Physik und Didaktik in Schule und Hochschule
PhyDid 1/13 (2014), S. 53 – 61

Das Force Concept Inventory: Vergleich einer offenen und einer geschlossenen Version

Hendrik Härtig

Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN)
Olshausenstraße 62, D-24118 Kiel
haertig@ipn.uni-kiel.de

Kurzfassung

Das Force Concept Inventory (FCI) wird seit mehr als zwanzig Jahren zur Erfassung des Verständnisses Newton'scher Mechanik verwendet. Schon recht kurz nach der Veröffentlichung entstand eine Diskussion vor allem darüber, ob eine schlichte Auswertung über den Summenscore geeignet sei, das Verständnis zu bemessen. Darüber hinaus wurde hinterfragt, ob mit Multiple-Choice-Aufgaben das Verständnis angemessen überprüft werden kann. Bis heute lassen sich diese Bedenken nicht vollständig entkräften, trotzdem wird das FCI international und national sehr häufig verwendet, z. B. um die Qualität einer neuen Lehrveranstaltung zur Mechanik zu überprüfen.
In der hier vorliegenden Studie wird Studierenden im Fach Physik ein Teil der Aufgaben aus dem FCI sowohl in offenem Antwortformat als auch in geschlossenem Antwortformat zur Bearbeitung vorgelegt. Damit soll untersucht werden, inwieweit das geschlossene Antwortformat die Einschätzung des Verständnisses beeinflusst. Es ergeben sich zwei zentrale Befunde: Zum einen findet sich Evidenz dafür, dass sich das Antwortformat tatsächlich auf das Testergebnis auswirkt – Aufgaben im offenen Format führen zu anderen Lösungen als Aufgaben im geschlossenen Format. Zum anderen zeigt sich der Effekt der Kontexte, mit denen das Verständnis überprüft wird, nur beim offenen Antwortformat, bei geschlossenem Antwortformat scheint sich der Kontext nicht auf die Lösung auszuwirken.

1. Das Force Concept Inventory zusammengefasst

Das Force Concept Inventory (FCI) gibt es bereits seit mehr als 20 Jahren. Sowohl in der Originalfassung von 1992 als auch in einer überarbeiteten Version zielt das FCI auf eine Erfassung des Verständnisses Newton'scher Mechanik, mit speziellem Fokus auf das Konzept der *Kraft* [1]. Das FCI umfasst in der heute üblicherweise verwendeten Fassung 30 Multiple-Choice-Items, die den Bereichen Kinematik, 1. bis 3. Newton'sches Axiom, Superposition und Kraftarten zugeordnet werden [1, 2]. Bereits kurz nach der Veröffentlichung des Originalbeitrags von Hestenes und Kollegen entwickelt sich eine rege Diskussion über die Aussagekraft des FCI [3, 4, 5], vor allem hinsichtlich des häufig genutzten Summenscores. Zentrale Frage der Diskussion ist, ob sich der Summenscore des FCI eignet, das Verständnis von Schülerinnen und Schülern oder Studierenden zu bemessen [6]. Um diese Diskussion nachvollziehen zu können, wird im Folgenden die Entstehungsgeschichte des FCI betrachtet.

1985 veröffentlichen Halloun und Hestenes ein Instrument, das das Verständnis Studierender im Bereich Mechanik erfassen soll [7]. Ausgehend von den Inhalten der Eingangsvorlesungen konstruieren die Autoren zunächst ein Testinstrument mit ausschließlich offenen Aufgaben. Anhand der Antworten der Studierenden auf diese offenen Aufgaben werden falsche (und richtige) Multiple-Choice-Antworten zu den entsprechenden Fragen generiert. Die Testentwicklung umfasst mehrere Zyklen mit dem Ziel, die Validität und Reliabilität des Tests sicherzustellen [7]. Die als falsch zu wertenden vier Antwortalternativen stellen mehr oder weniger übliche Alltagsvorstellungen dar, die nicht mit der fachlich als richtig zu wertenden Antwortalternative in Einklang zu bringen ist [1, 7]. Hinsichtlich der Auswertung empfehlen die Autoren, den Summenscore einer dichotomisierten (richtig/falsch) Auswertung heranzuziehen [7]. Ferner wird Evidenz für die prädiktive Kraft dieses Summenscores präsentiert: Der Studienerfolg in Physik korreliert stark mit der Abschlussnote in entsprechenden Kursen [7]. Weiterhin diskutieren Hestenes, Wells und Swackhammer, dass für die Gestaltung eines Lernprozesses die Auswahl bestimmter falscher Antwortalternativen mehr Informationen für die Lehrenden enthält [1, 4]. Trotzdem griffen in der Vergangenheit die Mehrheit der Autorinnen und Autoren auf den Summenscore zurück [vgl. 2]. Ein Grund dafür mag darin liegen, dass die Auswertung bei einer Multiple-Choice-Gestaltung der Aufgaben verbunden mit der Verwendung des Summenscores sehr ökonomisch und objektiv erfolgt [1, 6].

Am Summenscore verankern verschiedene Autoren ihre Kritik. Für Hestenes und Halloun ist der Summenscore als *coherence measurement* ein Messwert für ein verallgemeinertes Verständnis des Kraftkon-

zepts [4]. Ein Teil der Kritik bezieht sich auf die eingeschränkte Generalisierbarkeit der Prädiktivität des Summenscores. So stellt Henderson fest, dass das FCI zwar positiv prädiktiv wirkt, das mangelhafte Abschneiden im entsprechenden Studium zum Teil aber auch mit hervorragenden FCI-Ergebnissen einhergeht [6]. Huffman und Heller stellen hingegen vor allem die interne Konsistenz des FCI in Frage: Ihre Befunde einer Faktorenanalyse deuten darauf hin, dass es keine zugrunde liegenden Konstrukte gibt, die die Testschwierigkeit angemessen erklären [3]. Diese Befunde erklären die Autoren damit, dass das Verständnis sehr situationsbezogen vorliegt und es sich bei dem FCI unter Umständen um eine Abfrage einer sehr begrenzten Zahl solcher Situationen handelt [3]. Demzufolge wäre der Summenscore tatsächlich nicht mehr als die Summe dieser Teile und damit auch kein Messwert für das verallgemeinerte Verständnis [5]. Da sich aber die Faktorenanalyse von Huffman und Heller wiederum auf dichotomisierte Daten bezieht, ist vor allem diese schlichte richtig/falsch-Auswertung als problematisch anzusehen [8].

