Jurik Stiller[1]
Philipp Straube[2]
Sabrina Mathesius[1]
Stefan Hartmann[2]

[1] Humboldt-Universität zu Berlin
[2] Freie Universität Berlin

Ko-WADiS[1] | Vorläufige Ergebnisse der Pilotierung

Methodik

Nachdem die im vorhergehenden Beitrag beschriebenen Schritte zur Generierung von Testitems absolviert waren, wurden die 167 Items in Ko-WADiS einer ersten empirischen Überprüfung zugeführt.

Stichprobe

Es wurde insgesamt 578 Studierenden ein Testheft vorgelegt. 378 studieren an der Freien Universität und 104 an der Humboldt-Universität. Die übrigen 96 Studierenden studieren an verschiedenen Universitäten in Deutschland und Österreich. 84 % (484 Studierende) befinden sich im Bachelorstudium, 76 % im Lehramtsstudium. Der überwiegende Teil der Studierenden studiert im Erst- oder Zweitfach Biologie. Die Studierenden waren zum Testzeitpunkt im Schnitt 24,4 Jahre alt, 60 % der Studierenden sind weiblich.

Testadministration und Auswertungsmethoden

Für die Testadministration wurde das Multi-Matrix-Design verwendet. Die einzelnen Item-Blöcke wurden auf unterschiedliche Testhefte verteilt, die Testhefte jedoch über Ankerung immer systematisch in Beziehung gesetzt. Die Auswertung basierte auf Methoden der klassischen und probabilistischen Testtheorie. Ziel war die Generierung von Itemkennwerten, auf deren Basis eine Testrevision bzw. Itemselektion geschehen konnte.
Nach erfolgter Präpilotierung konnte die Itemselektion dann anhand des Anwahlverhaltens der Distraktoren sowie der Item-Fits, der gewichteten Abweichungsquadrate (wMNSQ) und der T-Werte durchgeführt werden. Für die Itemselektion wurde zudem die klassische Trennschärfe herangezogen. Auch inhaltliche Aspekte wurden natürlich während des gesamten Prozesses berücksichtigt. Der resultierende Itempool für die Pilotierung umfasste noch 113 Items, die sich relativ gleichmäßig auf die verschiedenen Zellen des Kompetenzmodells verteilen (siehe Abb. 1).

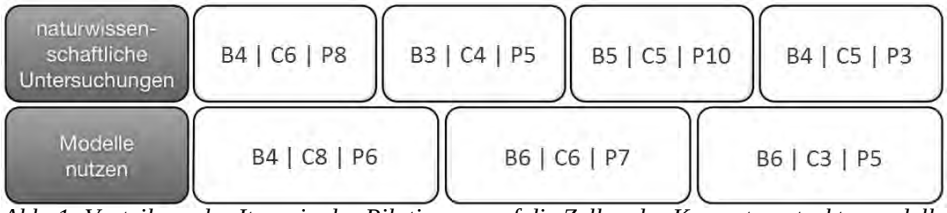

Abb. 1: Verteilung der Items in der Pilotierung auf die Zellen des Kompetenzstrukturmodells

Die EAP/PV-Reliabilität der IRT-Modellierung des Gesamtmodells mit den 113 Items der drei Fächer ist zum aktuellen Zeitpunkt niedrig (Biologie: $Rel_{EAP/PV}$ = .455; Chemie:

[1] Kompetenzmodellierung und -erfassung zum Wissenschaftsverständnis über naturwissenschaftliche Arbeits- und Denkweisen bei Studierenden (Lehramt) in den drei naturwissenschaftlichen Fächern Biologie, Chemie und Physik

$Rel_{EAP/PV}$ = .512; Physik: $Rel_{EAP/PV}$ = .471). Auf Basis dieser niedrigen Skalenhomogenität sind alle weiteren Analysen entsprechend mit Vorsicht zu interpretieren.

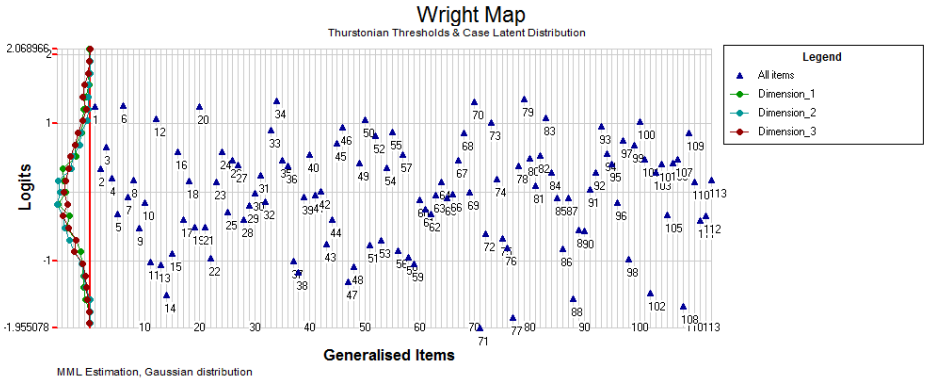

Abb. 2: Person-Item-Map
(Dimension 1: Biologie, Dimension 2: Chemie, Dimension 3: Physik)

Die Verteilung der Items über den gesamten Schwierigkeitsbereich ist gleichmäßig (Abb. 2). Einzelnen Fähigkeitsbereichen „entsprechen" nicht in allen Fächern ausreichend Items, eine ausreichend genaue Messung ist jedoch anhand der zu erwartenden noch ansteigenden Reliabilität (siehe Diskussion) wahrscheinlich sichergestellt.
Die Korrelationen zwischen den einzelnen Fächern sind hoch (Abb. 2; Cohen, 1988). Die Varianz besonders der Chemie ist im Vergleich zu den beiden anderen Fächern niedrig.

