Wann sollten computerbasierte Verfahren zur Messung von Kompetenzen anstelle von papier- und bleistift-basierten Verfahren eingesetzt werden?

Andreas Frey · Johannes Hartig

Zusammenfassung: Kompetenzen werden zurzeit meistens mit Papier- und Bleistift-Tests erhoben. Die Messung von bestimmten Kompetenzen stellt aber in dreierlei Hinsicht neue Herausforderungen an die pädagogisch-psychologische Diagnostik, denen nur schwer mit Papier- und Bleistift-Tests begegnet werden kann. Für a) Kompetenzen, bei denen eine zeitkritische Ausführung komplexer Interaktionen zentral ist, b) komplexen Kompetenzkonstrukten sowie c) Kompetenzen, deren Ausübung an einen Computer gebunden ist, ist eine computerbasierte Erfassung als deutlich vorteilhafter einzuschätzen als eine papier- und bleistift-basierte Messung. Generell gibt es jedoch kein per se überlegenes Testmedium. Vielmehr sollte jeweils das Testmedium genutzt werden, mit dem die jeweilige theoretische Konzeption eines Kompetenzkonstrukts am besten operationalisiert werden kann. Erscheinen sowohl Papier- und Bleistift als auch Computer prinzipiell zur Messung eines Kompetenzkonstrukts geeignet, sollte kalkuliert werden, welches der beiden Testmedien ökonomischer wäre. Bei der Kalkulation sind nicht nur die Mehrkosten durch den Computereinsatz zu beachten, sondern auch spezifische Kosten, die nur bei Papier- und Bleistift-Tests anfallen.

Schlüsselwörter: Kompetenzen · Messen · Testen · Computerbasiertes Testen

In which settings should computer-based tests be used instead of paper and pencil-based tests?

Abstract: Currently, competencies are mostly assessed with paper and pencil tests. Nevertheless, for some competencies, using paper and pencil tests is problematic for three reasons. For (a) competencies requiring complex interactions under time constraints, (b) complex competence

© Springer Fachmedien Wiesbaden 2013

Prof. Dr. A. Frey (✉)
Institut für Erziehungswissenschaft, Friedrich-Schiller-Universität Jena,
Am Planetarium 4, 07737 Jena, Deutschland
E-Mail: andreas.frey@uni-jena.de

Prof. Dr. J. Hartig
Bildungsqualität und Evaluation, Deutsches Institut für Internationale Pädagogische Forschung,
Schloßstr. 29, 60486 Frankfurt am Main, Deutschland
E-Mail: hartig@dipf.de

constructs, and (c) competencies requiring using a computer, a computer-based assessment provides several advantages compared to a paper and pencil-based assessment. Nevertheless, none of the two testing media is superior to the other in a general way. Thus, the testing medium which provides the best opportunities for a proper operationalization of the competence construct at stake should be selected. If both paper and pencil and computer seem applicable for measuring a competence construct, precisely costing out and comparing the two alternatives is recommended. Within this comparison, not only the additional costs caused by using computers but also the specific costs of paper and pencil-based assessment should be considered.

Keywords: Competencies · Computer-based testing · Measurement · Testing

1 Können alle Kompetenzen valide mit papier- und bleistift-Verfahren gemessen werden?

Die standardisierte pädagogisch-psychologische Diagnostik bedient sich seit ihren Anfängen zu Beginn des 20. Jahrhunderts vornehmlich papier- und bleistift-basierter Messverfahren. Mit diesem Testmedium konnte für eine Vielzahl an Konstrukten adäquate und nützliche Messinstrumente konstruiert werden. Auch in der Kompetenzdiagnostik ist der papier- und bleistift-basierte Zugang momentan vorherrschend. Die Messung von bestimmten Kompetenzen stellt aber in dreierlei Hinsicht neue Herausforderungen an die pädagogisch-psychologische Diagnostik, denen nur schwer mit dem herkömmlichen Zugang mit Papier- und Bleistift-Tests begegnet werden kann.

Erstens rücken zunehmend Kompetenzen in den Blick, die im teilweise zeitkritischen Ausführen komplexer Interaktionen bestehen und/oder die an realistische Anwendungssituationen gebunden sind. Gerade solche Kompetenzen sind im Hinblick auf beruflichen und sozialen Erfolg besonders relevant. Der Fokus liegt bei diesen Kompetenzen weniger auf der Wiedergabe oder der Verwendung isolierter curricularer Inhalte, sondern in deren integrierter Anwendung bei der Lösung realer Aufgabenstellungen und Probleme in beruflichen und sozialen Kontexten. Die Quantifizierung von Interaktionsmustern, die hochinferente Protokollierung von Bearbeitungszeiten und die Darstellung realistischer Kontexte ist mit dem statischen Testmedium Papier- und Bleistift in der Regel nur sehr begrenzt oder mit sehr großem Aufwand möglich. Computer bieten hier vielfältigere Möglichkeiten, da in computerbasierten, interaktiven Testumgebungen Daten verhaltensnah und zeitkritisch erhoben werden. Dies ermöglicht die Messung von Konstrukten wie dynamisches Problemlösen (vgl. z. B. Funke 2003; Wirth und Klieme 2003; Leutner et al. 2004), wobei auch Bearbeitungszeiten auf einfache Weise aufgezeichnet und mit aktuellen Methoden der Item-Response-Theorie zur Anreicherung der Testergebnisse verwendet werden (vgl. z. B. Goldhammer und Klein Entink 2011; van der Linden 2007). Durch die Darbietung von graphischen (hochauflösende Abbildungen, Animationen, Videosequenzen usw.) und auditiven Stimuli kann zudem ein hoher Grad an Realitätsnähe erreicht werden. So kann beispielsweise authentisches Tonmaterial als Stimulus bei einem Hörverstehenstest eingesetzt werden.

