Vorhersagemodell zum Studienerfolg im Fach und im Lehramt Physik: Erste Schritte zur Entwicklung eines Online-Self-Assessment

Nikola Schild, Luzy Heindel (geb. Krüger), Philipp Straube, Daniel Rehfeldt & Volkhard Nordmeier

Freie Universität Berlin, Fachbereich Physik, Arnimallee 14, 14195-Berlin

nikola.schild@fu-berlin.de, luzy.krueger@fu-berlin.de, philipp.straube@fu-berlin.de, daniel.rehfeldt@fu-berlin.de, volkhard.nordmeier@fu-berlin.de

Kurzfassung

Im Rahmen der Projekte HeLP (**H**ochschul**e**rfolg im **L**ehramt **P**hysik) und HelP (**H**ochschul**e**rfo**l**g in **P**hysik) soll ein bestehendes Vorhersagemodell zum Studienerfolg erweitert, geprüft und in einem Online-Self-Assessment (OSA) umgesetzt werden. Dieses OSA soll schließlich dazu dienen, Studieninteressierte bei der Entscheidung für ein Physik- (Lehramts-)Studium fundiert und evidenzbasiert zu beraten, um langfristig den hohen Abbruchquoten entgegenzuwirken. Um dies zu ermöglichen, wurden ein kognitiver Kompetenzstest in Mathematik und Physik und ein nicht kognitiver Fragebogen mit Selbsteinschätzungen zu Verhaltensweisen im Studium entwickelt. Beide Testteile sollen in einem Längsschnitt mit StudienanfängerInnen angewendet werden. Hierbei wird *Studienerfolg* hauptsächlich über *Studienzufriedenheit* und den *Verbleib im Studium* operationalisiert. Um die Testergebnisse sinnvoll interpretieren zu können, wurde der kognitive Testteil theoriebasiert in mehrschrittigen Verfahren entwickelt und daraufhin überprüft, inwiefern die Itemschwierigkeiten das Spektrum der Personenfähigkeiten sinnvoll abdecken. Inhalt dieses Artikels soll die Entwicklung und Validierung des kognitiven Testteils sein.
Die Aufgaben sind so konzipiert worden, dass sie die Grundanforderungen in den ersten Semestern repräsentativ abdecken. Im Multiple-Choice-Fragebogenformat wurden die Aufgaben in einem Pilotierungsdurchlauf getestet (N = 313), die Aufgabenschwierigkeiten nach der Item Response Theory (IRT) geschätzt und unpassende Items überarbeitet oder entfernt. Im nächsten Schritt der Testentwicklung wurde die überarbeitete Version des Tests in einer zweiten Pilotierung an der FU Berlin eingesetzt. Die Analyse der Aufgabenschwierigkeiten und Personenfähigkeiten mit Hilfe von 1PL-Modellen ergab eine gute Passung zwischen Aufgaben und Probanden, was für eine Zielgruppenadäquatheit des Instruments spricht. Daher ist davon auszugehen, dass sich dieser Testteil dazu eignet, ihn in der Hauptstudie auf seine Vorhersagekraft zum Studienerfolg prüfen zu können.

1. Zielstellung

Bundesweit lassen sich in den Studiengängen Physik und Physik Lehramt hohe Abbruchquoten verzeichnen [1]. Dies bedeutet für die Studierenden einen persönlichen Rückschlag – auch in Form von Zeit- und Einkommensverlust – und ebenso eine Fehlinvestition von Seiten der Universität [2], sowie einen gesellschaftlichen Malus: einen daraus resultierenden Mangel an Physik-Fach- und Lehrkräften. Daher besteht das Ziel der Forschungsprojekte *HeLP* und *HelP* zunächst darin, passende Erhebungsinstrumente zu konstruieren und die Interpretierbarkeit der Messergebnisse mit Blick auf ihre Vorhersagekraft zum Studienerfolg zu prüfen. In einem zweiten Schritt ist Ziel der Projekte, die Messergebnisse in der Hauptstudie dafür einzusetzen, Studienerfolg vorherzusagen. Auf diesen Ergebnissen aufbauend soll ein Online-Self-Assessment (OSA) für Studieninteressierte gestaltet werden (s. auch Heindel et al., 2015 [3]). Dieser Beitrag fokussiert auf die Entwicklung des *fachspezifischen Kompetenztests*, und zwar insbesondere mit Blick darauf, welche Schritte hinsichtlich der späteren validen Interpretierbarkeit der Ergebnisse berücksichtigt worden sind.

2. Theoretischer Hintergrund

2.1. Studienerfolg und Studienmisserfolg

Der Begriff „Studienerfolg“ wird oft sehr unterschiedlich definiert und bei seiner Erfassung ebenso unterschiedlich operationalisiert. Eine der national verbreiteten Auffassungen wurde von der HIS GmbH geprägt. Heublein et al. [4] definieren den Studienerfolg darüber, dass ein Hochschulstudium abgeschlossen wird. Dies bedeutet, dass Hochschul- oder Fachwechsler ebenso zu „erfolgreich“ Studierenden zählen, wie Standort- und Fachtreue. Allerdings ist es i.d.R. nicht möglich, den Verbleib einzelner Individuen nachzuvollziehen, und damit den Hochschuler-

folg manifest und personenbezogen zu erfassen. Werden also nur diejenigen Personen erfasst, die fach- und standorttreu geblieben sind und das Studium abgeschlossen haben, so ergibt sich, dass der sog. „Schwund“ nicht einkalkuliert wird. Mit „Schwund“ werden hier diejenigen Studierenden bezeichnet, die nicht mehr in demselben Studiengang zu finden sind, aber das Hochschulsystem nicht verlassen haben. Heublein et al. ([5], [6], [4], [1]), begegnen dieser Problematik durch eine bundesweite Messung, die im Querschnitt AbsolventInnen- und AnfängerInnenzahlen vergleicht. Durch dieses Verfahren werden Standort- und Fachwechsler miterfasst, können aber nicht genau identifiziert werden. Kleinere Studien hingegen müssen Hochschulerfolg anders operationalisieren. Ein populärer Ansatz ist hier die Erfassung von Zwischen- und Examensnoten (z. B. [7]). Hierdurch kann Hochschulerfolg quantitativ recht gut erfasst werden, wobei es bisher keine bekannten Evidenzen für die Gleichwertigkeit von guten Noten im Studium und einem Erreichen des Abschlusses gibt.

