Tanja Mutschler[1] Universität Potsdam[1]
David Buschhüter[1] Universität Helsinki[2]
Inkeri Kontro[2]
Andreas Borowski[1]

Beyond FCI: Internationale Vergleiche mithilfe eines erweiterten FCI

Motivation

Als Folge der Globalisierung hat der Ausbau interkultureller Vergleichsforschung stark zugenommen und so den Diskurs über Vergleiche zwischen unterschiedlichen kulturellen Kontexten weiter angeregt (Curtarelli & van Houten, 2018). Trotz ihrer großen Bedeutung für die globalisierte Gesellschaft hat die internationale Vergleichsforschung immer noch stark mit methodischen Problemen zu kämpfen (Curtarelli & van Houten, 2018; Sekaran, 1983). Denn um die von der Bologna-Erklärung geforderten Kernziele, wie die Schaffung eines gemeinsamen europäischen Hochschulraums, der die Anerkennung von Studienleistungen und Studienabschlüssen beinhaltet (Bologna Process Committee, 1999), zu erreichen, bedarf es Testinstrumente, die einen *gültigen* Vergleich zwischen den verschiedenen kulturellen Kontexten zulassen und in der Lage sind, Lernergebnisse zu messen. Laut CALOHEE (2018, S. 1) erlauben bestehende Instrumente aber weder einen transnationalen Vergleich, noch die Messung des Gelernten. Aus diesem Grund betont das EU-geförderte Projekt die Wichtigkeit der Entwicklung dieser Testinstrumente und fokussiert dabei unter anderem auch die Physik-Hochschulausbildung.

In diesem Fachbereich gibt es eine Vielzahl von Tests, die Leistungen bzw. Fähigkeiten in verschiedenen Fachwissensbereichen messen. Das Force Concept Inventory (FCI) von Hestenes et al. (1992) gehört dabei zu den bekanntesten. Es gilt als ein jahrzehntelang erprobtes und international verwendetes, standardisiertes Diagnoseinstrument „zur Bestimmung des qualitativen Verständnisses der newtonschen Mechanik“ (Wilhelm & Heuer, 2005, S. 1) und hätte damit die Voraussetzungen als Vergleichsinstrument innerhalb der frühen, universitären Physikausbildung eingesetzt zu werden. Da dieses Instrument in Populationen einiger Länder Deckeneffekte aufweist (vgl. Persson, 2015), bedarf es aber weiterer Items mit höheren Aufgabenschwierigkeiten, um valide Vergleiche zwischen entsprechenden Gruppen zu ermöglichen.

Ziel der vorliegenden Studie ist es deshalb, neue Items für den FCI zu entwickeln und diese mithilfe etablierter Verfahren zu übersetzen. In einem ersten Schritt wird mithilfe einer Think-Aloud-Studie die Vergleichbarkeit englisch- und deutschsprachiger Items überprüft und in einer ersten Version präpilotiert. Auf Basis der Ergebnisse wird diskutiert, inwiefern sich die Testitems für internationale Vergleiche eignen.

Theoretischer Hintergrund

Der FCI gilt als der „bekannteste Test zur Messung des konzeptuellen Verständnisses in den grundlegenden Gebieten der Physik“ (Enders, 2016, S. 4). Trotz einiger Kritik an seiner Aussagekraft (u.a. Huffman & Heller, 1995; Schecker & Gerdes, 1999), wird der FCI im Schul- und Hochschulbereich eingesetzt, um Lehrveranstaltungen zu evaluieren (u.a. Maries & Singh, 2016) oder Vorstellungen von Lernenden zu erfassen (u.a. Savinainen & Scott, 2002). Da der Multiple-Choice-Test ursprünglich aber für den Einsatz an High Schools gedacht war, kommt es beim Einsatz an Universitäten teils zu Deckeneffekten (Persson, 2015). Aufgrund der hohen Akzeptanz und seiner vielfachen Verwendung ist der Einsatz des FCI als Instrument für internationale Vergleiche dennoch vorstellbar.