Selbst auf der Ebene der einzelnen Items wird über Inkonsistenzen berichtet, auch wenn das Instrument in seiner Gesamtheit zur Evaluation verwendet wird [vgl. 9]. So finden Lasry und Kollegen, dass Probanden bei erneuter Vorlage des gleichen Instruments mitunter beliebig zwischen verschiedenen Falschantworten hin und her wechseln [10]. Bereits Schecker und Gerdes stellen fest, dass die Zuordnung einzelner Antwortalternativen zu bestimmten Subskalen – bis dahin zumindest – wenig untersucht sind [11]. Auf der Suche nach einer möglichen Ursache haben sie die Kontextabhängigkeit der Items diskutiert und empirisch untersucht, wobei als Kontext mal die Darbietungsform der Sachinformationen, mal die in der Aufgabe dargestellte Situation betrachtet wurde. Es zeigt sich, dass jegliche Form der Kontextvariation die Testergebnisse teilweise deutlich beeinflussen kann [11-14].

Schecker und Gerdes [11] berichten weiterhin, dass sich bei den Multiple-Choice-Aufgaben das Antwortmuster, und damit das fachliche Verständnis durch den Kontext (z. B. Golfball oder Fußball) beeinflussen lässt. Dies lässt sich eventuell unter anderem darauf zurückführen, dass Probanden auch bei gleichen zugrunde liegenden physikalischen Beschreibungen sehr unterschiedliche Lösungsansätze bemühen [vgl. 12]. Stewart, Griffin und Stewart [13] verorten diese Befunde in einem größeren Forschungskontext: Demnach ist (neu erworbenes) physikalisches Verständnis bei Novizen durch eine Unbestimmtheit ausgezeichnet, sie lassen sich leicht durch Oberflächenmerkmale der Kontexte irritieren. Erst mit zunehmender Expertise zeigt sich das tiefe Verständnis dadurch, dass Kontexte nicht mehr anhand ihrer oberflächlichen Merkmale, sondern entlang der physikalischen Begründung geordnet werden. Die Autoren nehmen anschließend verschiedene Modifikationen gleichzeitig vor, vom Kontext der Aufgabe bis hin zum Antwortformat, und finden Effekte auf Einzelitemebene, können diese aber aufgrund des Designs nicht bestimmten Veränderungen zuordnen [vgl. 13]. Ähnlich zeigt sich auch in anderen Studien eine gewisse Wechselwirkung zwischen Kontext und Antwortformat, das heißt, dass die Probanden durch die geschlossenen Antwortformate unter Umständen beeinflusst werden [vgl. auch 14]. Es fehlt jedoch bislang eine Untersuchung, die beide Effekte getrennt voneinander systematisch untersucht.

Neueste Arbeiten widmen sich anderen Zugängen der Auswertung auf Item- oder Antwortalternativenebene. Konzentrieren sich frühere Arbeiten dabei noch auf klassische Testanalysen [z. B. 15], werden inzwischen verschiedene alternative Ansätze vorgeschlagen, die alle (sowohl richtige als auch falsche) Antwortalternativen berücksichtigen (können). Dedic, Rosenfield und Lasry verwenden das *Latent Markov Chain Modelling* und zeigen so erstmalig, dass den verschiedenen falschen Antwortalternativen möglicherweise eine gewisse Hierarchie unterliegt [16]. Diese Befunde werden durch die Analyse des FCI mittels Item-Response-Theorie (IRT) tendenziell bestätigt. Zunächst beschränkte sich die Anwendung der IRT auf den Summenscore, es konnte aber gezeigt werden, dass der Test insgesamt bei Dichotomisierung und unter Verwendung der IRT eine reliable Skala darstellt [17, 18]. Neumann und Kollegen haben in der Folge für einzelne Items des FCI – zunächst mittels Expertenrating – eine theoretische Annahme für eine zugrunde liegende Hierarchie der Antwortalternativen im Sinne einer *learning progression* vorgeschlagen [2]. Darüber hinaus konnten Neumann und Kollegen auch zeigen, dass diese theoretische Annahme eine gewisse Gültigkeit besitzt, womit unter Verwendung der IRT und Zugrundelegen der Hierarchie zumindest einige Items erstmals plausibel voll ausgewertet werden konnten [19].

Trotz der dargestellten Diskussionen findet sich bis heute keine ausführliche und fundierte Überarbeitung des FCI, wie sie bereits u. a. von Schecker und Gerdes vorgeschlagen wird [11]. Neue methodische Ansätze konzentrieren sich auf eine Auswertungsstrategie, die ohne Dichotomisierung auskommt und auf diese Weise mehr Informationen über die Probanden generiert. Letztlich bleibt das FCI als Multiple-Choice-Test aber an die Antwortalternativen gebunden. Im folgenden Abschnitt wird zunächst allgemein auf die grundlegende Diskrepanz zwischen offenen und geschlossenen Antwortformaten eingegangen, um diese Diskrepanz schließlich im Kontext des FCI zu betrachten.

2. Offene vs. geschlossene Antwortformate

Im Rahmen größerer Erhebungen wird vor allem aufgrund der Testökonomie und der Auswerteobjektivität ein geschlossenes Antwortformat bevorzugt.

So kann die Testung an vielen Orten und bei vielen Probanden höchst objektiv und effizient durchgeführt werden [20]. Multiple-Choice-Tests ließen sich schon vor der Einführung von Scannern und entsprechender Verarbeitungssoftware sehr schnell und objektiv auswerten [21]. Zurzeit werden zahlreiche Concept Inventories verwendet – wie beispielsweise das FCI [1] – oder verbessert bzw. neu entwickelt – z. B. durch Urban-Woldron und Hopf im Bereich der Elektrizitätslehre [22]. Der überwiegende Teil der Concept Inventories ist aus bereits erwähnten Gründen im geschlossenen Antwortformat gestaltet. Allerdings wird seit längerem vermutet, dass verschiedene Antwortformate unterschiedliche kognitive Prozesse erfordern [23]. Dies würde bedeuten, dass durch die Engführung des Formats Informationen über das Verständnis verloren gehen würden.