Die theoretische Grundlage für das hier vorgestellte Vorhaben bildet das Studienerfolgsmodell von Thiel, Veit, Blüthmann, Lepa [8], adaptiert an Physik- und Physiklehramtsstudiengänge von Albrecht [9]. Dieses Modell unterscheidet kategorisch in verschiedene Einflussdimensionen, die zum Studienerfolg oder Studienmisserfolg führen können. Hier wird Studienerfolg in Form von „Studienzufriedenheit“ definiert. Allerdings zeigt sich auch hier das Problem einer nicht hinreichenden Operationalisierung des Studienerfolgs.

Eine weitere Definition des Studienerfolgs ermöglicht die Betrachtung des Studienabbruchs: Nach Heublein et al. [4] finden 63% des Studienabbruchs in Bachelorstudiengängen in den ersten zwei Semestern und nach durchschnittlich 2.3 Semestern statt. Auch Albrecht [9] konnte mit seiner Abbrecherbefragung ähnliche Werte im Bereich Physik (Lehramt) finden (Abbruch nach 2.7 Sem, SD = 1.02). Die vorliegenden Daten legen nahe, dass ein erfolgreiches Absolvieren der ersten drei Semester mit hoher Wahrscheinlichkeit einen Studienabschluss zur Folge hat. Dies bedeutet auch, dass nach drei Semestern ein Studienabschluss (oder auch -abbruch) recht genau geschätzt werden kann.

2.2. Ursachen für einen Studienabbruch

Die Ursachen für die Entscheidung zum Studienabbruch sind vielfältig. Grob lassen sich diese Ursachen in externale und internale unterscheiden. Zu externalen Gründen zählen hier beispielsweise Finanzierungsschwierigkeiten, familiäre Verpflichtungen oder Krankheit, wohingegen internale Ursachen eine mangelnde kognitive Leistungsfähigkeit, fehlende leistungsfördernde Arbeitshaltungen oder mangelnde Leistungsmotivation darstellen können [10]. Durch eine Intervention, wie sie in diesem Projekt angestrebt ist, lässt sich allerdings auf keinerlei externale Ursachen zum Studienabbruch Einfluss nehmen.

Da oft nicht ein einzelner Grund, sondern eine Kombination aus mehreren verschiedenen zum Abbruch führt [9], erscheint es durchaus sinnvoll, eine Art „Frühwarnsystem“ in Form eines OSA zu entwickeln, um möglichst vielen Abbruchgründen frühestmöglich entgegenzuwirken. Im Rahmen eines OSA, das bereits vor Beginn des Studiums bearbeitet werden soll, ist es nicht möglich, Entwicklungen während des Studiums zu berücksichtigen. Hier können lediglich die Eingangsvoraussetzungen der Studieninteressierten erfasst und diese auf ihre Vorhersagekraft zum Studienerfolg geprüft werden. Das Studienerfolgsmodell von [9] führt folgende Eingangsvoraussetzungen auf, die als Einflussfaktoren zum Studienerfolg wirken können:

- Kognitive Fähigkeiten
- Tätigkeit vor Studienbeginn
- Studienwahlmotive
- Informiertheit
- Soziodemografische Variablen (z. B. Alter, Geschlecht, Kinder, angestrebter Abschluss, Leistungskurswahl, Finanzierung des Studiums)

Es hat sich z. B. gezeigt, dass eine zentrale Ursache für die niedrige Studienerfolgsquote im Fach und Lehramt Physik die mangelnde Passung zwischen den Erwartungen der Studieninteressierten und der Studienrealität ist. Dies lässt sich vor allem auf eine mangelnde Informiertheit zu Studienbeginn und daraus resultierende Erwartungen, die nicht erfüllt werden konnten, zurückführen [9]. Diese Bereiche, also Erwartungen, Informiertheit und soziodemographische Variablen werden als *nicht-kognitive Faktoren* der Eingangsvoraussetzungen aufgeführt.

Eine weitere Ursache für einen Studienabbruch sind die inhaltlichen Studienanforderungen [9], was sich auf mangelnde kognitive Fähigkeiten zurückführen lassen könnte. In der Abbrecherbefragung [9] hat sich gezeigt, dass der am häufigsten genannte Grund für einen Studienabbruch Leistungsschwierigkeiten sind. Neben der Note der Hochschulzugangsberechtigung (HZB-Note) erlauben fachspezifische Kompetenztests eine genauere Einschätzung der kognitiven Eingangsvoraussetzungen [11] und werden in unserem Projekt als *kognitive Faktoren* ins Vorhersagemodell übernommen.

Der Hauptabbruchgrund im Physik Fach- und Lehramtsstudium waren Leistungsschwierigkeiten [9]. Da ein Physik Fach- und Lehramtsstudium fachlich besonders stark physikalische und mathematische Grundkompetenzen voraussetzt, sind diese auch für ein Vorhersagemodell relevant und können als Prädiktor zum Studienerfolg angenommen werden. Da bisher ausschließlich die HZB-Note für die Vorhersage von Studienerfolg erhoben wurde [9], ist offen, inwiefern ein fachspezifischer Kompetenztest zur Vorhersage von Studienerfolg geeignet ist.

3. Methoden

3.1. Operationalisierung des Studienerfolgs

Aufgrund der zuvor erläuterten Lücken in der Operationalisierung von Studienerfolg soll in diesem Forschungsvorhaben der Studienabschluss über den Studienverbleib nach dem dritten Studiensemester, die Studienzufriedenheit und die Modulnoten zum selben Zeitpunkt den *Studienerfolg* definieren. Auch hier können Fach- und Standortwechsler, sowie nach dem ersten Semester hinzugekommene Studierende nicht miterfasst werden, allerdings können hierdurch zumindest beständig Studierende identifiziert werden.

