Litzenberger, M., Punter, J. F., Gnambs, T., Jirasko, M. & Spiel, C. (in Druck). Qualitätssicherung bei der Studierendenauswahl mittels lernpsychologisch fundierter Wissensprüfung. In A. Kluge & K. Schüler (Hrsg.), *Qualitätssicherung und -entwicklung an Hochschulen: Methoden und Ergebnisse*. Lengerich: Pabst.

Qualitätssicherung bei der Studierendenauswahl mittels lernpsychologisch fundierter Wissensprüfung

Margarete Litzenberger, Joachim F. Punter, Timo Gnambs,
Marco Jirasko & Christiane Spiel

Zusammenfassung (Deutsch): An der Fakultät für Psychologie der Universität Wien wurde ein Auswahlverfahren für Psychologie-Studierende eingerichtet, welches die Absolvierung der Ringvorlesung „Psychologie als Wissenschaft“ umfasst. Als Prüfungsform wurde eine lehrzielorientierte und EDV-unterstützte Prüfung mit Multiple-Choice-Antwortformat eingeführt. Die Itemkonstruktion orientierte sich an mehreren Kriterien, die der Qualitätssicherung des Auswahlverfahrens dienen sollten. Zunächst wurden – in Anlehnung an die Taxonomie kognitiver Lernziele (Bloom, 1956) – Richtlinien zur Gestaltung von Multiple-Choice-Aufgaben erarbeitet. Darauf aufbauend erfolgte die Itementwicklung im Rahmen eines mehrstufigen Prozesses der Qualitätskontrolle. Dieses Procedere wird in diesem Beitrag ausführlich und beispielhaft dargestellt. Die Ergebnisse der Itemanalyse nach erfolgter Studierendenauswahl lassen auf eine zufriedenstellende Güte der Bewertung schließen.

Schlagwörter: Multiple-Choice, Studierendenauswahl, Auswahlverfahren

Abstract (Englisch): The Faculty of Psychology (University of Vienna) established a student selection procedure based on the lecture “Psychology as a Science”. The examination was designed as a multiple-choice exam, based on educational objectives, and can be analyzed computer-assisted. The item-construction follows defined criteria to guarantee the quality of the selection procedure. In the first step, scientific guidelines for the construction of fair multiple-choice-questions were elaborated, based on the taxonomy of cognitive educational objectives (Bloom, 1956). In the second step, the item-development was carried out in a stepwise process of quality control. This procedure is presented in detail. The results of the item analysis provide satisfactory psychometric properties.

Key Words: multiple-choice, student selection, proficiency assessment procedure

1. Ausgangslage

Im Zusammenhang mit dem Urteil des Gerichtshofes der Europäischen Gemeinschaften vom 07.07.2005 zu den „Voraussetzungen des Zugangs zum Hochschulstudium – Diskriminierung“ wurde an österreichischen Universitäten mit einem möglichen Ansturm ausländischer Studierender aus dem EU-Raum gerechnet. Diese Regelung schien insbesondere für Studierende aus Deutschland eine gangbare Alternative, den im Heimatland notwendigen *Numerus clausus* (vor allem in den Studienfächern Psychologie, Pharmazie und Medizin) zu umgehen. Eine Verordnung des Rektorats der Universität Wien[1] ermöglichte schließlich die Durchführung eines Auswahlverfahrens zur Auswahl der am geeignetsten erscheinenden BewerberInnen. Da der Numerus clausus aus Fairnessgründen nicht in Frage kam, jedoch hoher Zeitdruck bestand (Studienbeginn: 1. Oktober 2005), wurde als zukunftsorientierte Maßnahme eine in vorhergehenden Jahren bewährte Prüfung zu einer Ringvorlesung als Auswahlkriterium festgelegt.