Tab. 1: Übersicht über die Korrelationen zwischen den Fächern und Varianz

	Biologie	Chemie	Physik
Biologie			
Chemie	0.898		
Physik	0.833	0.892	
Varianz	0.440 (0.026)	0.240 (0.014)	0.381 (0.023)

Zur Erhöhung der Genauigkeit der Messung wurde ein latentes Regressionsmodell mit Hintergrundvariablen spezifiziert. Als abhängige Variable wurde hier der logit-Wert festgesetzt. Als unabhängige (Hintergrund-)Variablen wurden unter anderem verschiedene demographische Daten (wie Alter, Geschlecht, Anzahl der studierten naturwissenschaftlichen Fächer, Universität) angewendet. Der Vergleich der Modellpassung des eindimensionalen Regressionsmodells mit der Modellpassung des dreidimensionalen Regressionsmodells (Dimensionen: Fächer) zeigt, im Gegensatz zur anhand der hohen Korrelationen zu vermutenden Struktur, eine deutlich bessere Passung des dreidimensionalen Modells auf die Daten (χ^2 (24, N = 562, p = .01). Die Modellparameter sind in Tabelle 2 aufgeführt.

Tab. 2: Übersicht über die Modellparameter (3dim: Fächer)

	Deviance	*AIC*	*BIC*
1dim	9290	13375	9316
3dim	9067	13384	9275

Diskussion

Der Kompetenztest zu fachmethodischen Kompetenzen repräsentiert ein objektives und ökonomisches Erhebungsinstrument. Insbesondere in Bezug auf die noch geringe Reliabilität gibt es aber noch Handlungs- und Verbesserungsbedarf. Adams (2005) beschreibt als möglichen Grund den *measurement design effect*. Dieser führt besonders bei zu geringer Item-Anzahl (pro Dimension), zu geringer Populationsgröße, zu homogener Stichprobe, einem heterogenem Konstrukt oder zu geringer Trennschärfe zu entsprechend reduzierter Reliabilität. Im Fall der beschriebenen Pilotierungsstudie ist denkbar, dass sich mehrere dieser Effekte auswirken. Die beiden Dimensionen *Modell nutzen* und *naturwissenschaftliche Untersuchungen durchführen* sind mit jeweils mindestens 14 Items pro Fach (Abb. 1) wahrscheinlich ausreichend abgedeckt. Die Populationsgröße jedoch ist mit 578 Studierenden sicherlich mit ausschlaggebend für die niedrige Reliabilität. Zusätzlich ist die Stichprobe insbesondere in der Chemie in Bezug auf die getesteten Kompetenzen sehr homogen (niedrige Varianz). Dies dürfte hier die Reliabilität zusätzlich reduzieren. Die Annahme eines heterogenen Konstrukts erscheint zudem plausibel, auch dies führt zu einer niedrigeren Reliabilität. Die Trennschärfe der einzelnen Items hingegen ist akzeptabel und somit als Ursache eher nicht anzunehmen.

Ausblick

Die Person-Item-Map zeigte, dass „unterversorgte“ Schwierigkeitsbereiche bzw. Bereiche der Personenfähigkeit, die bisher schlecht aufgelöst werden können, innerhalb des Kompetenztest existieren. Hier wird mit einer gezielten Nachkonstruktion nachgesteuert. Auch sind die Items insgesamt etwas zu leicht, was vor allem in Anbetracht des geplanten Längsschnittes ggf. problematisch sein könnte. Im Laufe des Studiums müssen ansteigende Fähigkeit erwarten werden, während die Pilotierungskohorten in großer Mehrzahl aus dem Bachelorstudium stammten.

Geplant ist aktuell auch eine Erhebung der kognitiven Grundfähigkeiten, des Wissens im Bereich *Natur der Naturwissenschaften* sowie der Fähigkeiten in einem standardisierten Test zum *analytischen Problemlösen.* Diese Daten sollen mit in das Hintergrundmodell einfließen.

Es steht zudem die querschnittliche Erhebung der Kompetenzen im Bereich Erkenntnisgewinnung im kommenden Semester an, das Projekt selbst hat aber auch die längsschnittliche Modellierung der Kompetenzen zum Ziel. Dies wird insbesondere im Rahmen der angestrebten zweiten Projektförderphase erreicht werden. Zeitnah lassen sich anhand der aktuellen und im kommenden Wintersemester 2013/2014 erhobenen Daten auch noch verschiedene andere Modelle spezifizieren, insbesondere Modelle, die die beiden Kompetenzfacetten unterscheiden, haben in ersten Modellierung eine gute Repräsentation der Daten gezeigt. Ob und inwiefern sich auch noch zellgenaue Modelle spezifizieren lassen, ist aktuell noch nicht endgültig abzusehen.

Literatur

Adams, R. (2005). Reliability as a measurement design effect. Studies in Educational Evaluation, 31,162-172.
Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.