Softwareinformation

Kriteriumsorientiertes adaptives Testen mit der KAT-HS-App

Aron Fink[1], Christian Spoden[2], Andreas Frey[1,3] und Patrick Naumann[1]

[1]Institut für Psychologie, Goethe-Universität Frankfurt, Frankfurt am Main
[2]Deutsches Institut für Erwachsenenbildung, Leibniz-Zentrum für Lebenslanges Lernen e. V., Bonn
[3]Centre for Educational Measurement (CEMO), University of Oslo, Norwegen

Zusammenfassung: In dieser Softwareinformation werden die Möglichkeiten zur Konstruktion, Administration und Auswertung kriteriumsorientierter, computerisierter adaptiver und nicht-adaptiver Tests mit der R-basierten open-source KAT-HS-App erläutert. Die App ermöglicht unter anderem auch die Anwendung der kontinuierlichen Kalibrierungsstrategie von Fink, Born, Spoden und Frey (2018).

Schlüsselwörter: R-Software, computerbasiertes Testen, computerisiertes adaptives Testen, kontinuierliche Kalibrierung, Testentwicklung′

Criterion-Referenced Adaptive Tests Using the KAT-HS App

Abstract: This software demonstration presents the possibilities for the construction, administration, and evaluation of criterion-referenced, computerized adaptive and nonadaptive tests with the R-based open-source KAT-HS app. This app enables users to apply the continuous item calibration strategy of Fink, Born, Spoden, and Frey (2018).

Keywords: R software, computer-based testing, computerized adaptive testing, continuous item calibration, test development

Die Nutzung digitaler Technologien im Rahmen der psychologischen Diagnostik ermöglicht den Einsatz innovativer Itemformate und hat das Potential die Effizienz des Testprozesses erheblich zu steigern. Zudem bietet der Einsatz computerbasierter Verfahren die Möglichkeit, verschiedene leistungsfähige psychometrische Methoden in die Testpraxis zu integrieren. Besonders hervorzuheben sind hier Methoden auf Basis von Modellen der Item-Response-Theory (IRT; z. B. van der Linden, 2016). Die einzelnen Methoden sind gut untersucht und dokumentiert. Abseits der Grundlagenforschung und Anwendungen bei groß angelegten Vergleichsstudien wie PISA, IGLU oder TIMSS findet man jedoch nach wie vor kaum IRT-basierte Tests, obgleich sie das Potential haben in zahlreichen weiteren Anwendungsbereichen die Qualität von Testungen erheblich zu steigern. Ein Beispiel für einen solchen Anwendungsbereich mit hohem Testaufkommen sind Hochschulklausuren. Hier existiert eine deutliche Lücke zwischen psychometrischem Kenntnisstand und Testpraxis. Problematisch an derzeitigen Hochschulklausuren sind nach Spoden, Frey, Fink und Naumann (2020) im Wesentlichen vier Aspekte: 1. Lernziele werden nicht angemessen durch die genutzten Aufgaben operationalisiert. 2. Klausuren sind nicht als kriteriumsorientierte Verfahren konzipiert, so dass Ergebnisse nicht als Ausmaß des Erreichens von Lernzielen interpretiert werden können. 3. Testzeitpunkte werden nicht statistisch verlinkt, so dass die Unabhängigkeit der Ergebnisse von Kohortenleistungsfähigkeit und Klausurschwierigkeit nicht gewährleistet ist. 4. Die Messpräzision schwankt über den Merkmalsbereich mit typischerweise deutlich niedrigerer Messpräzision an den Rändern der Kompetenzverteilung. Spoden und Frey (im Druck) beschreiben in ihrem Konzept psychometrisch fundierter Hochschulklausuren, wie diesen Problemen durch die zielgerichtete Nutzung IRT-basierter Methoden begegnet werden kann. Ausgehend von diesem Konzept wurde die hier vorgestellte KAT-HS-App (KAT-HS = kriteriumsorientiertes adaptives Testen in der Hochschule) entwickelt. In ihrer Anwendbarkeit ist die App allerdings nicht auf Hochschulklausuren beschränkt. Vielmehr hat sie zum Ziel, IRT-basierte Methoden für neue Anwendungsbereiche zu erschließen. Hierfür werden verschiedene Elemente, für deren Anwendung normalerweise unterschiedliche Softwarepakete einzusetzen sind, in einem Programm gebündelt. Ihre Kernelemente sind: a) computerbasierte Testadministration; b) IRT-Skalierung; c) Methoden zur Überprüfung der psychometrischen Qualität des Tests; d) Online-Kalibrierung bei wiederholten Anwendungen eines Tests; e) Kontrolle von Itempositionseffekten (IPE);