Es wird vor allem festgestellt, dass offene Antwortformate eine detailliertere Erfassung der zugrunde liegenden Lösungsprozesse, und damit einen tieferen Einblick in die aufgabenrelevanten kognitiven Konstrukte ermöglichen können [24]. Unter anderem erreichen Probanden in offenen Antwortformaten weniger Punkte als in geschlossenen Antwortformaten; je wichtiger der Test für die Probanden ist, desto größer ist dieser Unterschied [25]. Die genaue Höhe der Differenz zwischen offenen und geschlossenen Aufgaben fällt jedoch in Abhängigkeit von der Ähnlichkeit der Aufgabenstämme sehr unterschiedlich aus, wie in einer Studie gezeigt werden konnte [26].

In der Literatur werden für diese Unterschiede zwischen Ergebnissen in offenen und geschlossenen Tests verschiedene Erklärungen diskutiert. Ein möglicher Grund scheint zu sein, dass die Antwortalternativen in Multiple-Choice-Tests manchen Probanden helfen, die richtige Lösung auszuwählen; so zeigt sich, dass bei freien Antwortformaten neben fachlich adäquaten Vorstellungen auch Alltagsvorstellungen erwähnt werden [27]. Vergleicht man zusätzlich Interviewergebnisse mit Multiple-Choice-Tests und mit Befragungen mit offenen Antwortformaten, unterscheidet sich der Multiple-Choice-Test mitunter deutlich von den anderen beiden, die sehr ähnlich ausfallen [27]. Dabei gehen derartige Untersuchungen davon aus, dass Interviews am ehesten das Verständnis erfassen. Ferner wird diskutiert, ob Probanden durch die Präsentation von falschen Antwortalternativen in Multiple-Choice-Tests während des Tests Fehlvorstellungen erwerben, die sie zur falschen Lösung führen [28, 29].

Die verschiedenen Formate haben auch hinsichtlich der Auswertung der Tests Konsequenzen. Bei Multiple-Choice-Tests werden meistens nur die Ergebnisse ausgewertet, bis hin zur ausschließlichen Verwendung des Summenscores [7]. Somit werden fast zwingend auch allgemeine Lösungsstrategien mit erfasst und die Abbildung verschiedener Lösungsprozesse gelingt nur selten [30]. Gerade die Dichotomisierung und Bildung von Summenscores wird hier als Defizit diskutiert. So schließen Briggs und Kollegen, dass mit Multiple-Choice-Tests die Einschätzung des realen Verständnisses mitunter nur sehr eingeschränkt gelingt [31].

Speziell für das FCI wurden in der Vergangenheit zwei der hier aufgegriffenen Fragestellungen in Teilen untersucht. Rebello und Zollman verglichen in einer Studie vier Items des FCI in geschlossenem Antwortformat mit den gleichen vier Items im offenen Antwortformat [14]. Ein einzelner Proband erhielt allerdings nicht zweimal die gleichen Aufgaben. In dieser Studie unterschieden sich die Summenscores zwar nicht bedeutsam, allerdings berichten die Autoren, dass die Probanden andere falsche Lösungen generierten als durch die Antwortalternativen angeboten [14]. Dies deutet auf Effekte auf Individualebene hin, die aber nicht untersucht wurden – was vermutlich auch der kleinen Stichprobe ($N = 87$) und dem unvollständigen Design geschuldet war. Ein weiterer Grund für die fehlende Differenz könnte aber auch die Auswahl der Aufgaben sein, die nicht entlang möglicher Teilkonstrukte erfolgte [vgl. 14]. Savinainen und Viiri verglichen das FCI dagegen mit inhaltlich verwandten Interviews [32]. Sie stellten fest, dass die Ergebnisse im FCI nur bedingt mit denen im Interview zusammenhingen. Während Probanden mit geringem Verständnis in beiden Instrumenten schlecht abschnitten, gab es einen relevanten Anteil an Personen, die im FCI selbst sehr gute Ergebnisse erzielten, dies in den Interviews aber nicht reproduzieren konnten [32]. Ähnliche Befunde zum Einfluss des Erhebungsformats finden sich auch in anderen Studien, die allerdings nicht das FCI verwenden [z. B. 33].

Demnach ist davon auszugehen, dass zumindest für einen Teil der Schülerinnen und Schüler, Studentinnen und Studenten das Antwortformat des FCI einen Einfluss auf das Testergebnis haben kann. In der hier vorgelegten Studie wird dies systematisch für einen Teil des FCI untersucht. Dabei wird sowohl der Aspekt des Antwortformats als auch der Aspekt des Kontextes berücksichtigt, um beide viel diskutierten Aspekte angemessen aufzugreifen.

3. Forschungsfragen

Vor dem Hintergrund der dargestellten Befunde werden im Folgenden zwei Forschungsfragen untersucht:

a) Inwiefern beeinflusst das Antwortformat des Force Concept Inventories die Ergebnisse der Probanden?

b) In welchem Ausmaß wirkt sich eine Modifikation des Kontextes auf die Bearbeitung durch die Probanden aus?

Beide Fragen wurden in der Vergangenheit zwar schon in Teilen untersucht, wir möchten damit aber die Diskussion um das FCI systematisch mit Evidenz bereichern. In der bisherigen Forschung wurde diskutiert, inwiefern das FCI ein kohärentes Verständnis erfasst, oder ob es aus mehreren Teilskalen

besteht. Ergebnisse dieser Diskussion fanden jedoch keinen Eingang in die Untersuchung der Aufgabeneffekte. Hierdurch konnte es zu einer Konfundierung verschiedener Effekte kommen, sodass letztlich nicht klar war, wodurch die Befunde (nicht) generiert wurden. Insbesondere in der Studie von Rebello und Zollman könnte die Auswahl von inhaltlich und gestalterisch sehr unterschiedlichen Aufgaben dazu geführt haben, dass sich die Ergebnisse aus offenen und geschlossenen Tests nicht unterschieden. Aus diesem Grund wird in der hier vorgelegten Studie zunächst eine sehr enge Inhaltsbegrenzung auf einen Teilaspekt des Kraftverständnisses (relevante Kraftarten) vorgenommen. Darüber hinaus wurden in der Vergangenheit die Kontexte der Aufgaben diskutiert, manchmal ohne die verschiedenen Aspekte sauber zu trennen. In Anlehnung an die Arbeiten von Schecker und Gerdes [11] wird dies hier systematisiert untersucht**.**