Die Ringvorlesung „Psychologie als Wissenschaft“ ist eine einführende Lehrveranstaltung, die Grundwissen in den insgesamt zwölf Prüfungsfächern des Diplomstudiums Psychologie an der Universität vermittelt und in der ersten Semesterwoche jedes Wintersemesters in Form einer Blocklehrveranstaltung angeboten wird. Auf diese Weise können angehende Psychologie-Studierende bereits frühzeitig – im Sinne des Konzepts der *Berufswahlreife* (*vocational maturity*; vgl. Crites, 1974 bzw. Seifert, 1984) – umfassende Informationen über das Studienfach erwerben und so ihre Studienwahlentscheidung bestärken bzw. revidieren. Die positive Absolvierung der „Psychologie als Wissenschaft“ (zweiteilige Lehrveranstaltung zu je 1 SWS) ist laut Studienplan Voraussetzung für die Teilnahme an allen Lehrveranstaltungen mit immanentem Prüfungscharakter (Übungen, Proseminare). Seit dem Studienjahr 2005/06 ist nun an diese Prüfung zusätzlich das Auswahlverfahren gekoppelt.

Seit Einführung dieser Eingangslehrveranstaltung wurden jährlich etwa 1500 Prüfungen (mit freiem Antwortformat) pro Studienjahr abgenommen; die investierte Arbeitszeit für die Abwicklung betrug etwa 1000 Stunden und konnte nur mit Unterstützung aller FakultätsmitarbeiterInnen jeweils zeitgereicht bewerkstelligt werden. Zur Handhabung des (infolge des EuGH-Urteils) erwarteten hohen Ansturms von Studierenden aus dem EU-Raum und der enormen Prüfungsbelastung, wurde von der Fakultät für Psychologie die Prüfung zur Vorlesung „Psychologie als Wissenschaft“ vom freien Antwortformat auf das Multiple-Choice-Antwortformat umgestellt. Ziel dieser Umstellung war damit auch die Implementierung eines fairen Auswahlverfahrens durch die Konstruktion von eindeutigen Multiple-Choice-Aufgaben. Mit der Umsetzung wurde ein Projektteam beauftragt.

[1] Verordnung des Rektorats der Universität Wien vom 08.09.2005 (Mitteilungsblatt, Studienjahr 2004/2005, ausgegeben am 08.09.2005, 39. Stück)

2. Lernpsychologische und messtheoretische Grundlagen

Entsprechend der oben genannten Ziele und Vorgaben wurden zunächst, durch Aufarbeitung und Dokumentation der einschlägigen Literatur, Grundlagen für die Erstellung von computerisiert auswertbaren Multiple-Choice-Prüfungsfragen erarbeitet, die das reine „Wiedererkennen von Faktenwissen“ übersteigen. Besonderes Anliegen bei der Implementierung des Auswahlverfahrens war die lernpsychologische Fundierung der angestrebten Wissensprüfung. Technische und administrative Notwendigkeiten für die Durchführung derartiger Prüfungen wurden ermittelt und schließlich ein ausreichend umfangreicher Itempool für die Prüfung „Psychologie als Wissenschaft“ konstruiert. Zur Sicherung der hohen Qualitätsansprüche eines Auswahlverfahrens wurde ein mehrstufiges Kontrollregulativ zur Prüfung der formalen und inhaltlichen Qualität der Items implementiert. Diese Vorgehensweise zur Qualitätssicherung wird im Folgenden vorgestellt.

2.1. Taxonomien zur Generierung von Wissensfragen

Taxonomien, um die kognitive Fähigkeit zur Lösung eines Testitems zu spezifizieren, wurden verschiedentlich entwickelt. Die bekannteste stammt von Bloom (1956), eine weitere, welche der Itemkonstruktion zugrunde liegt, wurde von Roid und Haladyna (1982) veröffentlicht. Während die Bloomsche Taxonomie eher ein theoretisches Gerüst darstellt und zur Klassifikation von Items geeignet ist (Legg, 1991), bietet sich letztere Taxonomie speziell als Grundlage für den Prozess der Itemkonstruktion von Multiple-Choice-Tests an.

Bloom unterscheidet drei Bereiche (kognitiv, affektiv und psychomotorisch), denen er eine Reihe hierarchisch gegliederter Lernziele zuordnet. Leistungsbeurteilungen im Rahmen von Wissenstests, wie es die Eingangsprüfung zur Studierendenauswahl in der Psychologie darstellt, sind primär dem kognitiven Bereich zuzuordnen. Es handelt sich hierbei um Prozesse des Wissens, Verstehens, Begründens u. ä. Affektive und psychomotorische Lernziele finden deshalb im Folgenden keine Berücksichtigung (ohne deren Bedeutung, auch für die psychologische Ausbildung, schmälern zu wollen). Im Detail unterscheidet Bloom (1956) im kognitiven Bereich sechs Lernziele:

1. *Wissen (Knowledge)*: Informationen (von spezifischen Fakten bis zu ganzen Theorien) sollen aus dem Gedächtnis wiedergegeben werden, z. B. Reproduktion von Fachausdrücken, spezifische Fakten, Grundprinzipien.
2. *Verständnis (Comprehension)*: Die Bedeutung des Erlernten soll erfasst werden ohne es notwendigerweise auf andere (neue) Bereiche zu übertragen oder Implikationen daraus abzuleiten, z. B. Graphiken und Diagramme interpretieren, verbales Material in mathematische Formeln übertragen, Konzepte und Prinzipien verstehen.
3. *Anwendung (Application)*: Erlerntes (Regeln, Methoden, Konzepte, Theorien etc.) soll in neuen und konkreten Situationen angewendet werden, z. B. Konzepte in neuen Sachlagen anwenden, Gesetze und Theorien auf praktische Anwendungsfelder umlegen.
4. *Analyse (Analysis)*: Material soll in seine Bestandteile zerlegt werden, z. B. unausgesprochene Annahmen oder logische Fehlschlüsse erkennen, den Aufbau des Materials zerteilen können.

5. *Synthese (Synthesis)*: Aus Teilen, Elementen etc. soll etwas neues Ganzes zusammengestellt werden, z. B. Untersuchungsdesigns abwandeln, Forschungsabläufe in die richtige Reihenfolge bringen.
6. *Beurteilung (Evaluation)*: Die Güte eines vorgegebenen Materials soll für einen bestimmten Zweck eingeschätzt werden, z. B. Korrektheit von Interpretationen aufgrund der Datenlage beurteilen, Material anhand interner bzw. externer Standards beurteilen.

Die Bloomschen Lernziele sind nicht unabhängig von einander zu betrachten. Sie stehen vielmehr in einer geordneten Rangreihe zu einander. Wurde ein untergeordnetes Lernziel (wie z. B. Wissen) nicht erfolgreich erreicht, wird auch ein übergeordnetes Lernziel eher nicht zu bewältigen sein. Jemand der z. B. nicht weiß, was eine unabhängige Variable ist (*Wissen*), wird auch nicht in der Lage sein, diese in einem konkreten Beispiel zu identifizieren (*Anwendung*).

Ergänzend bzw. als Überbau der Testkonstruktion schlagen Roid und Haladyna (1982) eine kriteriumsbezogene Testentwicklung vor, wobei „kriteriumsbezogen“ so verstanden wird, dass die Items eines Tests eine repräsentative Auswahl der Lehrziele eines Fachs darstellen. Dementsprechend bedeutet ein Testscore von 85%, dass der Studierende 85% des Lehrstoffs beherrscht. Zur Entwicklung eines kriteriumsbezogenen Tests schlagen die Autoren einen fünfstufigen Prozess vor:

1. *Abklärung der Lehrziele*: Der Lehrende/Prüfende sollte eine klare Vorstellung über die Lehrziele haben, diese können eher abstrakt (z. B. „die Funktionsweise des menschlichen Gehirns verstehen“) oder auch sehr konkret (z. B. „den Grundaufbau einer Nervenzelle beschreiben können“) sein.
2. *Operationalisierung der Lehrziele*: Hier geht es um die Spezifikation der zuvor definierten Lehrziele und der Festlegung konkreter Regeln zur Itemkonstruktion mit dem Zweck, Lehrziele messbar zu machen.
3. *Itementwicklung*: Je nach Lehrziel und Schwierigkeitsniveau der Items können verschiedene Techniken zur Itemkonstruktion angewandt bzw. Typen von Multiple-Choice-Aufgaben (z. B. Mehrfachwahl-Aufgaben, Lückentext, Zuordnungsaufgaben) realisiert werden.
4. *Itemrevision*: In diesem Schritt sollen die Items einer logischen und einer empirischen Itemrevision unterzogen werden. Die generierten Items werden einerseits von ExpertInnen des Fachs hinsichtlich fachlicher Fehler bzw. der Konsistenz zwischen Item und Lehrziel geprüft und andererseits einer empirischen Kontrolle unterzogen um Mehrdeutigkeiten auszuschließen.
5. *Testentwicklung*: Nach der ersten Erprobung und Revision der Items liegt schließlich ein Itempool vor, aus dem zufällig eine Itemauswahl für eine kriteriumsbezogene Testerstellung im oben beschriebenen Sinn getroffen werden kann.