f) computerisiertes adaptives Testen (z. B. Frey, 2020); und g) kriteriumsorientiertes Testen. Zudem setzt die App auf eine intuitive Benutzerführung mit sinnvollen Voreinstellungen, die es dem Anwender möglichst einfach macht IRT-basierte Tests auf professionelle Weise einzusetzen. Das Ziel ist somit die Zugänglichkeit IRT-basierter Methoden zu verbessern und gleichzeitig die methodisch angemessene Nutzung sicherzustellen. Die psychometrischen Grundlagen werden im Konzept von Spoden und Frey (im Druck) beschrieben, zu dem ein Workshop kostenfrei als Video abgerufen werden kann (https://kat-hs.uni-frankfurt.de/materialien/workshop/). Die KAT-HS-App ist eine in R programmierte Shiny-App (Chang, Cheng, Allaire, Xie & McPherson, 2019) und somit kostenfrei und open-source. Sie greift neben eigens programmierten Funktionen auf Routinen der R-Pakete mirt (Chalmers, 2012), mirtCAT (Chalmers, 2016) und equateIRT (Battauz, 2015) zurück. Die App ist nach Registrierung über die Website https://kat-hs.uni-frankfurt.de/materialien/software/ für Forschung und Lehre kostenfrei erhältlich. In der jetzigen Form ist die App ausschließlich für die Verwendung auf Windowssystemen geeignet. Im Folgenden werden die grundlegenden Funktionalitäten der App vorgestellt. Eine ausführlichere Dokumentation der KAT-HS-App ist im zugehörigen Benutzerhandbuch zu finden.

Testkonstruktion

Die Nutzung der KAT-HS-App setzt einen Itempool voraus, der ein eindimensionales Merkmal inhaltsvalide abbildet. Die Items sollen zudem konform mit den Annahmen des verwendeten IRT-Modells sein. Inwieweit dies zutrifft, wird bei der Nutzung der App berechnet und Vorschläge zum Umgang mit nicht hinreichend fittenden Items gegeben. Die Iteminformationen (z. B. Item-ID, Stimulus, Antwortoptionen, Inhaltsbereich, etc.) müssen in einer Itemdatenbank im XLSX-Format zusammengestellt werden. Durch die Möglichkeit, Itemstämme und etwaige Antwortoptionen in HTML-Code zu spezifizieren, existieren zahleiche Möglichkeiten der Itemgestaltung. So ist beispielsweise auch das Einbinden von Multimediadateien möglich. Es können automatisch auswertbare und von Hand zu kodierende Antwortformate genutzt werden. Die App kann nur mit dichotom bewerteten Items arbeiten. Existierende Itemparameterschätzungen können zu Beginn in der Itemdatenbank eingefügt oder bei wiederholten Testungen im laufenden Betrieb durch Nutzung der kontinuierlichen Kalibrierungsstrategie (KKS; Fink, Born, Spoden & Frey, 2018; Frey & Fink, in press) ergänzt werden.