4. Anlage der Studie

Diese Studie unterliegt den gleichen Restriktionen wie alle vorher zitierten: Die Kodierung offener Antworten ist mit erheblichem zeitlichen und finanziellen Aufwand verbunden. Aus diesem Grund beschränkt sich die Studie vorerst auf eine Teilauswahl von Items des FCI. Diese Teilauswahl sollte allerdings nach Möglichkeit inhaltlich verwandte Items erfassen, um tatsächlich den Einfluss des Antwortformats untersuchen zu können. Somit ist das Ziel, nicht die gesamte Breite des Kraftkonzepts zu erfassen, sondern nur eine Teilfacette. Schecker und Gerdes gelang es, eine solche Teilskala von sieben Items aus der Originalversion des FCI zu isolieren [11]. In den Items wird nach Kräften gefragt, die in einer bestimmten Situation relevant sind. Ausgehend von dieser Teilskala sowie der zugrunde liegenden Teilaspekte, die Hestenes und Kollegen vorschlagen [1], wurden fünf Aufgaben aus dem Original-FCI von 1992 in der überarbeiteten Übersetzung von Neumann und Kollegen nach Schecker und Gerdes ausgewählt (vgl. [2, 11]). Es wurde nicht die revidierte Fassung des FCI benutzt, da das Golf-Item, das von Schecker und Gerdes als das zentrale Item herausgestellt wird, darin nicht mehr vorkommt [vgl. 11]. Die fünf Items sind (Nummerierung nach [11]):

(1) Zwei Kugeln unterschiedlicher Masse fallen vom Dach.

(5) Eine Stahlkugel wird nach oben geworfen.

(12) Ein Buch liegt auf einem Tisch.

(17) Ein Stein fällt von einem Dach.

(22) Ein Golfball wird abgeschlagen.

Es wird davon ausgegangen, dass diese Items einen sehr ähnlichen Aspekt der Newton‘schen Mechanik erfassen. Es handelt sich in allen Fällen um Kraftarten, aber nicht um die gleichen. Dies soll der Beantwortung der ersten Forschungsfrage dienen. Dabei wurden einige Items mit gleichem inhaltlichen Fokus nicht berücksichtigt, weil sie im offenen Antwortformat eine zeichnerische Antwort erfordert hätten. Die hier vorgestellte Studie ist Teil eines größeren Projekts, dessen Ziel die Software-basierte (und vollautomatische) Evaluation offener Antwortformate ist. Diese basiert jedoch auf latenten semantischen Analysen schriftsprachlicher Antworten, weswegen zeichnerische Lösungen nicht analysiert werden können. Ferner soll die Erhebung mittels eines Online-Surveys durchgeführt werden; auch hier sind zeichnerische Antwortformate nicht ohne größeren Aufwand im Rahmen eines Kleinprojektes zu realisieren.

Die zweite Forschungsfrage geht auf die heterogene Befundlage zur Kontexteinbindung der Aufgaben ein: Wie dargestellt, scheint die kontextuelle Einbindung der physikalischen Inhalte selbst bei identischem Frageinhalt zu unterschiedlichen Lösungen zu führen. Aus diesem Grund wurde für ein Item, das Golf-Item (22), der Kontext bei gleichem Inhalt variiert (vgl. Abbildung 1). Neben dem Golf-Kontext gab es somit vier weitere Kontexte: Fußball, Weitsprung, Angry Birds© und der Wurf eines Schlüssels. Die Fragestellung an die Probanden war immer exakt identisch (vgl. Abbildung 1). Im Idealfall würde man erwarten, dass die Probanden auf alle fünf Fragen exakt gleich antworten, da die physikalische Beschreibung fast identisch ist (auf die Problematik des Magnus-Effekts bzw. des Auftriebs beim Golfball wurde in der Vergangenheit bereits hingewiesen, sie war weder im geschlossenen noch im offenen Antwortformat Teil der verlangten Musterlösung).

Alle Items waren ursprünglich Multiple-Choice-Items. Für das offene Antwortformat mussten daher leichte sprachliche Modifikationen an den Aufgabenstämmen vorgenommen werden, die dann wiederum auch in die geschlossenen Aufgaben übernommen wurden. Das gesamte Testinstrument enthält damit 18 Items, wovon 9 ein offenes Antwortformat und 9 ein geschlossenes Antwortformat haben. Je eine offene und eine geschlossene Aufgabe sind vollkommen identisch. Ferner lassen sich die Aufgaben einteilen in

- fünf Original-FCI-Fragen mit *ähnlichem* Inhalt und
- eine Original-FCI-Frage sowie vier ergänzte Fragen mit *gleichem* Inhalt aber unterschiedlichem Kontext.

Um im Gegensatz zu früheren Studien ein vollständiges Design zu gewährleisten, werden den Probanden immer zuerst die offenen Aufgaben präsentiert. Außerdem dürfen die Probanden nie vor- oder zurückblättern, sonst könnten die Antwortalternativen das Antwortverhalten beeinflussen. Dies stellt jedoch eine Einschränkung des Designs dar.

Der Test wurde sowohl als Papier-und-Bleistift-Instrument für die Erhebung vor Ort vorbereitet als auch für eine darauffolgende Online-Erhebung. Unklar war im Vorhinein, ob die beiden Erhebungstypen überhaupt ohne Weiteres vergleichbar sind. Dies sollte anhand der deskriptiven Resultate geprüft werden. Für die Auswertung wurde in dieser Studie nur auf die dichotomisierten Antworten zurückgegriffen. Das Ziel war, zunächst die Vergleichbarkeit von offenen und geschlossenen Antworten zu untersuchen. Da aber für die Multiple-Choice-Antworten bereits Alternativen, zwischen denen die Probanden wählen *müssen,* vorgegeben waren, blieb unklar, ob diese jenseits einer richt-falsch-Kodierung interpretiert werden können. Dafür wurde, analog zu den meisten vorangegangenen Studien, der Papier-und-Bleistift-Test dichotomisiert und über den Summenscore ausgewertet. Die Antworten für die Fragen des offenen Antwortformats wurden dabei mittels der Antwortalternativen des geschlossenen Antwortformats als Kategorien kodiert. Auf diese Weise wurde vor allem sichergestellt, dass offene und geschlossene Antworten identisch bewertet werden. Der Rater hatte zu entscheiden, welcher Antwortalternative aus dem Multiple-Choice-Test einer schriftlichen Antwort am ehesten entspricht. Die Güte dieser Kodierung wurde mittels einer Zufallsstichprobe und Doppelkodierung überprüft. Die entsprechende Interrater-Reliabilität lag für alle Items im guten bis sehr guten Bereich (0,627 < Cohens κ < 1,000).