Aufgrund der inkrementellen Validität fachspezifischer Kompetenztests [11] und der besonderen Relevanz von mathematischen und physikalischen Grundkompetenzen im Physikstudium, wird der fachspezifische Test Grundkompetenzen in Mathematik und Physik abfragen.

3.2. Vorhersage des Studienerfolgs durch kognitive Faktoren

In einer Längsschnittstudie sollen StudienanfängerInnen des Fachs und des Lehramts Physik bis zum Ende des dritten Semesters begleitet werden. Die kognitiven Fähigkeiten sollen in einem kurzen fachspezifischen Kompetenztest im Multiple-Choice-Format zu Studienbeginn erhoben und zum Ende des dritten Fachsemesters mit einem Verbleib im Studium, der Studienzufriedenheit und den Modulnoten verglichen werden. Um dies zu ermöglichen, wird in jeder Befragung ein personenspezifischer Code miterhoben. Diese Vorgehensweise ermöglicht es, diejenigen Studierenden, die zu allen Erhebungszeitpunkten anwesend sind zu erfassen. Problematisch an dieser Herangehensweise ist, dass eine nichtvorhandene Befragung zum letzten Erhebungszeitpunkt nicht zwangsläufig als Studienmisserfolg interpretiert werden kann, da für eine Nichtteilnahme zum zweiten Erhebungszeitpunkt ganz verschiedene Gründe vorliegen können (Studienabbruch, Fachwechsel, Standortwechsel, Auslandssemester, Elternzeit, Krankheit, keine Bereitschaft, an der Befragung teilzunehmen u.v.m.). Daher können nur Datensätze, die sich einander zuordnen lassen, als Erfolg interpretiert und keine Aussage über Misserfolg getroffen werden.

Da in dieser Studie Leistungsdaten erhoben werden sollen, stellt sich das Fragebogenformat hier als das geeignetste dar. Zusätzlich wird seit einigen Jahren an der Freien Universität Berlin eine Längsschnittstudie zur Studienzufriedenheit durchgeführt, die in Form eines Fragebogens vorliegt. Die Kombinierbarkeit beider Befragungen stellt einen ökonomischen Vorteil dar.

3.3. Konzeption des Kompetenztests

Wie oben beschrieben, wurden kognitive Fähigkeiten in den Bereichen Mathematik und Physik als wichtige Eingangsvoraussetzung für das Studium angenommen. Bei der Konzeption der Mathematik- und Physikaufgaben wurde a priori die Annahme zugrunde gelegt, dass es sich bei der zukünftigen Zielgruppe um Personen handelt, die über eine Hochschulzugangsberechtigung oder vergleichbare Qualifikation verfügen, aber keine darüber hinaus gehenden mathematischen oder physikalischen Kompetenzen haben, da davon ausgegangen werden kann, dass ein großer Anteil der StudienanfängerInnen direkt nach dem Schulabschluss ein Studium antritt [9]. Daher sollte bei der Konzeption der Aufgaben das Abiturwissen nicht überschritten werden. Um den Übergang zwischen Schule und Hochschule möglichst gut abbilden zu können, wurden die KMK-Bildungsstandards [12], einige aktuelle Rahmenlehrpläne (u.a. [13], [14]), Schulbücher (u.a.[15], [16], [17], [18]) und die Mathematikzentralabituraufgaben (Grundkurs, Berlin Brandenburg) der vergangenen Jahre (Mathematik Grundkurs, 2011; 2012, 2013) untersucht, sowie die Inhalte verschiedener Brückenkurse und der Anfangsvorlesungen der Physik (erstes und zweites Semester) an der Freien Universität Berlin. Die Schnittmenge der Rechercheergebnisse der schulischen und der universitären Seite wurden dann als relevant angenommen. Dieser relevante Inhaltsbereich lässt sich folgendermaßen darstellen:

Mathematik: Analysis, analytische Geometrie, Algebra (im Sinne der Anwendung von Rechnungen)

Physik: Mechanik, Elektrodynamik, Thermodynamik

Einige wichtige Themenbereiche der Mathematik und der Physik der Oberstufe wurden nicht in den relevanten Kanon übernommen, wie beispielsweise die Stochastik oder die Optik. Beide Themengebiete spielen zwar in der Schule eine große Rolle, kommen aber in der Studieneingangsphase nur marginal vor, sind daher gemäß dem „Drei-Semester-Kriterium“ (s.o.) nicht studienerfolgskritisch.

Zur Konzeption der Aufgaben wurden drei verschiedene Dimensionen berücksichtigt (Abb. 1): Der *Themenbereich*, das *Anforderungsniveau* und die postulierte *Aufgabenschwierigkeit*.

Unter *Themenbereich* werden hier die relevanten Inhaltsbereiche verstanden. Das *Anforderungsniveau* kennzeichnet (wie in den KMK-Standards vorgegeben), die Unterteilung zwischen deklarativem, angewandtem und transferiertem Wissen. Anhand der Recherche der Bildungsstandards und Rahmenlehrpläne konnten wir die Aufgaben sinnvoll in deklaratives und Anwendungswissen unterteilen. Da die StudienanfängerInnen sehr unterschiedliche Hintergründe haben, ist Transferwissen vermutlich nicht valide und ökonomisch messbar und wird daher in diesem Test nicht berücksichtigt. A priori wurde versucht, verschiedene *Aufgabenschwierigkeiten* zu erzeugen, diesich durch die Anzahl der Lösungsschritte und den