Zweitens handelt es sich bei Kompetenzen häufig um sehr komplexe Konstrukte. Sie sind in der Regel kontextspezifisch und handlungsbezogen definiert und basieren zumeist auf differenzierten theoretischen Konzeptionen. Während beispielsweise klassische, in

der Psychologie betrachtete kognitive Leistungskonstrukte wie Merkfähigkeit teilweise mit einem einzelnen Satz definiert werden können, sind die theoretischen Beschreibungen von Kompetenzen zumeist deutlich umfassender. Die reliable und valide vollständige Abbildung solch komplexer Konstrukte erfordert in der Regel eine umfangreiche empirische Datenbasis, um die relevanten Inhalte in der theoretisch spezifizierten Breite angemessen zu repräsentieren. Diese umfangreiche Datenbasis kann entweder durch die Vorgabe von sehr vielen Items an einzelne Testpersonen oder durch die Untersuchung von sehr großen Stichproben sichergestellt werden. Aufgrund der auch im zweiten Fall nötigen großen Anzahl an Items macht in der Regel die Nutzung unvollständiger Testheft-designs (vgl. Frey et al. 2009; Gonzalez und Rutkowski 2010; Yousfi und Böhme 2012) notwendig, mit denen zwar unverzerrte Populationskennwerte berechnet, aber keine Aussagen auf Individualebene getroffen werden können. Zur Messung komplex definierter Kompetenzen wie beispielsweise mathematischer Kompetenz oder Lesekompetenz sind die Anforderungen an Testlänge (Items pro Person) oder Stichprobengröße so hoch, dass sich eine Ableitung von hinreichend präzisen individuellen Kompetenzwerten mit Papier- und Bleistift-Verfahren aus praktischen Gründen ausschließt. Diese Problematik lässt sich durch computerisiertes adaptives Testen (CAT; vgl. z. B. Frey 2012) erheblich verringern. Bei CAT werden in der Regel nur halb so viele Items benötigt, um genauso präzise Messwerte zu erhalten wie bei herkömmlichen sequenziellen Tests. Adaptives Testen ist außerdem für die Kompetenzdiagnostik ausgesprochen nützlich, da das ursprünglich eindimensionale CAT-Konzept in den letzten Jahren auch für mehrdimensionale Messansätze (Überblick in Frey und Seitz 2009; hypothetische Anwendung bei PISA in Frey und Seitz 2011) und kognitive Diagnosemodelle (vgl. z. B. Rupp et al. 2010) erweitert wurde (vgl. Cheng 2009, 2010).

Drittens werden zunehmend Kompetenzen relevant, deren Ausübung an die Benutzung von Informationstechnologie gebunden ist. Beispiele hierfür sind die Kompetenz im Umgang mit Informations- und Kommunikationstechnologien selbst (ICT Literacy), das Lesen elektronischer Texte, die Verwendung des Computers als Werkzeug zur Lösung mathematischer oder naturwissenschaftlicher Probleme, die computerbasierte Durchführung umfangreicher oder iterativer Berechnungen oder die Erstellung von technischen Zeichnungen und Animationen. Die genannten Kompetenzen sind entweder gänzlich an den Computer gebunden oder werden in praktischen Betätigungsfeldern fast ausschließlich mit dem Computer ausgeübt. Für die Zukunft ist damit zu rechnen, dass weitere schulisch vermittelte Kompetenzen im Berufsleben vor allem unter Benutzung von Computern zunehmend relevant werden. Es bietet sich folglich an, dass eine Messung solcher Kompetenzen ebenfalls am Computer stattfindet. Dies lässt eine höhere prognostische Validität erwarten als bei einer papier- und bleistift-basierten Messung, weil die Handlung, auf die aufgrund der Testwerte geschlossen werden soll, am Computer realistischer getestet werden kann.