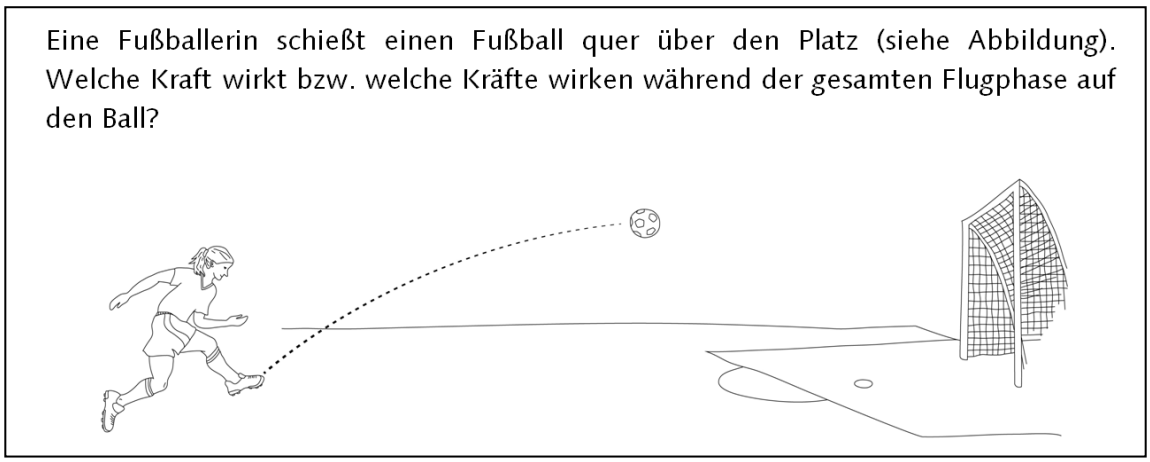

Abb. 1: Beispielaufgabe aus dem modifizierten FCI

5. Ergebnisse

Die Papier-und-Bleistift-Erhebung fand im Wintersemester 2012/13 an der Christian-Albrecht-Universität Kiel statt. Alle Teilnehmer waren Bachelor-Studierende im Fach Physik (sowohl Ein-Fach-Bachelor als auch Zwei-Fach-Bachelor) im 4. und 6. Fachsemester. Die Erhebung fand in der Vorlesungszeit statt, sie nahm 30 Minuten in Anspruch, war aber nach Absprache mit den Verantwortlichen keine Pflicht für die Studierenden. Insgesamt konnten 80 vollständig ausgefüllte Testhefte zur Untersuchung herangezogen werden.

***N* = 80**	**alle Items**	**offene**	**geschlossene**
Cronbachs α	0,91; $p < 0,001$	0,83; $p < 0,001$	0,89; $p < 0,001$
Mean (max = 18)	10,51	5,33	5,19
SD	5,3	2,63	3,16

Tab. 1: Ergebnisse der Papier-und-Bleistift-Erhebung

Aus Tabelle 1 lässt sich entnehmen, dass der Test sowohl insgesamt als auch in Teilskalen eine gute Reliabilität erzielt. Allerdings sind die Daten um den Mittelwert nicht normalverteilt, optisch gleichen die Resultate eher einer Bimodalverteilung: Es gibt eine große sehr leistungsstarke Gruppe und eine etwas kleinere leistungsschwache Gruppe, das Mittelfeld ist kaum besetzt. Diese Aufteilung lässt sich auch post-hoc durch kein erfasstes Merkmal erklären (z. B. Lehramt oder Semester).

Die Online-Erhebung fand noch im selben Semester statt. In Deutschland, Österreich und der Schweiz wurden über Kollegen, Fachschaften, Verbände und soziale Netzwerke Studierende mit dem Fach Physik aufgefordert, sich zu beteiligen. Als Anreiz wurden Gutscheine verlost. Aus der Online-Erhebung konnten insgesamt 292 vollständig ausgefüllte Fragebögen verwendet werden, wobei diese Erhebung naturgemäß unter sehr unkontrollierten Bedingungen stattfand.

Auch hier fiel die Reliabilität gut aus, wenngleich alle Werte geringer waren als für die Papier-und-Bleistift-Erhebung. Der Mittelwert war jeweils höher, bei kleinerer Standardabweichung (vgl. Tabelle 2), und wies wiederum eine Bimodalverteilung auf, die der aus der Paper-Pencil-Studie sehr ähnelt.

N = 292	alle Items	offene	geschlossene
Cronbachs α	0,89; p < 0,001	0,79; p < 0,001	0,87; p < 0,001
Mean (max = 18)	13,27	6,47	6,80
SD	4,70	2,45	2,70

Tab. 2: Ergebnisse der Online-Erhebung

Aufgrund der beiden ungewöhnlichen Verteilungen fiel auch ein Kolmogorow-Smirnow-Test signifikant aus. Es handelt sich also nicht um normalverteilte Daten. Somit sollten Vergleichstests auf der Basis von Summenscores mittels robusterer, non-parametrischer Verfahren durchgeführt werden. Die unterschiedlichen deskriptiven Ergebnisse erzeugen mittels Mann-Whitney-U-Test einen signifikanten Effekt des Erhebungsformats (*geschlossenes Antwortformat: $U = 3{,}845$; $p < 0{,}001$*; *offenes Antwortformat: $U = 3{,}429$; $p < 0{,}001$*). Aus diesem Grund können die beiden Datensätze nicht gemeinsam, sondern müssen einzeln zur Beantwortung der Fragestellungen herangezogen werden.

5.1. Forschungsfrage 1: Das Antwortformat

Im Rahmen der ersten Forschungsfrage ist zu klären, ob sich das Antwortformat auf die Testergebnisse auswirkt. Hierzu wird die dichotomisierte Auswertung herangezogen. Die mittleren Werte lassen sich für die Teilskalen aus den Tabellen 1 und 2 ablesen. Zwei Zugänge sind möglich: In früheren Studien wurde eine Korrelation der Summenscores zwischen offenen und geschlossenen Antworten vorgenommen. Die Bimodalverteilung der realen Daten spricht für die Nutzung von *Spearmans rho* als korrelativem Maß. Dieses Vorgehen berücksichtigt allerdings nicht die Tatsache, dass sich verschiedene Items unterschiedlich verhalten können, da der Summenscore alle Items zusammenfasst.