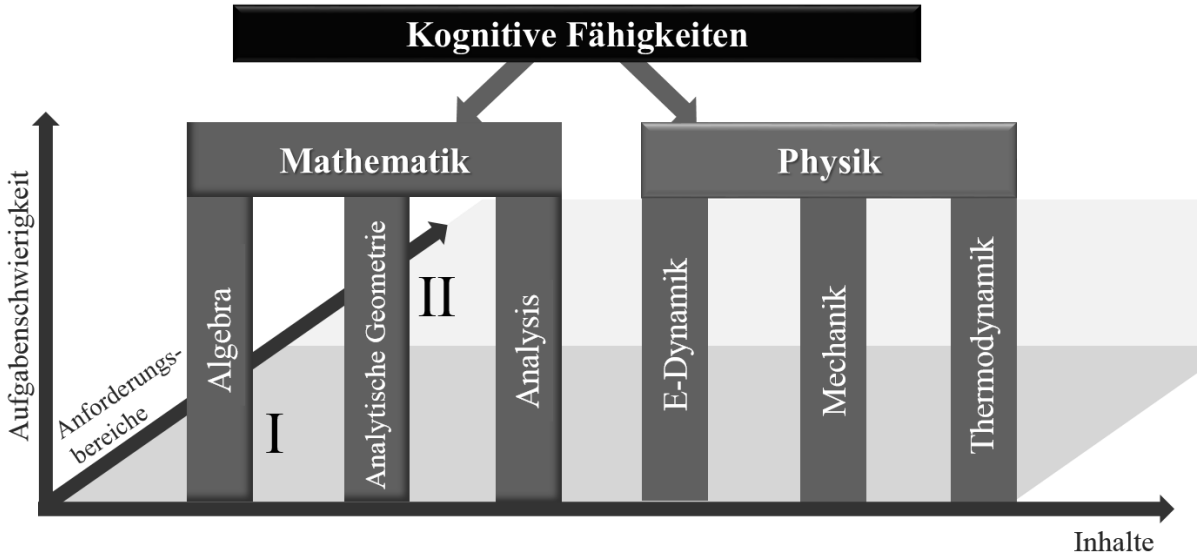

Abb. 1: Modell der drei Dimensionen der Testaufgaben: *Themenbereich*, *Anforderungsniveau* und postulierte *Aufgabenschwierigkeit*.

Stellenwert im Lehrplan auszeichnete. Hier gab es eine Klassifizierung in drei Schwierigkeitsbreiche von „leicht“ (z. B. einschrittige Aufgabe zum Kürzen eines Bruchs), „mittel“ (Termumformung in etwa 3 Schritten) und „schwer“ (Ableiten einer Funktion in mehreren Schritten). Die Einstufung in Schwierigkeiten sollte hier als grobes Maß gelten, da aufgrund der Itemfülle zunächst zwei verschiedene Fragebögen eingesetzt werden mussten und sichergestellt werden sollte, dass beide in etwa den gleichen Anspruch haben. Die tatsächliche Itemschwierigkeit sollte im Anschluss über die Ergebnisse berechnet werden. Die Aufgaben wurden entsprechend der drei genannten Dimensionen kontinuumsabdeckend konzipiert und in ein Multiple-Choice Format [19] mit je zwei bis fünf Antwortmöglichkeiten überführt. Um die Ratewahrscheinlichkeit zu reduzieren, wurde bei jeder Aufgabe die Möglichkeit, „weiß nicht“ anzukreuzen, eingeräumt.

Da in den Naturwissenschaften eine Abbruchquote von 39% zu erwarten ist [1] und hinzu noch ein Schwund von Studierenden kommt, kann damit gerechnet werden, dass zwischen 30 und 60 % der zu Studienbeginn eingeschriebenen Personen keinen Abschluss in demselben Studiengang erreichen [9]. Geht man davon aus, dass mathematische und physikalische Grundkompetenzen eine Vorhersage zum Studienerfolg leisten, ist zu erwarten, dass der Ausgang von sehr guten und von sehr schlechten Ergebnissen recht klar ist. Daher ist gerade im mittleren Leistungsbereich ein sehr genaues Maß nötig, um vermeintliche AbbrecherInnen von nicht-AbbrecherInnen zu differenzieren. Somit ist es sinnvoll, einen Test zu entwickeln, der insgesamt eine mittlere Schwierigkeit für die Befragten aufweist.

3.4. Vorgehen bei der zweistufigen Pilotierung des Kompetenztests

Der Kompetenztest sollte in einem zweistufigen Verfahren pilotiert werden. Im ersten Pilotierungsschritt sollte eine starke Itemselektion stattfinden, um den Test ökonomisch gut umsetzbar zu machen und überprüft werden, inwieweit die Items mathematische und physikalische Grundkompetenzen abbilden. Hierzu wurden an der TU Berlin StudienanfänerInnen aus technischen Studiengängen befragt. Im zweiten Pilotierungsschritt wurden aus den Resultaten der ersten Pilotierung selektierte und veränderte Items im 1. Fachsemester Physik (Fach und Lehramt) an der FU Berlin eingesetzt. Aus den erhobenen Daten wurden dann die letztendlichen Aufgabenschwierigkeiten geschätzt.

3.5. Auswertung mit der IRT

Um die Itemgüte in beiden Pilotierungsschritten zu schätzen, wurden die erhobenen und bereinigten Daten mit dem Rasch-1PL-Modell modelliert. Hierzu wurde ConQuest 2.0 eingesetzt. Da jede Teilstichprobe die minimale Anzahl an Probanden (100 Personen) überschritt, konnte das Verfahren zur Bestimmung der Aufgabenschwierigkeiten angewendet werden [20]. Der Test soll kognitive Fähigkeiten messen, wurde aber bereits bei der Konzeption in zwei Teilbereiche, nämlich Mathematik und Physik, unterteilt. Um zu überprüfen, ob beide Testteile dasselbe Konstrukt messen, wurde ein Dimensionsvergleich vorgenommen [21]. Im ersten Modell wurde angenommen, dass beide Teile auf einen latenten Faktor laden, im zweiten Modell wurde angenommen, dass beide Teile zwei unterschiedliche latente Faktoren laden. Hierzu wurden für beide Testhefte je mehrere Vergleichskriterien berücksichtigt (AIC, BIC, CAIC, χ^2-Differen-

zentest [21]. Für beide Testhefte zeigten alle vier Vergleichskriterien, dass ein zweidimensionales Modell zu bevorzugen ist. Die beiden χ^2-Differenzentests waren mit Werten << .05 sogar höchstsignifikant [21]. In beiden Tests korrelieren die Dimensionen *Mathematik* und *Physik* mit jeweils > .6. Dies kann dahingehend interpretiert werden, dass es sich um zwei korrelierte Konstrukte handelt, die aber beide einen eigenen Beitrag zur Varianzaufklärung leisten. Alle berichteten Ergebnisse beruhen deshalb auf dem zweidimensionalen Modell.

