Offene Antworten und Multiple Choice Tests im direkten Vergleich

Theoretischer Hintergrund

Aufgrund der Testökonomie und der Auswerteobjektivität wird häufig ein geschlossenes Antwortformat, insbesondere Multiple-Choice Tests (MC) bevorzugt (Bühner, 2006). Seit längerem gibt es Überlegungen, ob verschiedene Antwortformate unterschiedliche kognitive Prozesse zur Bearbeitung erfordern (Martinez, 1999). So wird diskutiert, dass offene Antwortformate eine detailliertere Erfassung der zugrunde liegenden Lösungsprozesse und damit einen tieferen Einblick in die aufgabenrelevanten Konstrukte ermöglichen können (Williamson, Xi, & Breyer, 2012). Ferner erreichen Probanden in offenen Antwortformaten weniger Punkte als in geschlossenen Antwortformaten; je wichtiger, der Test für die Probanden ist, desto größer ist dieser Unterschied (DeMars, 2000). Es zeigt sich, dass bei freien Antwortformaten neben fachlich adäquaten Vorstellungen auch Alltagsvorstellungen erwähnt werden (Opfer, Nehm, & Ha, 2012). Vergleicht man zusätzlich Interviews mit MC Tests und mit offenen Antwortformaten, unterscheiden sich die MC Test mitunter deutlich von den anderen beiden, die sehr ähnlich ausfallen (Opfer, Nehm, & Ha, 2012).

Mit dem Force Concept Inventory veröffentlichten Halloun und Hestenes (1985) ein Instrument, das das Verständnis Studierender im Bereich Mechanik erfassen soll. Die Autoren konstruierten dafür zunächst ausschließlich offene Aufgaben. Ausgehend von den Antworten auf diese wurden dann falsche (und richtige) MC Antworten generiert. Die Testentwicklung umfasste mehrere Zyklen mit dem Ziel, die Validität und Reliabilität sicherzustellen (Halloun & Hestenes, 1985). Die als falsch zu wertenden Antwortalternativen stellen somit mehr oder weniger übliche Alltagsvorstellungen dar (Hestenes, Wells, & Swackhammer, 1992; Halloun & Hestenes, 1985). Rebello und Zollman (2004) vergleichen in einer Studie vier Items des FCI in geschlossenem Format mit denselben vier Items im offenen Antwortformat, wobei allerdings ein einzelner Proband nicht dieselben Aufgaben zweimal erhält. Die Autoren berichten, insgesamt würden von den Probanden andere falsche Lösungen generiert, als durch die Antwortalternativen angeboten (Rebello & Zollmann, 2004). Dies deutet auf Effekte auf der Einzelebene hin, die aber nicht untersucht wurden. Ein weiterer Grund für die fehlende Differenz könnte aber auch die Auswahl der Aufgaben sein, die nicht entlang möglicher Teilkonstrukte erfolgte (ebd.). Savinainen und Viiri (2008) vergleichen den FCI mit Interviews und stellen fest, dass die Ergebnisse im FCI nur bedingt repliziert werden. Grund dafür ist ein relevanter Anteil an Personen, die im FCI sehr gute Ergebnisse erreichen, dies in den Interviews aber nicht reproduzieren können.

Fragestellung

Die hier vorgestellte Studie stellt eine Erweiterung der Studien von Härtig (2014a) dar. Es werden zwei Forschungsfragen untersucht:

- (1) Inwiefern beeinflusst das Antwortformat des Force Concept Inventories die Ergebnisse der Probanden?
- (2) In welchem Ausmaß wirkt sich eine Modifikation des Kontexts auf die Bearbeitung durch die Probanden aus?
- (3) Lässt sich die Auswertung offener Antwortformate automatisieren (vgl. Härtig 2014b)?

Zunächst wurde eine sehr enge Inhaltsbegrenzung auf einen Teilaspekt des Kraftverständnisses (relevante Kraftarten) vorgenommen. Darüber hinaus wurden in der Vergangenheit die Kontexte der Aufgaben diskutiert, weswegen dies hier in Anlehnung an die Arbeiten von Schecker und Gerdes (1999) systematisiert untersucht wird.