Zusätzlich zur Korrelation soll daher, die Diskussion zur Entstehung des FCI aufgreifend, eine Faktorenanalyse herangezogen werden. Die Faktoren sind a priori festgelegt, es wird daher eine konfirmatorische Faktorenanalyse verwendet. Im Hinblick auf die spezifische Gestaltung des Tests sollen zwei verschiedene Teilskalen betrachtet werden: 1. Es wird der Einfluss des Antwortformats bei den fünf Items untersucht, die exakt das gleiche Konzept erfassen, nur in unterschiedlichen Kontexten. Hier sind Kontext und Antwortformat tatsächlich die einzig variierten Größen. 2. Es wird außerdem der Einfluss des Antwortformats bei den fünf Items untersucht, bei denen das Konzept leicht variiert wird. Zu vergleichen sind also jeweils zwei Modelle: Das Modell mit nur einer Skala geht davon aus, dass alle 10 Items unabhängig vom Antwortformat das Konzeptverständnis erfassen. Das Modell mit zwei Skalen geht hingegen davon aus, dass sich Aufgaben mit offenem Antwortformat von Aufgaben mit geschlossenem Antwortformat unterscheiden. Welches Modell die Ergebnisse der Studierenden besser erklärt, wird anhand sogenannter Fitindizes entschieden.

Vergleicht man die Summenscores mittels Korrelation, sind die Papier-und-Bleistift-Erhebung und die Online-Erhebung sehr ähnlich. In beiden Fällen zeigt sich nur eine mittlere Korrelation (Papier und Bleistift: Spearmans ρ ($N = 80$) = 0,652; $p < 0{,}001$ und Online: Spearmans ρ ($N = 292$) = 0,649; $p < 0{,}001$). Dies deutet auf der Ebene des Gesamttests darauf hin, dass durch offene und geschlossene Aufgaben unterschiedliche Aspekte erfasst werden.

Die konfirmatorische Faktorenanalyse wurde mithilfe der Statistiksoftware R (Analyse mit lavaan) durchgeführt. Die beiden Tabellen 3 und 4 enthalten die Ergebnisse dieser Analysen. Zu vergleichen sind jeweils die Fitindizes. Dabei ist festzustellen, dass bei den Aufgaben, bei denen nur der Kontext variiert wird, sich das zwei-Skalen-Modell in der Papier-und-Bleistift-Studie als tragfähiger erweist. Der Unterschied ist aber nicht signifikant und nicht alle Fitindizes erfüllen die üblichen Kriterien.

N = 80	df	p (X^2)	CFI	RMSEA	SRMR	AIC	SBIC
Gleiches Konzept, unterschiedlicher Kontext							
offen / geschl.	34	< 0,001	0,852	0,227	0,070	377,888	361,690
eine Skala	35	< 0,001	0,737	0,298	0,134	486,527	471,100
Unterschiedliches Konzept, unterschiedlicher Kontext							
offen / geschl.	34	< 0,001	0,808	0,115	0,103	808,293	792,095
eine Skala	35	< 0,001	0,779	0,122	,100	812,646	797,219

Tab. 3: Konfirmatorische Faktorenanalyse für den Papier-und-Bleistift-Te

$N = 292$	df	p (X^2)	CFI	RMSEA	SRMR	AIC	SBIC
Gleiches Konzept, unterschiedlicher Kontext							
offen / geschl.	34	< 0,001	0,921	0,159	0,046	724,059	734,675
eine Skala	35	< 0,001	0,721	0,293	0,185	1353,507	1363,618
Unterschiedliches Konzept, unterschiedlicher Kontext							
offen / geschl.	34	< 0,001	0,672	0,146	0,091	2872,913	2883,529
eine Skala	35	< 0,001	0,640	0,150	0,096	2892,355	2965,890

Tab. 4: Konfirmatorische Faktorenanalyse für den Online-Test

In der Online-Studie hingegen zeigt sich, dass sich offene und geschlossene Items unterscheiden. Auch die Fitindizes für das Modell liegen nun überwiegend im zufriedenstellenden Bereich. Der CFI liegt über der Anforderung von 0,900 [34]. Der RMSEA liegt zwar oberhalb der Grenze von 0,080 [35] allerdings ist er bekanntermaßen anfällig für die Wahl der Stichprobe (Größe und Verteilung). Hier ist der SRMR aussagekräftiger, weil er robust gegenüber der Stichprobenverteilung ist – der Kennwert liegt unter der geforderten Grenze von 0,100 [36]. Daraus folgt, dass in der Online-Erhebung bei Aufgaben gleichen Inhalts aber mit verschiedenem Kontext offene Antwortformate andere Lösungen hervorbringen als Multiple-Choice-Antworten.

In beiden Erhebungen erreichen das ein- und das zwei-Skalen-Modell gleiche Fitindizes für die Aufgaben, in denen das Konzept leicht variiert. Es wäre folglich naheliegend anzunehmen, dass es sich um eine zugrunde liegende Fähigkeit handelt, die sich in offenen und geschlossenen Antworten nicht unterscheiden lässt.

5.2. Forschungsfrage 2: Der Kontext

Mit der zweiten Forschungsfrage sollte geklärt werden, inwieweit der Kontext sich auf die Bearbeitung auswirkt. Aufgrund der deutlich größeren Stichprobe wurde dies nur anhand der Daten der Online-Erhebung überprüft. Da hier auch der Einfluss des Antwortformats innerhalb der ausschließlich kontextvariierten Aufgaben nachgewiesen werden konnte, wurden die Ergebnisse getrennt nach beiden Antwortformaten ausgewertet.

Bei der Auswertung konnte mittels der Kontexte das Verständnis eingeschätzt werden. Ein Item wurde als ein Rating betrachtet und die Ähnlichkeit der Ratings ermöglichte Rückschlüsse auf die Ähnlichkeit der Kontexte. Diese Ähnlichkeit wurde über die Interrater-Übereinstimmung für die Gesamtheit der Probanden untersucht; da es sich um fünf Items/Ratings bei dichotomen Daten handelte, wurde Fleiss' κ zur Auswertung herangezogen.

Es zeigte sich, dass die zweite Forschungsfrage für die offenen und geschlossenen Antwortformate unterschiedlich zu beantworten ist. Während die Übereinstimmung für die Multiple-Choice-Items perfekt ausfiel (Fleiss' $\kappa = 0{,}861$), lag sie für die offenen Antworten an der unteren Grenze einer substantiellen Übereinstimmung (Fleiss' $\kappa = 0{,}660$). Dies spricht dafür, dass sich der Kontexteinfluss in dieser Studie ausschließlich im offenen Antwortformat zeigt.