Mit dem zweidimensionalen Modell wurden anschließend die Itemschwierigkeiten und Personenfähigkeiten geschätzt. Es wurden pro Person je fünf *Plausible Values* geschätzt. Es wurden die itemcharakteristischen Kurven und die Kennwerte der Aufgaben auf Auffälligkeiten untersucht. Hierbei wurden Aufgaben als auffällig erachtet, wenn sie eine Schwierigkeit von > $|\mathbf{2}|$ hatten, Infit- und Outfitmaße, sowie T-Werte stark voneinander abwichen oder außerhalb der angegebenen Grenzen waren oder die charakteristische Kurve stark vom Erwartungswert abwich. Gab es pro Aufgabe mehrere Auffälligkeiten oder eine sehr ähnliche Aufgabe mit besseren Kennwerten, so wurden die Aufgaben aus dem Itempool entfernt.

Im Fall der ersten Pilotierung wurde der Itempool, wenn es keine alternative Aufgabe gab, im Rahmen eines Expertengesprächs überarbeitet. Dieser neue Itempool wurde dann für die zweite Pilotierung in ein neues, verbessertes Testheft überführt.

3.6. Umgang mit Missings

Die Bereinigung der Daten wurde mit SPSS Statistics 22 durchgeführt. Bei den Missings wurde in die Kategorien „nicht ausgefüllt“ und „nicht eindeutig“ (z. B. zwei Kreuze gesetzt) unterschieden. Für die Auswertung mussten die Items dichotomisiert werden. Hierbei wurden alle richtig gelösten Items mit einer „1“ kodiert und die falschen Antworten als „0“. Dabei wurden auch die Antworten „weiß nicht“ und mehrdeutige Antworten als falsch interpretiert. Nicht ausgefüllte Items wurden in den Fällen als falsch interpretiert, in denen man nicht auf einen Zeitmangel schließen konnte. D. h., wurden mindestens die letzten beiden Items und die anschließenden Soziodemographie Items nicht ausgefüllt, wurde dies als „Zeitmangel“ interpretiert. Datensätze, auf die das zutraf, wurden aus der Analyse ausgeschlossen. Da es sich hierbei um je drei Personen pro Testheft, also weniger als 5% des Datensatzes, handelte, wird dieses Vorgehen als akzeptabel erachtet [22].

3.7. Pilotierung 1

Zur Pilotierung wurden die StudienanfängerInnen im Mathematikbrückenkurs der TU Berlin befragt. Der Auswahl der Kohorte lag die Annahme zugrunde, dass Mathematik- und Physikaufgaben für alle MINT-StudienanfängerInnen ähnlich schwierig sind, unabhängig davon, ob sie Physik oder einen Ingenieursstudiengang o. ä. belegen werden.

Die entwickelten Aufgaben wurden für die Pilotierung auf zwei Testhefte aufgeteilt, die je 12 Mathematik- und 9 Physikaufgaben beinhalteten. Beide Testhefte hatten keine gemeinsamen Aufgaben, wurden allerdings so zusammengestellt, dass sie die Inhalte, Anforderungsbereiche und das festgelegte Schwierigkeitsmaß möglichst gleichermaßen abdecken. Um Ermüdungs- und Reihenfolgeneffekte auszuschließen, wurde außerdem jedes Testheft in zwei verschiedenen Versionen (Mathematik- und Physikteil vertauscht) ausgegeben [21]. In der IRT-Modellierung wurden beide Testhefte getrennt voneinander ausgewertet, auch wenn die Testhefte zufällig in der gleichen Stichprobe verteilt wurden und somit anzunehmen ist, dass die mittlere Personenfähigkeit in beiden Gruppen sehr ähnlich ist.

3.8. Pilotierung 2

Die Ergebnisse der Pilotierung der Fragen wurden nun in ein neues Testheft mit 16 Physik- und 20 Mathematikaufgaben überführt. Der Test wurde wiederum randomisiert. Das Testheft wurde nun bei den StudienanfängerInnen des Fachs und Lehramts Physik an der FU Berlin eingesetzt (s. o.). Um möglichst eine Vollerhebung zu generieren, wurden zwei verschiedene Erhebungszeitunkte gewählt. Der erste Erhebungszeitpunkt war zu Beginn des Mathematikbrückenkurses für Physikstudierende, der zweite war zu Beginn der Vorlesungszeit in der Experimentalphysikvorlesung, an der alle Studierenden laut Studienverlaufsplan im 1. Semester teilnehmen. Die Testdauer betrug in etwa 25 Minuten.

4. Ergebnisse

4.1. Stichprobe der Pilotierung 1

Insgesamt wurden N = 313 StudienanfängerInnen befragt. Dabei haben N = 151 das erste Testheft und N = 162 das zweite Testheft ausgefüllt. In die Datenanalyse wurden dann N = 148 und N = 159 Personen übernommen (siehe 3.6.: Umgang mit Missings). Durchschnittlich waren die Studierenden 21 Jahre alt (SD = 4). 61% der Befragten waren männlich; 30% gaben an, weiblich zu sein. Dabei waren 7% der ProbandenInnen für den Studiengang Physik eingetragen, die restlichen für andere, hauptsächlich ingenieurswissenschaftliche Studiengänge.

4.2. Stichprobe der Pilotierung 2

An der Befragung haben N = 172 Personen teilgenommen. Von den 39 Personen, die an beiden Befragungen teilgenommen haben, wurde nur das erste ausgefüllte Testheft gewertet, da davon auszugehen ist, dass durch die Teilnahme am Brückenkurs nicht mehr die Studieneingangsvoraussetzungen gemessen werden. 8 Personen mussten aus der Analyse ausgeschlossen werden (siehe 3.6 Umgang mit Missings). Da es sich hier um 4.7% der Stichprobe handelt, kann dieses Vorgehen gerade noch akzeptiert werden [22]. Dies ergab eine auszuwertende Stichprobe von N = 164.