In diesem Zusammenhang betonen zahlreiche Studien (u.a. Curtarelli & van Houten, 2018; Hambleton & Kanjee, 1995) die besonderen Anforderungen an Testinstrumente, die für

Vergleiche zwischen kulturellen Kontexten eingesetzt werden. Um die Validität der Interpretationen der Testinstrumente nicht einzuschränken, muss laut Hult et al. (2008) Datenäquivalenz über die Dimensionen Konstrukt, Messung und Datenerhebung gewährleistet sein. Letztere ist dabei der am wenigsten problematische Aspekt. Bezüglich der Konstruktäquivalenz bietet Duit (2009) in diesem Zusammenhang einen guten Überblick. Die Übersetzungsäquivalenz – als Teil der Messäquivalenz – wird als zentralste Einheit angesehen (Poortinga, 1983). Die Implementierung dieser wird in den meisten Studien aber selten bis gar nicht dargestellt (Hult et al., 2008). Auch deutsche Übersetzungen des FCI wurden hinsichtlich ihrer Übersetzungsäquivalenz nur wenig diskutiert.
Um nun einen erweiterten FCI als Instrument für transnationale Vergleiche verwenden zu können, muss sichergestellt werden, dass der Test zum einen Fähigkeitsunterschiede von höchstleistenden Studierenden auflöst und zum anderen die (Mess-)Äquivalenz gewährleistet. Daraus ergeben sich die folgenden Forschungsfragen:

F1: Inwieweit sind die neuen Testitems in der Lage, Fähigkeitsunterschiede des oberen Leistungsbereiches aufzulösen?
F2: Inwiefern zeigen sich Unterschiede zwischen finnischen und deutschen Studierenden in der Bearbeitung der Aufgaben?

Methodik
Um sicherzustellen, dass Fähigkeitsunterschiede von höchstleistenden Studierenden abgebildet werden, wurden Items aus dem FCI entnommen und durch schwerere Items anderer Tests (Thornton & Sokoloff, 1998 [FMCE]; Rosenblatt & Heckler, 2011 [FVA]; Halloun, 2007 [IBMC]; Singh & Rosengrant, 2003 [EMCS]) bzw. selbstentwickelter Items ausgetauscht. Die resultierende FCI-Version setzt sich aus 17 ursprünglichen Items (wovon zwei stark überarbeitet wurden) und 14 neuen, teils offenen und teils geschlossenen Items (wovon sechs neu erstellt und acht aus anderen Tests übernommen wurden) zusammen. Um auch die curriculare Fachwissenserweiterung abzudecken, wurden die Themenbereiche *Energie*- und *Impulserhaltung* mit in das Testinstrument aufgenommen.
Sowohl die 14 neuen als auch die beiden stark überarbeiteten Items wurden mittels TRAPD-Methode (*Translation, Review, Adjudication, Pretesting, Documentation*) (Curtarelli & van Houten, 2018) übersetzt. Dafür wurden von zwei Expertinnen unabhängige Übersetzungen angefertigt, die dann in einer Reviewsession besprochen und anschließend von einem dritten Gutachter zusammengeführt wurden. Diese Übersetzung wurde dann durch eine Think-Aloud-Studie mit N=15 Teilnehmenden aus Deutschland und Finnland qualitativ validiert. Das neue Testinstrument wurde sowohl in Finnland (N=107; Universität Helsinki) als auch in Deutschland (N=20; Universität Potsdam) am Ende des ersten Semesters (und damit *nach* Teilnahme an den Einführungs-Mechanikveranstaltungen) präpilotiert.

Ergebnisse
Die quantitative Auswertung zeigt, dass die neuen Items Unterschiede im oberen Leistungsbereich auflösen und keine Deckeneffekte (nach Rost, 1996) erkennbar werden (vgl. Abb. 1 & Abb. 2). Somit lassen sich auch die Fähigkeiten höchstleistender Studierender in den vorliegenden Stichproben auflösen.
Die Think-Aloud-Studie deutet auf ein globales Verständnis der Items der Testteilnehmenden beider Nationen hin. Aufgetretene Unterschiede in den Problemlöseprozessen zeigten sich bei alltagssprachlichen Begriffen und kontextualisierten Situationen. Beispielsweise führte der Begriff „Hang“ (im Gegensatz zu „hill“) im Item Nr. 16 zu Assoziationen mit der Hangabtriebskraft bei den deutschen Teilnehmenden. Diese diskutierten das Item dann im Hinblick auf die wirkenden Kräfte – also mit einem dynamischen Ansatz. Finnische Studierende nutzten zum Lösen der Aufgabe jedoch mehrheitlich einen kinematischen Ansatz.