6. Diskussion

In der hier vorgelegten Studie wurde für eine ergänzte Auswahl von Aufgaben aus dem Force Concept Inventory (FCI) überprüft, ob sich das Antwortformat auf die Lösungen auswirkt. Dazu wurden in zwei Studien insgesamt 372 Studierende im Bachelor Physik sowohl mit offenen als auch mit geschlossenen Aufgaben befragt. Bei beiden Erhebungen zeigt sich zunächst, dass die Befragung mittels Onlinesystem zu signifikant besseren Ergebnissen führt als die Befragung mit Papier und Bleistift. Dieses Phänomen findet sich auch in Intelligenztests wieder [37]. Mögliche Gründe dafür sind in den unkontrollierten Bedingungen der Online-Erhebung zu suchen. Die Online-Erhebung führt einerseits vermutlich zu einer positiven Auswahl, andererseits stehen die Probanden unter keinem Zeitdruck [37]. Der Unterschied führt jedoch dazu, dass die beiden Datensätze hier nicht aggregiert, sondern getrennt ausgewertet werden müssen.

In der ersten Forschungsfrage wurde untersucht, inwieweit das Antwortformat das Testergebnis beeinflusst. Die einfachste Annahme wäre, dass Studierende mit einem entsprechenden Verständnis zu reproduzierbaren Ergebnissen gelangen, unabhängig vom Antwortformat. Auf Grundlage dieser Annahme wurde das FCI in der Multiple-Choice-Version aus einem offenen Instrument heraus entwickelt [1]. Die hier präsentierten Daten lassen vermuten, dass sich das Antwortformat in bestimmten Testsituationen auf die Ergebnisse auswirken kann: Variiert man bei gleichem zugrunde liegenden Konzept bzw. bei gleichem Inhalt systematisch den Kontext, erweist sich das Erklärungsmodell über das Antwortformat als tragfähig. Es ist sogar dem Modell, das von einem einheitlichen Verständnis ausgeht, überlegen. Allerdings lässt sich dieser Einfluss des Antwortformats in dieser Studie nicht mehr nachweisen, sobald sich die genutzten Aufgaben inhaltlich etwas mehr unterscheiden (vgl. Tab. 4). Ferner gelingt es nur in der größeren Online-Stichprobe den Effekt nachzuweisen.

Insgesamt muss man feststellen, dass die Gütekriterien nur in einem der vier getesteten Modelle die üblichen Kriterien erfüllen, vor allem der RMSEA

erweist sich in unserer Stichprobe als kritisch (vgl. [38]). Ausgehend von den hier vorliegenden Befunden wäre eine umfassendere Untersuchung vor allem mit einer größeren und heterogeneren Population ratsam. Die naheliegende Interpretation der Ergebnisse ist, dass die Variation des benötigten Verständnisses den Einfluss des Antwortformats überlagert, auch wenn er vorhanden ist. Dieser Frage könnte mit elaborierteren Modellen (*nested factor design*) nachgegangen werden. Der in dieser Studie erhobene Datensatz reichte für eine Schätzung entsprechender Modelle nicht aus.

Nichtsdestotrotz lohnt sich der Einsatz des offenen Antwortformats, wie die Analyse des Kontexteinflusses zeigt. Während im Multiple-Choice-Format unterschiedliche Kontexte kaum Auswirkungen haben, deutet die Analyse bei offenen Antworten auf einen Einfluss auf die Erfassung des Verständnisses hin. Dieser Befund steht im Einklang mit Ergebnissen früherer Studien und der Forschung zum Kompetenzerwerb. Wie im theoretischen Hintergrund diskutiert, lassen sich insbesondere Novizen durch Oberflächenmerkmale wie den Kontext leicht beeinflussen. Es zeigt sich, dass Probanden bei offenen Antworten dazu neigen, adäquate und inadäquate Vorstellungen gleichzeitig zu formulieren [27], während sie im geschlossenen Format dazu gezwungen werden, eine Option auszuwählen. Aus diesem Befund ergibt sich gleichermaßen eine Schwierigkeit wie auch eine Erkenntnis.

Einerseits stellt diese Kontextabhängigkeit die Erfassung konzeptuellen Verständnisses mittels kurzer Tests in Frage. Bereits für das vollständige FCI haben mehrere Autoren festgestellt, dass eine plausible Aussage über das Kraftverständnis am ehesten über den Gesamtscore möglich sei [7]. Lässt sich jedoch, wie hier ansatzweise dargestellt, der Gesamtscore durch vermeintlich unwesentliche Veränderungen des Kontextes entscheidend beeinflussen, stellt sich die Frage, welche Anzahl von Kontexten benötigt wird, um eine valide Einschätzung der Probandenfähigkeit zu treffen. Andererseits lässt sich dieser Befund als Ergänzung der Befundlage zur Expertiseforschung verstehen. Dass sich selbst Physikstudierende im 4. und 6. Fachsemester unter Umständen durch die Kontextmodifikationen verunsichern lassen, deutet darauf hin, dass sich durch die Einführungsvorlesung eventuell bei einem Teil der Studierenden kein stabiles Verständnis erzeugen lässt. Dies eröffnet prinzipiell Möglichkeiten für eine Individualisierung der Unterstützung im Fachstudium.

Sowohl die veränderten Items als auch die Rohdaten dieser Studie können per Mail beim Autor der Studie angefragt werden. Sie werden dann nach Rücksprache zur Verfügung gestellt.

7. Literatur

[1] Hestenes, D.; Wells, M. & Swackhammer, G. (1992): Force Concept Inventory. In: The Physics Teacher 30, 141-158.

[2] Neumann, I.; Fulmer, G. W. & Liang, L. L. (2013): Analyzing the FCI based on a Force and Motion Learning Progression. In: Science Education Review Letters, 8-14.

[3] Huffman, D. & Heller, P. (1995): What Does the Force Concept Inventory Actually Measure? In: The Physics Teacher 33, 138-143.

[4] Hestenes, D. & Halloun, I. (1995): Interpreting the Force Concept Inventory. In: The Physics Teacher 33, 502-506.

[5] Heller, P. & Huffman, D. (1995): Interpreting the Force Concept Inventory. In: The Physics Teacher 33, 503-511.