Estimate	Error	wMNSQ	T
-2.10	0.24	.99	0.0

Tab. 1*: Kennwerte des Beispielitems Version 1*

Eine Person mittlerer Physikfähigkeit hätte eine Wahrscheinlichkeit von fast 90%, diese Aufgabe richtig zu lösen (siehe Abb. 3)

(mit: wMNSQ: $0.80 \leq \text{wMNSQ} \leq 1.20$; T:$-2{,}00 \leq \text{T} \leq 2{,}00$ [23])

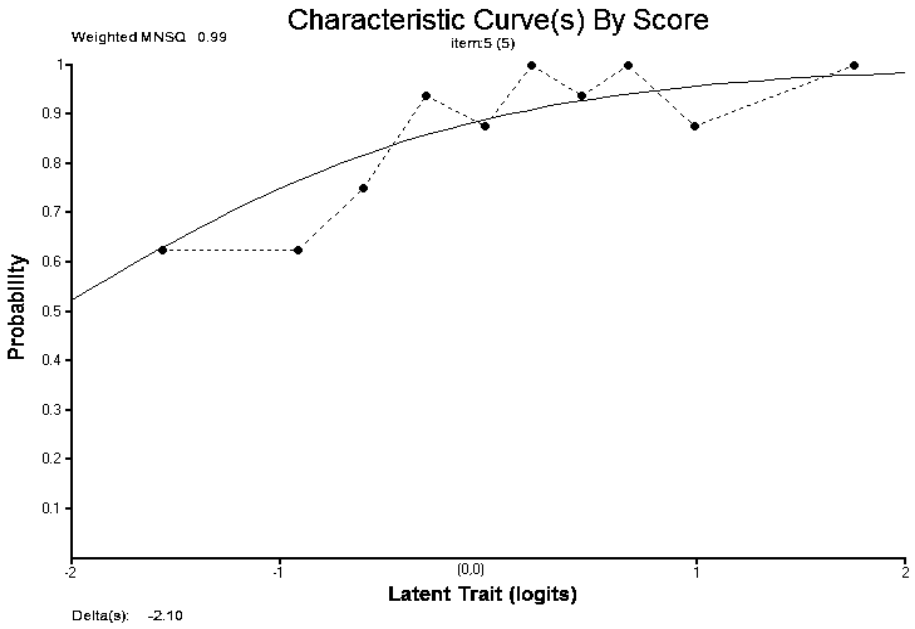

Abb. 2: Itemcharakteristische Kurve zum Beispielitem Version 1

Der Altersdurchschnitt war 21 Jahre (*SD* = 4). Es gaben 59% an männlich zu sein und 27% weiblich. 29% der TeilnehmerInnen gaben an, Physik im Mono Bachelor zu studieren, 26% Physik im Lehramt und 39% ergaben sich aus sonstigen Studiengängen, hauptsächlich Meteorologie und geologische Wissenschaften.

4.3. Ergebnisse der Pilotierung 1

Insgesamt wurden 52% aller Aufgaben richtig gelöst. Diese deskriptive Statistik liefert einen ersten Hinweis darauf, dass es sich bei dem vorliegenden Kompetenztest um ein mittelschweres Format handelt. Die Analyse der Ergebnisse ergab eine Reduktion auf 16 Physik- und 20 Mathematikitems.

4.4. Beispielaufgabe der Pilotierung 1

Eine Physikaufgabe, die nach dem dritten Newtonschen Axiom fragte (sinngemäß: „Welche der beiden Figuren erfährt die größere Kraft?“), hatte eine besonders niedrige Itemschwierigkeit.

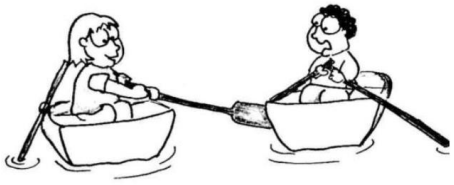

Abb. 3: Darstellung zum Beispielitem Version 1

Die Kennwerte in *Tabelle 1* weisen darauf hin, dass dieses Item zwar eine Messung ermöglicht, aber, wie schon in der Abb. 3 sichtbar, sehr einfach ist (Estimate: -2.10). Daraufhin wurde das Item in seinem Schwierigkeitsgrad erhöht, indem eine weitere Person einem der Boote hinzugefügt wurde (Abb. 4), was zu einer Erhöhung der Itemschwierigkeit führte (vgl. Abschnitt „Ergebnisse der geänderten Aufgabe“).

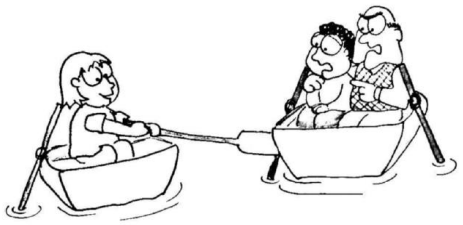

Abb. 4: Darstellung zum Beispielitem Version 2

4.5. Ergebnisse der Pilotierung 2

Insgesamt wurden 49% der Aufgaben richtig gelöst. Die Raschanalyse ergab, dass sich zwei Aufgaben nicht sinnvoll interpretieren lassen. Die charakteristischen Kurven dieser Aufgaben lieferten Hinweise darauf, dass teilweise eine höhere Personenfähigkeit zu einer geringeren Lösungswahrscheinlichkeit der Aufgabe führen kann. Hierbei handelte es sich bei einer der beiden Aufgaben um die einzige Aufgabe im Be-

reich Thermodynamik in der Physik. Da die Thermodynamik im Vergleich zu Mechanik und Elektrodynamik in der Studieneingangsphase eine eher untergeordnete Rolle spielt, konnte diese Aufgabe zugunsten der späteren Messbarkeit der anderen Themenbereiche aus dem Itempool ausgeschlossen werden. Bei der anderen Aufgabe, die keine sinnvollen Kennwerte lieferte, handelte es sich um eine Mathematikaufgabe zur analytischen Geometrie, die durch ihren Ausschluss das inhaltliche Spektrum des Instruments nicht beeinträchtigte. Zusätzlich wurde eine weitere Mathematikaufgabe aus dem Itempool ausgeschlossen, da diese inhaltlich einer anderen Aufgabe stark ähnelte. Da die andere Aufgabe besser ins Raschmodell passte, wurde die schlechter passende entfernt, insbesondere um die Testlänge zu reduzieren.