In Abb. 2 ist zu erkennen, dass sich bei diesem Item (gelbe Markierung) auch die relativen Lösungshäufigkeiten stark unterscheiden und ein Zusammenhang deswegen naheliegend ist. Ein anderer Begriff, der unterschiedlich von den Studienteilnehmenden verstanden wurde, ist „rollen“ bzw. „rolling“ (Item Nr. 17). So setzten die deutschen Studierenden Reibung als Bedingung für diese Bewegungsart voraus. Diese zwar fachlich korrekte Aussage führte im Rahmen des Tests aber dazu, dass der Lösungsweg deutlich komplexer wurde und dadurch seltener zur richtigen Antwort führte. Dieses Ergebnis zeigt sich ebenfalls in Abb. 2 (grüne Markierung). Insgesamt konnten durch die Think-Aloud-Studie fünf Items herausgearbeitet werden, deren Formulierungen angepasst werden müssen.
In Abb. 2 ist ein Streudiagramm der klassischen Aufgabenschwierigkeit über die beiden Stichproben dargestellt. Ausreißende Items können zum Teil mit den Ergebnissen der Think-Aloud-Studie in Zusammenhang gebracht werden (vgl. Items Nr. 16 & 17, Abb. 2). Andere Items, die stark von der Positionierung entlang des Graphen abweichen, geben Hinweise auf nicht-übersetzungsbedingte Unterschiede (z.B. andere curriculare Fokussierung) und sollten dahingehend untersucht werden. Nicht alle Unterschiede in den Problemlöseprozessen der Teilnehmenden, die in der Think-Aloud-Studie aufgedeckt wurden, werden in der quantitativen Darstellung sichtbar.

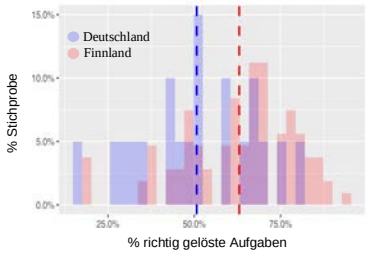

Abb. 1 - Scoreverteilung für die Stichproben DEU - FIN

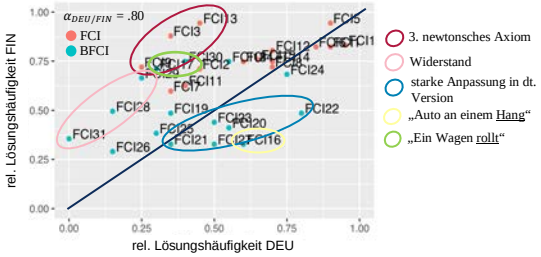

Abb. 2 - Itemverteilung betreffend ihrer relativen Lösungshäufigkeit im Vergleich DEU - FIN

Diskussion

Die Ergebnisse zeigen, dass der weiterentwickelte Test durchaus als internationales Vergleichsinstrument innerhalb der frühen Physikausbildung eingesetzt werden kann, ohne, dass Deckeneffekte auftreten. Im nächsten Schritt sollten dafür – basierend auf den Erkenntnissen der Think-Aloud-Studie – bestehende, offene Items geschlossen und die Übersetzung kritischer Formulierungen überarbeitet werden. Zusätzlich müssten Lösungsansätze dafür integriert werden, dass Aufgaben, die Übersetzungsäquivalenz aufweisen und für die curriculare Validität wichtig sind, wahrscheinlich DIF (Differential Item Functioning) zeigen.
Es konnte weiterhin herausgestellt werden, dass die durch die Think-Aloud-Studie ermittelten Unterschiede in der Bearbeitung der Items weder durch Expert*innenbegutachtungen innerhalb des TRAPD-Verfahrens noch durch die Auswertung der quantitativen Daten vollständig aufgedeckt werden konnten. Es wird daher angeraten, qualitative Validierungen von Übersetzungen mittels Think-Aloud-Studien ergänzend bei der Entwicklung von internationalen Vergleichsinstrumenten einzusetzen (vgl. OECD, 2017), um die von Hult et al. (2008) geforderte Messäquivalenz zu gewährleisten und individuelle Problemlöseprozesse genauer abzubilden.

Danksagung

Wir danken Prof. Ian Bearden (Universität Kopenhagen) und Prof. Ann-Marie Pendrill (Universität Lund) für die Zusammenarbeit an der Entwicklung der BFCI-Items.

Literatur

Bologna Process Committee. (1999). Joint declaration of the European Ministers of Education convened in Bologna on 19 June 1999 (The Bologna Declaration). Verfügbar unter: http://www.ehea.info/media.ehea.info/file/Ministerial_conferences/02/8/1999_Bologna_Declaration_English_553028.pdf

CALOHEE. (2018). *Tuning CALOHEE Assessment Reference Framework for Civil Engineering, Teacher Education, History, Nursing, Physics.* Groningen: University of Groningen.