Vorerst werden papier- und bleistift-basierte Verfahren voraussichtlich ein zentraler Zugang der Kompetenzdiagnostik bleiben. Bei den drei oben skizzierten Arten komplexer Kompetenzen, die nur schwer oder unter großem Aufwand mit Papier- und Bleistift-Verfahren zu messen sind, werden computerbasierte Verfahren jedoch künftig erheblich an Bedeutung gewinnen. Für die Testpraxis stellt sich allerdings nicht die Frage nach dem besten Testmedium per se, sondern danach, mit welchem Testmedium die jeweilige

theoretische Konzeption eines Kompetenzkonstrukts am besten operationalisiert werden kann. Zentral ist dabei die Frage der Validität und somit danach, in welchem Ausmaß die ermittelten Testwerte im Sinne des theoretischen Kompetenzverständnisses interpretiert werden können. Bei vielen Kompetenzkonstrukten ergibt sich die Wahl des am besten geeigneten Testmediums direkt aus der jeweiligen theoretischen Konzeption. Erscheinen sowohl papier- und bleistift-basierte als auch computerbasierte Testverfahren als prinzipiell geeignet, valide interpretierbare Testwerte zu generieren, sollte kalkuliert werden, welche der beiden Administrationsformen vorteilhafter ist. Dies ist notwendig, da bestimmte Rahmenbedingungen erfüllt sein müssen, damit computerbasiertes Testen auf ökonomische Weise eingesetzt werden kann. Diese Rahmenbedingungen werden im folgenden Abschnitt skizziert.

2 Unter welchen Voraussetzungen sind computerbasierte Administrationsformen auf ökonomische Weise zu realisieren?

Ein häufig genanntes Argument gegen computerbasiertes Testen sind die damit verbundenen hohen Kosten. In der Tat können erhebliche Kosten bei computerbasierten Kompetenztestungen anfallen. Nicht selten sind aber auch Rahmenbedingungen gegeben, unter denen eine computerbasierte Testung ökonomisch zu realisieren ist, teilweise sogar deutlich ökonomischer als Papier- und Bleistift-Tests. Deshalb sollten die gegebenen Rahmenbedingungen vor der Entscheidung für ein Testmedium sorgfältig analysiert und die jeweils anfallenden Kosten kalkuliert werden. Eine ökonomische Durchführung von computerbasierten Verfahren ist vor allem bei

- vorhandener Testsoftware,
- Low-Stakes-Testungen und Internetzugang am Testort,
- großen Stichproben,
- häufig wiederholten Routinetestungen sowie
- am Testort verfügbaren Computern

zu erwarten (vgl. auch Parshall et al. 2002).

Soll eine Entscheidung zwischen papier- und bleistift-basierter Testung und computerbasierter Testung getroffen werden, ist es wichtig, auch für Erstere alle Kosten zu beachten. Beispielsweise werden beim Vergleich von papier- und bleistift- mit computerbasierten Verfahren häufig vor allem die Mehrkosten betrachtet, die direkt durch den Computereinsatz entstehen. Die bei einigen Studien erheblichen Kosten für die Verwaltung des Itempools, den Satz und Druck von Testheften, die Verbringung von gedruckten Testmaterialien an den Testort, die Eingabe oder das Einscannen von Antworten sowie die Lagerung von Testbögen im Rahmen der gesetzlichen Archivierungsfrist werden dabei teilweise übersehen. Werden solche Aspekte in die allfälligen Kalkulationen miteinbezogen, fallen die Kosten bei Papier- und Bleistift-Verfahren nicht selten höher aus, als bei computerbasierten Verfahren.

Danksagung: Diese Veröffentlichung wurde ermöglicht durch Sachbeihilfen der Deutschen Forschungsgemeinschaft (Kennz.: FR 2552/2-2 und HA 5050/2-2) im Schwerpunktprogramm

„Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (SPD 1293).

Literatur

- Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, 74, 619–632.
- Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balanced attribute coverage: the modified maximum global discrimination index method. *Educational and Psychological Measurement*, 70, 902–913.
- Frey, A. (2012). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2., akt. und überarb. Aufl., S. 275–293). Berlin: Springer.
- Frey, A., & Seitz, N. N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*, 35, 89–94.
- Frey, A., & Seitz, N. N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in PISA. *Educational and Psychological Measurement*, 71, 503–522.
- Frey, A., Hartig, J., & Rupp, A. (2009). Booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, 28, 39–53.
- Funke, J. (2003). *Problemlösendes Denken*. Stuttgart: Kohlhammer.
- Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, 39, 108–119.
- Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 3, 125–156.
- Leutner, D., Klieme, E., Meyer, K., & Wirth, J. (2004). Problemlösen. In PISA-Konsortium Deutschland (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* (S. 147–175). Münster: Waxmann.
- Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: theory, methods, and applications*. New York: Guilford Press.
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287–308.
- Wirth, J., & Klieme, E. (2003). Computer-based assessment of problem solving performances. *Assessment in Education: Principles, Policy & Practice*, 10, 329–345.
- Yousfi, S., & Böhme, H. (2012). Principles and procedures of considering item sequence effects in the development of calibrated item pools: Conceptual analysis and empirical illustration. *Psychological Test and Assessment Modeling*, 54(4), 366–396.