Item	Estimate	wMNSQ	T
P1	-1,381	1,03	0,3
P2	-0,812	0,92	-1,0
P3	-1,987	0,94	-0,4
P4	0,567	**0,82**	**-2,5**
P5	-0,919	1,07	0,8
P6	0,114	0,93	-1,0
P7	-1,597	1,06	0,5
P8	-0,338	0,93	-0,9
P9	1,046	1,11	1,2
P10	-0,777	1,00	0,0
P11	1,191	1,10	1,1
P12	1,267	0,83	-1,9
P13	1,191	1,02	0,3
P14	1,118	0,98	-0,2
P15	0,242	**1,15**	**2,1**
M1	-2,599	1,01	0,1
M2	1,068	0,96	-0,4
M3	-2,260	1,09	0,6
M4	1,254	0,99	-0,1
M5	1,939	0,96	-0,3
M6	0,198	1,13	1,8
M7	-0,992	0,89	-1,3
M8	-0,609	1,09	1,1
M9	1,178	1,17	1,8
M10	-3,247	0,95	-0,1
M11	-1,102	1,03	0,3
M12	-0,410	1,06	0,8
M13	0,390	**0,83**	**-2,4**
M14	-0,216	1,04	0,5
M15	1,750	1,03	0,3
M16	0,294	**0,84**	**-2,3**
M17	1,178	0,86	-1,5
M18	0,822	0,91	-1,1
M19	-0,025	**1,20**	**2,6**

Tab. 2: *Ergebnisse der Raschanalyse der Pilotierung 2 nach Ausschluss zweier Items*

Die in der *Tabelle 2* dargestellten Daten liegen hauptsächlich im gut interpretierbaren Bereich [23]. Zwei Items mussten, wie oben beschrieben, aus der Analyse ausgeschlossen werden und sind daher in der vorliegenden Darstellung nicht mehr vorhanden. Lediglich drei Items weisen T-Werte ($-2{,}00 \leq T \leq 2{,}00$) [23] im Overfit-Bereich auf [23]. Dies bedeutet, dass sie einen Aufklärungsbeitrag leisten, wenn auch einen schlechteren, als Items, die einen T-Wert im Normbereich aufweisen. Daher können diese drei Items beibehalten werden. Zusätzlich zeigten sich auffälligen Werte beim letzten Item in der Tabelle (M19). Da hier das Infitmaß ($0.80 \leq \text{wMNSQ} \leq 1.20$) [23] gerade noch innerhalb der Grenze ist und durch den Ausschluss dieses Items der Test inhaltlich beeinträchtigt werden würde, wurde dieses Item beibehalten.

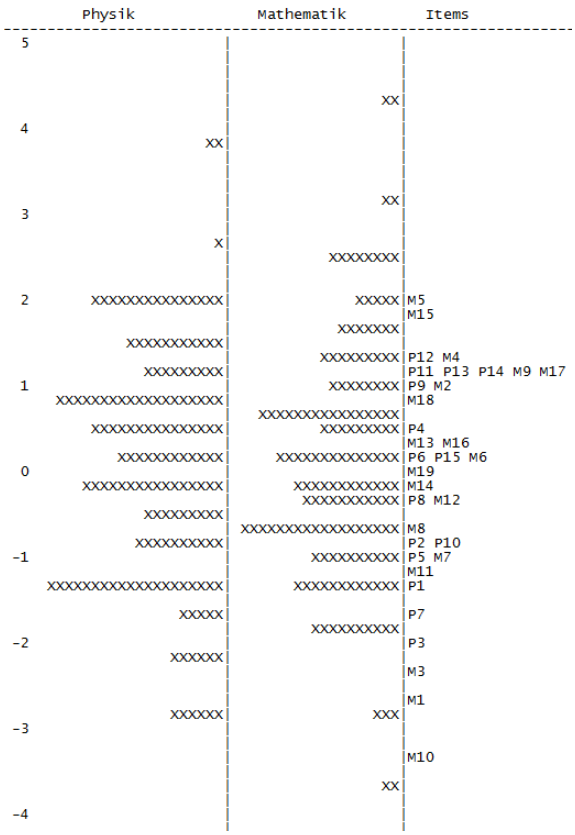

Abb. 5: Darstellung der Personenfähigkeiten in den Dimensionen *Mathematik* und *Physik*, sowie die Aufgabenschwierigkeiten in einer Wrightmap.

Aus der Wrightmap (Abb. 5) wird deutlich, dass besonders der mittlere Fähigkeitsbereich bei den Probanden, sowie der mittlere Schwierigkeitsbereich bei den Items besonders dicht abgedeckt sind. Insbesondere der obere Randbereich wird durch die Aufgaben nicht gut abgedeckt, was jedoch mit Blick auf die Vorhersage unproblematisch ist, da eine Differenzierung im mittleren Leistungsbereich angestrebt wird.

4.6. Ergebnisse der geänderten Aufgabe

Die oben beschriebene Aufgabe wurde aufgrund ihres zu niedrigen Schwierigkeitsgerades nach der ersten Pilotierung geändert. In *Tabelle 3* sind die Kennwerte der Aufgabe vor der Änderung (Messung an der TU Berlin) und nach der Änderung (Messung an der FU Berlin) dargestellt. Die Aufgabe ist Annahmen-konform schwieriger geworden, hat nun eine Schwierigkeit von Estimate_1 = 1,60 statt Estimate_2= -2,10. Die weiteren Werte liegen im gut interpretierbaren Bereich.