[6] Henderson, C. (2002): Common Concerns About the Force Concept Inventory. In: The Physics Teacher 40, 542-547.

[7] Halloun, I. A. & Hestenes, D. (1985): The initial knowledge state of college physics students. In: Am. J. Phys. 53, 1043-1048.

[8] Gerdes, J. & Schecker, H. (1999): Der Force Concept Inventory. In: MNU 52, 283-288.

[9] Savinainen, A. & Scott, P. (2002): Using the Force Concept Inventory to monitor student learning and to plan teaching. In: Physics Education 37, 53-58.

[10] Lasry, N.; Rosenfield, S.; Dedic, H.; Dahan, A., & Reshef, O. (2011): The puzzling reliability of the Force Concept Inventory. In: Am. J. Phys. 79, 909-912.

[11] Schecker, H. & Gerdes, J. (1999): Messung von Konzeptualisierungsfähigkeit in der Mechanik – Zur Aussagekraft des Force Concept Inventory. In: ZfDN 5, 75-89.

[12] Nieminen, P.; Savinainen, A. & Virii, J. (2010): Force Concept Inventory-based multiple-choice test for investigating students‘ representational consistency. In: Physics Review Special Topics – Physics Education Research 6, 020109.

[13] Stewart, J.; Griffin, H. & Stewart, G. (2007): Context sensitivity in the force concept inventory. In: Physics Review Special Topics – Physics Education Research 3, 010102.

[14] 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.

[15] Martin-Blas, T.; Seidel, L. & Serrano-Fernández, A. (2010): Enhancing Force Concept Inventory diagnostic to indentify dominant misconceptions in first-near engineering physics. In: European Journal of Engineering Education 35, 597-606.

[16] Dedic, H.; Rosenfield, S. & Lasry, N. (2010): Are All Wrong FCI Answers Equivalent? In: AIP Conference Proceedings 1289, 125-128.

[17] Wang, J. & Bao, L. (2010): Analyzing Force Concept Inventory with Item Response Theory. In: Am. J. Phys. 78, 1064-1070.
[18] Planinic, M.; Ivanjek, L. & Susac, A. (2010): Rasch model based analysis of the Force Concept Inventory. In: Physics Review Special Topics – Physics Education Research 6, 010103.
[19] Neumann, I.; Fulmer, G.; Liang, L. L. & Neumann, K. (2012): Investigating Development on a Force and Motion Learning Progression. Paper presented at NARST.
[20] Bühner, M. (2006): Einführung in die Test- und Fragebogenkonstruktion (2nd ed.). München, Don Mills: Pearson Studium.
[21] Braun, H. I.; Bennett, R. E.; Frye, D. & Soloway, E. (1990): Scoring Constructed Responses Using Expert Systems. In: Journal of Educational Measurement 27, 93-108.
[22] Urban-Woldron, H. & Hopf, M. (2012). Entwicklung eines Testinstruments zum Verständnis in der Elektrizitätslehre. In: Zeitschrift für Didaktik der Naturwissenschaften 18, S. 201-227.
[23] Martinez, M. E. (1999): Cognition and the question of test item format. In: Educational Psychologist 34, 207-218. doi: 10.1207/s15326985ep3404_2.
[24] 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.
[25] DeMars, C. E. (2000): Test Stakes and Item Format Interactions. In: Applied Measurement in Education 13, 55-77. doi: 10.1207/s15324818ame1301_3.
[26] Rodriguez, M. C. (2003): Construct Equivalence of Multiple-Choice and Constructed Response Items: A Random Effects Synthesis of Correlations. In: Journal of Educational Measurement 40, 163-184. doi:10.1111/j.1745-3984.2003.tb01102.x
[27] 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
[28] Ha, M.; Nehm R. H.; Urban-Lurain, M. & Merrill, J. E. (2011): Applying Computerized-Scoring Models of Written Biological Explanations across Courses and Colleges: Prospects and Limitations. In: Cell Biology Education 10, 379-393. doi: 10.1187/cbe.11-08-0081
[29] Kang, S. H. K.: McDermott, K. B. & Roediger, H. L. (2007): Test format and corrective feedback modify the effect of testing on long-term retention. In: European Journal of Cognitive Psychology 19, 528-558. doi: 10.1080/09541440601056620
[30] Kuechler, W. L. & Simkin, M. G. (2010): Why Is Performance on Multiple-Choice Tests and Constructed-Response Tests Not More Closely Related? Theory and an Empirical Test. In: Decision Sciences Journal of Innovative Education 8, 55-73.
[31] Briggs, D.; Alonzo, A.; Schwab, C. & Wilson, M. (2006): Diagnostic Assessment With Ordered Multiple-Choice Items. In: Educational Assessment 11, 33-63. doi: 10.1207/s15326977ea1101_2
[32] 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.
[33] Steinberg, R. N. & Sabella, M.S. (1997): Performance on Multiple-Choice Diagnostics and Complementary Exam Problems. In: The Physics Teacher 35, 150-155.
[34] Homburg, C. & Baumgartner, H. (1998). Beurteilung von Kausalmodellen – Bestandsaufnahme und Anwendungsempfehlungen. In: Hildebrandt, L.; Homburg, C.(Hrsg.): Die Kausalanalyse: ein Instrument der empirischen betriebswirtschaftlichen Forschung, Stuttgart 1998, S. 343-369.
[35] Browne, M. W. & Cudeck, R. (1993). Alternative Ways of Assessing Model Fit. In: Bollen, K. A.; Long, J. S. (Hrsg.): Testing Structural Equation Models, Newbury Park et al. 1993, S. 136-162.
[36] Homburg, C.; Klarmann, M.; Pflesser, C. (2008). Konfirmatorische Faktorenanalyse. In: Herrmann, A.; Homburg, C.; Klarmann, M. (Hrsg.): Handbuch Marktforschung, 3. Aufl., Gabler, Wiesbaden, S. 271-304.
[37] Preckel, F. & Thiemann, H. (2003): Online- versus paper-pencil-version of a high potential intelligence test. In: Swiss Journal of Psychology 62, 131-138.
[38] Hu, L. & Bentler, P. M. (1999): Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. In: Structural Equation Modeling: A Multidisciplinary Journal 6, 1-55.

Danksagung

An dieser Stelle möchte ich der Deutschen Forschungsgemeinschaft für die Förderung dieses Projekts danken. Ferner gilt mein aufrichtiger Dank zwei unbekannten GutachterInnen für die hilfreiche Kritik.