Nach Import der Itemdatenbank können verschieden komplexe Testarten mit der App erstellt werden. Vom einfachen linearen Test mit nur einer Testversion bis zum volladaptiven Test mit KKS inklusive Kontrolle von IPE und kriteriumsorientierter Testwertinterpretation ergeben sich zahlreiche Möglichkeiten der Testzusammenstellung. Hierfür können in der App noch weitere Einstellungen getroffen werden (z. B. Testlänge, Bearbeitungszeit, Anzahl Itemcluster, Itemauswahlkriterium, Personenparameterschätzer, Content-Balancing, etc.). Für den Fall von wiederkehrenden Testungen ist die KKS in die App implementiert. Mit dieser können Items im laufenden Testbetrieb ergänzt und kalibriert werden. Die KKS weist folgende Kernelemente auf: (a) Nutzung von Itemantworten mehrerer Testanwendungen zur Kalibrierung mit einem IRT-Modell, (b) Beibehaltung der Berichtsmetrik über Testungen, (c) ansteigende Adaptivität und Präzision der Fähigkeitsschätzungen über Testungen sowie (d) Kontrolle von Itemparameter Drift (IPD) und (e) IPE. In der KKS sind drei Arten von Itemclustern zu unterscheiden. Das *adaptive Cluster*, welches Items enthält, die im Testverlauf adaptiv gewählt werden und der Erhöhung der Messpräzision dienen; das *Kalibrierungscluster*, welches neue Items ohne Schätzung enthält, die der Vergrößerung des Itempools dienen; schließlich das *Linking-Cluster*, welches Items enthält, die zum Linking zweier aufeinander folgender Testzyklen genutzt werden. Die Auswahl der Linkitems kann anhand bestimmter Kriterien (z. B. Schwierigkeitsverteilung, Inhaltsbereich) automatisiert durch die App erfolgen. Je nach Charakteristika des Itempools, gewünschtem Grad an Adaptivität, Anzahl neuer Items, et cetera können in den Testeinstellungen mehr oder weniger der entsprechenden Cluster spezifiziert und so verschiedene Testarten erstellt werden. In Tabelle 1 sind die verschiedenen Testarten dargestellt, die durch die Kombination der drei Clusterarten konstruiert werden können.

Sind mehr als ein Itemcluster angegeben, erstellt die KAT-HS-App verschiedene Testversionen. Als Testdesign wird die Struktur eines balancierten lateinischen Quadrats (Williams, 1949) genutzt, um die Balancierung der Clusterpositionen und der Clusterreihenfolgen über Personen zu erreichen. Dies erhöht zum einen die Testsicherheit und ermöglicht zum anderen die Ausbalancierung von IPE (Frey, Bernhardt & Born, 2017) und Carry-Over-Effekten erster Ordnung auf der Ebene von Item-Clustern.

Zusätzlich muss eine Tabelle mit den Personeninformationen (z. B. Personen-ID, Login-Daten, Testversion) in die App importiert werden.

Tabelle 1. Konfiguration unterschiedlicher Testarten mit der KAT-HS-App

Testart	Anzahl adaptiver Cluster	Anzahl Kalibrierungs-cluster	Anzahl Linking-cluster
Einzelne / erste Anwendung nicht-adaptiver Test	= 0	≥ 1	= 0
Wiederkehrende Anwendung nicht-adaptiver Test ohne Vergrößerung des Itempools	= 0	= 0	≥ 1
Verlinkung + Vergrößerung des Itempools	= 0	≥ 1	≥ 1
KKS ohne Vergrößerung des Itempools	≥ 1	= 0	≥ 1
KKS	≥ 1	≥ 1	≥ 1
Volladaptiver Test (Abbruchkriterium = Testlänge)	= 1	= 0	= 0

Anmerkung: KKS = Kontinuierliche Kalibrierungsstrategie.

Testadministration

Der Test kann im Standardbrowser des Testcomputers oder, für Szenarien die einen höheren Grad an Sicherheit benötigen, im Safe Exam Browser (ETH Zürich, 2019) gestartet werden. Letzterer versetzt den Computer in einen sogenannten Kioskmodus, welcher den Zugriff auf Hilfsmittel (z. B. externe Programme, Webseiten) einschränkt oder unterbindet. Der Safe Exam Browser muss vorher auf dem Testcomputer installiert und konfiguriert sein. Nach Beendigung des Tests speichert die Testanwendung die individuellen Ergebnisse als RData-Files. In diesen sind alle relevanten Informationen (bearbeitete Items, Rohantworten, kodierte Antworten, Bearbeitungszeiten etc.) gespeichert und können bei Bedarf in R eingesehen und weiterbearbeitet werden.

Auswertung

Schritt 1: Daten zusammenführen

Für die folgenden Auswertungsschritte benötigt die KAT-HS-App eine Antwortmatrix mit den dichotom kodierten Antworten der Testeilnehmer. Die App bietet die Möglichkeit zur automatisierten Erstellung einer solchen Matrix. Diese Antwortmatrix kann anschließend für die IRT-Skalierung genutzt werden. Zudem können die Antwortmatrizen aufeinander folgender Testanwendungen zusammengeführt sowie eine Tabelle mit den Rohantworten der Testpersonen exportiert werden.