Version	Estimate	Error	wMNSQ	T
1 (TU)	-2.10	0.24	.99	0.0
2 (FU)	-1.60	0.12	1.06	0.5

***Tab. 3**: Ergebnisse des Beispielitems vor und nach der Änderung*

5. Fazit

Die Ergebnisse der Vorstudie zeigen, dass sich die konzipierten Aufgaben dazu eignen, zwischen fachspezifischen Leistungen in Mathematik und Physik, die in der Studieneingangsphase essentiell sind, zu differenzieren. Da der entwickelte Test besonders viele Items mit mittlerer Itemschwierigkeit beinhaltet und die Vorhersage von Studienerfolg insbesondere bei mittelmäßig geeigneten Eingangsvoraussetzungen einer möglichst genauen Grundlage bedarf, ist zu vermuten, dass er als Prädiktor im Bereich kognitiver Fähigkeiten gut zwischen vermeintlichen AbbrecherInnen und NichtabbrecherInnen differenziert.

6. Ausblick

Der vorliegende fachspezifische Kompetenztest hat sich durch die verschiedenen Voruntersuchungen als sinnvoll erwiesen, um ihn als möglichen Prädiktor zum Hochschulerfolg zu nutzen. Der Test wird zu Beginn des Wintersemesters 15/16 zur Haupterhebung eingesetzt. Drei Semester später soll anhand des Verbleibs der Befragten im Studium und deren Studienzufriedenheit, sowie Modulnoten, Rückschlüsse auf die Vorhersagekraft dieses vorgestellten Tests gezogen werden.

7. Literatur

[1] Heublein, U., Richter, J., Schmelzer, R. & Sommer, D. (2012). Die Entwicklung der Schwund- und Studienabbruchquoten an den deutschen Hochschulen. Statistische Berechnungen auf der Basis des Absolventenjahrgangs 2010.

[2] Schiefele, U., Streblow, L. & Brinkmann, J. (2007). Aussteigen oder Durchhalten. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 39 (3), 127–140.

[3] Heindel, L., Schild, N., Rehfeldt, D. & Nordmeier, V. (2015). Entwicklung eines Online-Tools zur Studienfachwahl Physik / Lehramt Physik. PhyDid B, 2015 (im Druck)

[4] Heublein, U., Hutzsch, C., Schreiber, J., Sommer, D., Besuch, G., (2010). Ursachen des Studienabbruchs in Bachelor- und in herkömmlichen Studiengängen. Ergebnisse einer bundesweiten Befragung von Exmatrikulierten des Studienjahres 2007/08.(HIS: Forum Hochschule 2, 2010). Hannover: Hochschulinformations-System.

[5] Heublein, Schmelzer, R., Sommer, D., Wank, J. (2008). Die Entwicklung der Schwund- und Studienabbruchquoten an deutschen Hochschulen. Statistische Berechnungen auf der Basis des Absolventenjahrgangs 2006.(HIS: Projektbericht Mai 2008). Hannover: Hochschulinformations-System.

[6] Heublein, U., Spangenberg, H., Sommer, D. (2003). Ursachen des Studienabbruchs. Analyse 2002. Hannover: Hochschul-Informations-System.

[7] Freyer, K. (2013). Zum Einfluss von Studieneingangsvoraussetzungen auf den Studienerfolg Erstsemesterstudierender im Fach Chemie (Bd. 156). Berlin: Logos; Logos Berlin.

[8] Thiel, F., Veit, S. & Blüthmann, I. (2008). Ergebnisse der Befragung der Studierenden in den Bachelorstudiengängen an der Freien Universität Berlin Sommersemester (unveröffentlicht).

[9] Albrecht, A. (2011). Längsschnittstudie Identifikation von Risikofaktoren für einen erfolgreichen Studieneinstieg in das Fach Physik. Dissertation, Freie Universität Berlin.

[10] Blüthmann, I. (2012). Studierbarkeit, Studienzufriedenheit und Studienabbruch: Analysen von Bedingungsfaktoren in den Bachelorstudiengängen. Dissertation, Freie Universität Berlin.

[11] Kurz, G., Linser, M. & Oliveira-Vitt, L. de. (2008). Studienverlaufsuntersuchungen an der Hochschule Esslingen. Teil 1: Zulassungsverfahren und Eignungstests. In M. Rentschler (Hrsg.), Studieneignung und Studierendenauswahl. Untersuchungen und Erfahrungsberichte (Report - Beiträge zur Hochschuldidaktik, Bd. 42, S. 95–124). Aachen: Shaker.

[12] Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife (Beschluss der Kultusministerkonferenz vom 18.10.2012).

[13] RLP Phys. Sek II, Berlin 2006, Rahmenlehrplan Physik, Sekundarstufe II, Senatsverwaltung für Bildung, Jugend und Sport Berlin.

[14] RLP Math. Sek II, Berlin 2006 Rahmenlehrplan Mathematik, Sekundarstufe II, Senatsverwaltung für Bildung, Jugend und Sport Berlin.
[15] Boysen, G.; Glunde, H.; Heise, H.;(2000). Physik für Gymnasien, Cornelsen.
[16] Bigalke, A.; Köhler, N. (2007). Mathematik, Analysis. Band 1, Cornelsen.
[17] Bigalke, A.; Köhler, N. (2007). Mathematik Sekundarstufe II, Analytische Geometrie, Stochastik: Schülerbuch. Band 1, Cornelsen.
[18] Griesel, H.; Grundlach, A.; Postel, H. (2009). Elemente der Mathematik SII, Band 1 und 2; Schrödel.
[19] Moosbrugger, Kelava (Hrsg.) (2012), Testtheorie und Fragebogenkonstruktion, Springer Lehrbuch.
[20] Neumann (2014), Rasch-Analyse naturwissenschaftsbezogener Leistungstests. In Krüger, Parchmann, Schecker (Hrsg.), Methoden der naturwissenschaftsdidaktischen Forschung, Springer Spektrum.
[21] Rost, J. (2004). Lehrbuch Testtheorie-Testkonstruktion (2., vollst. Überarbeitete und erweiterte Aufl.). Bern: Hans Huber.
[22] Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. Psychologische Rundschau, 58(2), 103–117.
[23] Bond, T.; Fox, C. (2007). Applying the Rasch model: Fundermental measurement in the human sciences. Mahwah, NJ: Erlbaum, 2007.