Methode

Die Studie verwendet nur eine Auswahl von Items des FCI. Die Items fragen nach Kräften, die in einer bestimmten Situation relevant sind. Es wurden sieben Aufgaben aus dem Original FCI von 1992 in der überarbeiteten Übersetzung von Neumann und Kollegen nach Schecker und Gerdes (1999) ausgewählt. Diese Items erfassen vermutlich einen sehr ähnlichen Aspekt der Newtonschen Mechanik, im Sinne der Kategorie Kraftarten. Dies soll der Beantwortung der ersten Forschungsfrage dienen. Forschungsfrage zwei soll zusätzlich klären, inwieweit die kontextuelle Einbindung der physikalischen Inhalte selbst – bei ansonsten gleicher Frage – zu unterschiedlichen Lösungen zu führt. Aus diesem Grund wurde für zwei der Items der Kontext variiert. Neben dem Kontext Golf für eines der Items gab es vier weitere Kontexte: Fußball, Weitsprung, Angry Birds© und den Wurf eines Schlüssels. Neben dem Kontext Buch gab es ebenfalls zwei weitere Kontexte: Stuhl und Mädchen. Man würde erwarten, dass Probanden auf alle fünf bzw. drei Fragen exakt gleich antworten, da die physikalische Sichtweise extrem vergleichbar ist.

Alle Items wurden sprachlich so modifiziert, dass sie sowohl als MC Items (wie im FCI) als auch im offenen Antwortformat präsentiert werden können. Das gesamte Testinstrument enthält damit 26 Items, 13 mit offenem Antwortformat, 13 als MC Items. Je eine offene und eine geschlossene Aufgabe sind dabei vollkommen identisch. Außerdem umfasst jedes Antwortformatpaket fünf Original FCI Fragen mit ähnlichem Inhalt und zusätzlich zwei Original FCI Fragen mit jeweils zwei bzw. vier Kontextvariationen.

Allen Probanden wurden alle Items vorgelegt, immer die offenen Aufgaben zuerst, die Probanden dürfen nie vor oder zurück blättern. Der Test wurde sowohl als Papier und Bleistift Instrument als auch als Online Erhebung. Die Antworten für die offenen Antwortformate wurden mittels der Antwortalternativen des geschlossenen als Kategorien kodiert. Die Güte dieser Kodierung wurde mittels einer Zufallsstichprobe und Doppelkodierung überprüft. Die Interrater-Reliabilität hierfür liegt für alle Items im guten bis sehr guten Bereich ($0,800 < \text{Cohen's } \kappa < 1,000$).

Ergebnisse

An der Studie nahmen 610 Studierende aus Deutschland, Österreich und der Schweiz teil. 412 Personen studierten Physik als Hauptfach (auch Lehramt). Die Daten sind nicht normalverteilt, optisch gleichen die Resultate einer Bimodalverteilung. Diese Aufteilung lässt sich auch post-hoc durch kein erfasstes Merkmal erklären (z. B. Lehramt oder Semester). Entsprechend fällt ein Kolmogorow-Smirnow Test signifikant aus, daher werden die Analysen mittels robusterer, non-parametrischer Verfahren durchgeführt.

Zur Klärung der Forschungsfrage 1 werden dichotomisierte Auswertungen herangezogen und in einer konfirmatorischen Faktorenanalyse verwendet. Es sollen zwei verschiedenen Teilskalen betrachtet werden: Es wird der Einfluss des Antwortformats bei den Items untersucht. Zu vergleichen sind also drei Modelle: Ein Modell mit einer Skala, eines mit zwei Skalen aufgrund des Antwortformats und ein Modell mit dem Antwortformat als genesteter Faktoren. Der Modell-Vergleich mittels Chi-Quadrat Test zeigt dabei die Überlegenheit des Modells mit genesteten Faktoren für das Antwortformat gegenüber beiden anderen.