Schritt 2: IRT-Skalierung

Die Skalierung kann in der KAT-HS-App mit oder ohne Verankerung auf frühere Testzeitpunkten (freie und fixierte Skalierung) erfolgen. Die fixierte Skalierung ist für Testdesigns mit Linkitems geeignet. Hierbei werden die Parameterschätzungen der Linkitems auf die Werte des vorangegangenen Testzyklus fixiert und alle anderen Itemparameter frei geschätzt. So werden individuelle Testergebnisse auf derselben Metrik verortet und sind über die verschiedenen Testzeitpunkte hinweg direkt vergleichbar. Als Messmodell können die geläufigen logistischen Testmodelle mit einem (1PL) oder zwei Parametern (2PL) genutzt werden (van der Linden, 2016). Die Schätzung der Itemparameter erfolgt mit dem bei den meisten aktuellen IRT-Programmen genutzten Marginal-Maximum-Likelihood-Verfahren (Bock & Aitken, 1981). Es ist über einen breiten Anwendungsbereich einsetzbar, tätigt vergleichsweise wenige Annahmen, erlaubt die konsistente Itemparameterschätzung von ein- und mehrparametrischen IRT-Modellen und ist auch für unvollständige Kalibrierungsdesigns (wie z. B. unter Verwendung der KKS) geeignet.

Schritt 3: Überprüfung der psychometrischen Qualität der Skalierung

Die Skalierungsergebnisse können hinsichtlich verschiedener Kennwerte auf ihre psychometrische Güte überprüft werden. Die App bietet die Möglichkeit Modellfitstatistiken, Itemstatistiken der klassischen Testtheorie sowie IRT-basierte Itemfitstatistiken zu berechnen. Für wiederkehrende Testanwendungen mit Linkitems können diese zudem auf IPD überprüft werden. Dafür werden die Itemparameterschätzungen aus den Skalierungen zweier aufeinanderfolgender Testanwendungen durch Equating-Methoden (z. B. Born, Fink, Spoden & Frey, 2019) auf eine gemeinsame Skala gebracht und mittels Wald-Test auf IPD getestet. Der Test auf IPD erfolgt iterativ. Zudem gibt es die Möglichkeit RDS-Files aus der App zu exportieren und für weitere Analysen in R zu nutzen.

Schritt 4: Personenparameterschätzung

Basierend auf den Itemparameterschätzungen können nun die Personenparameter anhand verschiedener Schätzverfahren (z. B. Glas, 2016) bestimmt werden. Für die grafische Darstellung der Item- und Personenparameter gibt es

die Möglichkeit eine Wright Map zu erstellen. Darüber hinaus wird die Reliabilität berechnet.

Schritt 5: Kategorisierung von Testergebnissen
In einigen Anwendungsbereichen ist es nützlich, Testergebnisse nicht nur als numerische Werte, sondern auch in Form inhaltlich definierter Kategorien zurückzumelden. Dies kann kriteriumsorientierte Testwertinterpretationen erleichtern. Bei vorliegenden Grenzwerten auf dem latenten Merkmalskontinuum können die Kategorisierungen mit der App automatisch durchgeführt werden. Wichtig im Hinblick auf die Validität der abgeleiteten Kategorieinterpretationen ist eine sorgfältige Bestimmung der Grenzwerte, die bei Kompetenztests üblicherweise auf Standard-Setting-Prozeduren basiert.

Dokumentation des Tests

Mit Hilfe des R-Paketes knitr (Xie, 2015) bietet die App die Möglichkeit, automatisiert eine Dokumentation des Tests zu erstellen (vgl. Spoden & Buchwald, 2018). Hierbei werden alle relevanten Informationen zum Test zusammengestellt und als DOCX-Dokument exportiert.

