer Sicht mit überschaubarem Aufwand zu überwinden. Mit der frei verfügbaren KAT-HS-App (Fink, Spoden, Frey & Naumann, im Druck) existiert zudem eine auf den deutschen Hochschulkontext abgestimmte Software, mit der psychometrisch fundierte adaptive und nicht-adaptive Klausuren konfiguriert, administriert und ausgewertet werden können.

Wir plädieren dafür, die Potentiale computerbasierter Klausuren in bestmöglicher Weise zu nutzen. Dies beinhaltet neben der Integration innovativer multimedialer und simulationsbasierter Itemformate auch die Nutzung leistungsfähiger psychometrischer Methoden.

Literatur

- Dolan, R. P. & Burling, K. S. (2017). Computer-based testing in higher education. In C. Secolsky & D. B. Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education*, (2nd ed., pp. 370–384). New York, NY: Routledge. <https://www.routledgehandbooks.com/doi/10.4324/9781315709307.ch24>
- Fink, A., Born, S., Spoden, C. & Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling*, 60, 327–346. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam_3-2018_327-346.pdf
- Fink, A., Spoden, C., Frey, A. & Naumann, P. (im Druck). Kriteriumsorientiertes adaptives Testen mit der KAT-HS-App. *Diagnostica*.
- Frey, A. (2020). Computerisiertes adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3. Aufl., S. 501–525). Berlin: Springer. https://doi.org/10.1007/978-3-662-61532-4_20
- Frey, A., Bernhardt, R. & Born, S. (2017). Umgang mit Itempositionseffekten bei der Entwicklung computerisierter adaptiver Tests. *Diagnostica*, 63, 167–178. <https://doi.org/10.1026/0012-1924/a000173>
- Frey, A. & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. *Zeitschrift für Erziehungswissenschaft, Sonderheft 8*, 169–184. https://doi.org/10.1007/978-3-531-90865-6_10
- Frey, A. & Fink, A. (im Druck). Controlling for item position effects when adaptive testing is used in Large-Scale Assessments. In L. Khorramdel, M. von Davier, & K. Yamamoto (Eds.), *Innovative computer-based international large-scale assessments – foundations, methodologies and quality assurance procedures*. Springer
- Frey, A., Spoden, C., Fink, A. & Born, S. (2020). Kompetenzorientierte individualisierte Hochschulklausuren und deren prüfungsrechtliche Einordnung. *elead*, 13. urn:nbn:de:0009-5-51197
- Niehues, N., Fischer, E. & Jeremias, C. (2018). *Prüfungsrecht* (7. Aufl.). München: Beck.
- Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement*. Amsterdam: Elsevier.

Interessenskonflikt

Der korrespondierende Autor erklärt im Namen aller Autoren, dass kein Interessenkonflikt vorliegt.

Autorenschaft

Aron Fink, Pädagogische Psychologie mit Schwerpunkt Beratung, Diagnostik und Evaluation, Goethe-Universität, a.fink@psych.uni-frankfurt.de; Patrick Naumann, Pädagogische Psychologie mit Schwerpunkt Beratung, Diagnostik und Evaluation, Goethe-Universität, naumann@psych.uni-frankfurt.de; Christoph König, Pädagogische Psychologie mit Schwerpunkt Beratung, Diagnostik und Evaluation, Goethe-Universität, koenig@psych.uni-frankfurt.de. Aron Fink und Patrick Naumann teilen sich die Erstautorenschaft.

Aron Fink, MSc.

Pädagogische Psychologie mit Schwerpunkt Beratung, Diagnostik und Evaluation
Goethe-Universität
Theodor-W.-Adorno-Platz 1
60323 Frankfurt am Main
a.fink@psych.uni-frankfurt.de

<https://doi.org/10.1026/0033-3042/a000532>