Für Forschungsfrage 2 wird für die beiden kontextvariierten Items angenommen, jedes Items stelle ein Rating des Verständnis der Probanden dar. Damit lässt sich die Ähnlichkeit der Ratings mittels Fleiss' κ als Kohärenzanalyse verstehen. Es zeigt sich, dass die Effekte leicht

unterschiedlich ausfallen. Für die offenen Antwortformate zeigt sich durchgängig der Kontexteffekt, für die MC Items nur in einem der beiden Fälle (Tab 1).

Tabelle 1: Kontexteffekte für die Itemgruppen

(Golf etc.) MC	(Golf etc.) Offen	(Buch etc.) MC	(Buch etc.) Offen
Fleiss 'κ=.877	Fleiss' κ=.618	Fleiss 'κ=.683	Fleiss 'κ = .695

Folglich lohnt sich der Einsatz offener Antwortformate: Forschungsfrage 1 führt zu Belegen dafür, dass die Probanden hier andere Antworten generieren, bei im Mittel niedrigeren Scores. Forschungsfrage 2 zeigt auf, dass der Kontexteinfluss vor allem im offenen Antwortformat relevant wird. In Anlehnung an Arbeiten von Härtig (2014b) soll daher die automatische Auswertung offener Antwortformate auch hier zu einer ökonomischen und objektiven Auswertung führen. Dafür wurden mittels SPSS TA zunächst Regeln zur softwarebasierten Kodierung entwickelt und diese dann erprobt (vgl. Härtig 2014b für Details). Es zeigt sich im Vergleich zwischen Software und einer Gruppe menschlicher Experten, dass die Software zu vergleichbaren Befunden kommt – Fleiss 'κ fällt in der Gruppe nur menschlicher Experten genauso hoch aus wie bei der Gruppe menschliche Experten und Software. Damit kann hier ferner geschlossen werden, dass offene Items ökonomisch und objektiv auch in größeren Erhebungen zu verwenden sind.

Literatur

- Bühner, M. (2006). Einführung in die Test- und Fragebogenkonstruktion (2nd ed.). München, Don Mills: Pearson Studium.
- DeMars, C. E. (2000). Test Stakes and Item Format Interactions. In: Applied Measurement in Education, 13, 55–77. doi: 10.1207/s15324818ame1301_3.
- Halloun, I. A. & Hestenes, D. (1985). The initial knowledge state of college physics students. In: Am. J. Phys., 53, 1043-1048.
- Härtig, H. (2014a – im Druck). Der Force Concept Inventory Vergleich einer offenen und einer geschlossenen Version. PhyDid A.
- Härtig, H. (2014b). Software-basierte Evaluation freier Antwortformate. Zeitschrift für Didaktik der Naturwissenschaften. Doi: 10.1007/s40573-014-0012-6.
- Hestenes, D., Wells, M., & Swackhammer, G. (1992). Force Concept Inventory. In: The Physics Teacher, 30, 141-158.
- Martinez, M. E. (1999). Cognition and the question of test item format. In: Educational Psychologist, 34, 207–218. doi: 10.1207/s15326985ep3404_2.
- Opfer, J. E., Nehm, R. H., & Ha, M. (2012). Cognitive foundations for science assessment design: Knowing what students know about evolution. In: Journal of Research in Science Teaching, 4, 744–777. doi: 10.1002/tea.21028
- Rebello, N. S. & Zollman, D. (2004). The effect of distracters on student performance on the force concept inventory. In: Am. J. Phys., 72, 116-125.
- Savinainen, A. & Viiri, J. (2008). The Force Concept Inventory as a Measure of Students' Conceptual Coherence. In: International Journal of Science and Mathematics Education, 6, 719.
- Schecker, H. & Gerdes, J. (1999): Messung von Konzeptualisierungsfähigkeit in der Mechanik – Zur Aussagekraft des Force Concept Inventory. In: ZfDN, 5, 75-89.
- Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A Framework for Evaluation and Use of Automated Scoring. In: Educational Measurement: Issues and Practice, 31, 2–13.