Zusammenfassung und Ausblick

Mit der KAT-HS-App liegt eine kostenfreie Software zur Konstruktion, Administration und Auswertung psychometrisch fundierter, computerbasierter Tests vor. Sie verfügt über eine grafische Nutzeroberfläche und erleichtert so auch Nutzergruppen ohne Programmierkenntnisse in R die Bedienung. Darüber hinaus werden mit der Installation ein Benutzerhandbuch, Vorlagen für Item- und Personentabellen sowie Beispieldatensätze bereitgestellt. Da die App sowohl die computerbasierte Testadministration, als auch die Analysen in einem Programm ermöglicht, füllt sie eine relevante Lücke innerhalb verfügbarer Software. Weiterhin ist sie die erste Software, mit der die für viele Anwendungsbereiche attraktive KKS direkt genutzt werden kann. Durch die Konzentration auf zentrale IRT-basierte Verfahren, der für viele Anwendungsbereiche geeigneten Voreinstellung an Methoden (die bei Bedarf angepasst werden kann) und einer einfachen Benutzerführung eröffnet sie Anwendern die Nutzung von IRT-basierten Tests, für die die bislang verfügbaren Softwarepakete eine zu hohe Hürde darstellten. Sollten versierte Nutzer bisher nicht implementierte Methoden anwenden wollen, können R-Objekte exportiert und für weiterführende Analysen in R genutzt werden. Künftig soll der Funktionsumfang der App erweitert werden (z. B. ordinale IRT-Modelle; Person-Fit). Die Weiterentwicklung soll durch das Bilden einer aktiven Nutzercommunity über das Onlineportal https://kat-hs.uni-frankfurt.de gefördert werden.

Literatur

Battauz, M. (2015). equateIRT: An R package for IRT test equating. *Journal of Statistical Software, 68(7),* 1 – 22. https://doi.org/10.18637/jss.v068.i07

Bock, R. D. & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46,* 443 – 459. https://doi.org/10.1007/BF02293801

Born, S., Fink, A., Spoden, C. & Frey, A. (2019). Evaluating different equating setups in the continuous item pool calibration for computerized adaptive testing. *Frontiers in Psychology*, *10*, 1277. https://doi.org/10.3389/fpsyg.2019.01277

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48(6),* 1 – 29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71(5),* 1 – 39. https://doi.org/10.18637/jss.v071.i05

Chang, W., Cheng, J. Allaire, J. J., Xie, Y. & McPherson, J. (2019). *Shiny: Web Application Framework for R* (R package version 1.3.2) [Computer Software].

ETH Zürich, Lehrentwicklung und -technologie (2019). *Safe Exam Browser* (Version 2.2.3) [Computer Software]. Zürich: ETH Zürich, Lehrentwicklung und -technologie (LET).

Fink, A., Born, S., Frey, A. & Spoden, C. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling, 60,* 327 – 346.

Frey, A. (2020). Computerisiertes adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3., aktual. und überarb. Auflage, S. 501 – 525). Berlin, Heidelberg: 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. & Fink, A. (in press). 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.* New York, NY: Springer.

Glas, C. A. W. (2016). Maximum-likelihood estimation. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume two: statistical tools* (pp. 197 – 216). London: Chapman and Hall.

Spoden, C. & Buchwald, F. (2018). Diagnostische Tests mit R und knitr: Erstellung, Auswertung und Vorbereitung der Rückmeldung. *Diagnostica, 64*, 49 – 57. https://doi.org/10.1026/0012-1924/a000189

Spoden, C. & Frey, A. (Hrsg.). (im Druck). *Psychometrisch fundierte E-Klausuren für die Hochschule.* Lengerich: Pabst Science Publishers.

Spoden, C., Frey, A., Fink, A. & Naumann, P. (2020). Kompetenzorientierte elektronische Hochschulklausuren im Studium des Lehramts. In K. Kaspar, M. Becker-Mrotzeck, J. Hofhues, J. König & D. Schmeinck (Hrsg.), *Bildung, Schule und Digitalisierung* (S. 184 – 189). Münster: Waxmann.

Linden, W. J. van der (2016). *Handbook of item response theory, volume one: models.* London: Chapman and Hall. https://doi.org/10.1201/9781315374512

Williams, E. J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Chemistry, 2,* 149 – 168.

Xie, Y. (2015). *Dynamic documents with R and knitr* (The R series, 2nd ed.). Boca Raton, FL: Chapman & Hall / CRC.

Interessenskonflikt

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

Förderung

Die in dem Artikel berichtete Forschung wurde durch das Bundesministeriums für Bildung und Forschung (Ref: 16DHL1005) gefördert.

ORCID

Aron Fink
https://orcid.org/0000-0003-0624-1131

Aron Fink, M.Sc.
Goethe-Universität Frankfurt
Theodor-W.-Adorno-Platz 6
60323 Frankfurt am Main
a.fink@psych.uni-frankfurt.de