Curtarelli, M. & van Houten, G. (2018). Questionnaire translation in the European Company Survey: Conditions conducive to the effective implementation of a TRAPD based approach. *Translation & Interpreting Vol. 10 No. 2 (2018)*, pp. 34-54.

Duit, R. (2009). STCSE: Students' and Teachers' Conceptions and Science Education. Verfügbar unter: http://archiv.ipn.unikiel.de/stcse/download_stcse.html

Enders, J. (2016). Peer Instruction und Flipped Classroom in der Service-Lehre Physik. *Didaktik der Physik, Beiträge zur DPG Frühjahrstagung Hannover 2016*.

Halloun, I. (2007). Inventory of Basic Conceptions – Mechanics [IBMC] [Measurement Instrument]. Verfügbar unter https://www.physport.org/assessments/assessment.cfm?I=95&A=IBCM

Hambleton, R. & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: use of improved methods for test adaptions. *European Journal of Psychological Assessment, Vol. 11, No. 3,* pp. 147-157.

Hestenes, D., Wells, M. & Swackhammer, G. (1992). Force Concept Inventory. *The Physics Teacher, Vol. 30, March 1992*, pp. 141-158.

Huffman, D. & Heller, P. (1995). What Does the Force Concept Inventory Actually Measure? *The Physics Teacher, Vol. 33,* pp. 138-143.

Hult, T. et al. (2008). Data equivalence in Cross-Cultural International Business Research: Assessment and Guidelines. *Journal of International Business Studies, Vol. 39 (2008),* pp. 1027-1044.

Maries, A. & Singh, C. (2016). Teaching assistants' performance at identifying common introductory students difficulties in mechanics revealed by the Force Concept Inventory. *Physical Review Physics Education Research, Vol. 12, No. 1*, 010131(26).

OECD. (2017). PISA 2015 Technical Report. Verfügbar unter https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

Persson, J. (2015). Evaluating the Force Concept Inventory for different student groups at the Norwegian University of Science and Technology. Verfügbar unter: arXiv:1504.06099.

Poortinga, Y. (1983). Psychometric approaches to intergroup comparison: The problem of equivalence. In Irvine, S. & Berry, J. (Ed.). *Human assessment and cross-cultural factors*. New York: Plenum, pp. 237-258.

Rosenblatt, R. & Heckler, A. (2011). Systematic study of student understanding of the relationships between the directions of force, velocity, and acceleration in one dimension [FVA] [Measurement Instrument]. *Physical Review Special Topics – Physics Education Research, Vol. 7, No. 2*, 0201121-20. doi: 10.1103/PhysRevSTPER.7.020112

Rost, J. (1996). *Testtheorie & Testkonstruktion*. Bern: Hans Huber.

Savinainen, A. & Scott, P. (2002). Using the Force Concept Inventory to monitor student learning and to plan teaching. *Physics Education, Vol. 37, No. 1*, pp. 53-58.

Schecker, H. & Gerdes, J. (1999). Messung von Konzeptualisierungsfähigkeit in der Mechanik – Zur Aussagekraft des Force Concept Inventory. *Zeitschrift für Didaktik der Naturwissenschaften, Jg. 5, Heft 1, 1999*, S. 75-89.

Schecker, H., Wilhelm, T., Hopf, M. & Duit, R. (Hrsg.). (2018). *Schülervorstellungen und Physikunterricht*. Berlin Heidelberg: Springer.

Sekaran, U. (1983). Measurement issues in cross-national research. *Journal of International Business Studies, Vol. 26, No. 3,* pp. 597-619.

Singh, C. & Rosengrant, D. (2003). Multiple-choice test of energy and momentum concepts [EMCS] [Measurement Instrument]. *American Journal of Physics, Vol. 71, No. 6,* pp. 607-617. doi: 10.1119/1.1571832

Thornton, R. & Sokoloff, D. (1998). Assessing student learning of Newton's law: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula [FMCE] [Measurement Instrument]. *American Journal of Physics, Vol. 66, No. 4*, pp. 338-352. doi: 10.1119/1.18863

Wilhelm, T. & Heuer, D. (2005). Verständnis der newtonschen Mechanik bei bayerischen Elftklässlern – Ergebnisse beim Test „Force Concept Inventory" in herkömmlichen Klassen und im Würzburger Kinematik-/Dynamikunterricht. *Didaktik der Physik, Frühjahrstagung Berlin 2005*.