Die Itemkonstruktion im Rahmen der Implementierung der Multiple-Choice Prüfung zur „Psychologie als Wissenschaft“ orientierte sich vornehmlich an den beiden vorgestellten Taxonomien, wenngleich aus organisatorischen Gründen einige Abstriche gemacht werden mussten. Dennoch wurde die Testkonstruktion damit auf eine allgemein anerkannte und lernpsychologisch fundierte Basis gestellt.

2.2. Richtlinien zur Konstruktion von Multiple-Choice Aufgaben

Multiple-Choice stellt nicht ein spezifisches Antwortformat dar, sondern kann als Überbegriff verschiedener geschlossener Antwortformate verstanden werden. Grundsätzlich werden zwei Gruppen unterschieden: dichotome (*true/false* bzw. *alternate choice*) und polytome Antwortformate, bei denen eine (die beste) Antwort zu wählen ist. Ein wesentlicher Unterschied dieser beiden Ansätze besteht im Informationsgehalt der Antwortoptionen. Bei Richtig/Falsch-Items muss jede Antwortoption eindeutig und absolut richtig oder falsch sein, während bei Best-Answer-Items eine graduelle Abstufung der Korrektheit vorliegen kann.

Das dichotome Antwortformat wird sehr kontrovers diskutiert, wobei die Vorteile vor allem darin liegen, dass sie einfach zu konstruieren sind, mehr Inhalte in der gleichen Vorgabezeit als bei klassischem Multiple-Choice abgefragt werden können und auch höhere Lernziele erfassbar sind (vgl. Haladyna, 1992). Diese Vorteile sind jedoch durch eine Reihe von Studien (vgl. Downing, Baranowsk, Grosso & Norcini, 1995) in Frage gestellt. So können zwar mehr Items in diesem Format vorgegeben werden – Testscores von klassischen Multiple-Choice Items sind jedoch reliabler. Zudem zeigten sich Items als unterschiedlich schwierig, je nachdem ob der Stamm positiv oder negativ formuliert wurde und sie richtig bzw. falsch sind (Candida, Petersen & Petersen, 1976). Falsche Items weisen zudem eine größere Trennschärfe als Richtige auf (Ebel, 1972, zitiert nach Haladyna, 1992). Vergleichbare Kritik trifft multiple Richtig/Falsch-Items (d. h. ein Item besteht aus einem Aufgabenstamm und mehreren Antwortoptionen, die alle mit richtig/falsch zu beantworten sind). Während die Akzeptanz dieses Antwortformats bei einzelner Verrechnung der Antwortoptionen bei den Studierenden deutlich höher ist (Macher, 2005), dürften Ratewahrscheinlichkeiten und Antworttendenzen problematisch sein – empirische Studien hierzu fehlen allerdings bislang.

Best-Answer-Items sind in der Regel länger als klassische Multiple-Choice-Items, es muss daher mit längeren Bearbeitungszeiten gerechnet werden, wodurch in der gleichen Zeitspanne weniger Items vorgegeben und weniger Inhalte abgefragt werden können. Darüber hinaus kann hier Halbwissen (z. B. Ausschluss einzelner Antwortoptionen) helfen, die richtige Lösung zu finden, indem Distraktoren eliminiert werden können (d. h. das Format birgt Hinweise auf die Lösung für testerfahrene Studierende). Da die primären Anliegen der Itemkonstruktion eine automatisierte Auswertbarkeit und eine einfache Scorebildung (= Anzahl richtig bearbeiteter Items) waren, werden spezifische Typen von Best-Answer-Antwortformaten (z. B. Zuordnungsaufgaben, kontextbezogene Multiple-Choice-Aufgaben) hier nicht näher in Betracht gezogen.

Multiple-Choice mit Mehrfachwahl bestehen aus einem Aufgabenstamm und mehreren Antwortoptionen. Im Unterschied zu „klassischen“ Multiple-Choice-Items kann es dabei nicht nur eine, sondern mehrere richtige Antworten geben. Die Auswertung kann dabei einerseits in Anlehnung an multiple Richtig/Falsch-Items erfolgen (für jede richtigerweise ausgewählte bzw. nicht ausgewählte Antwortmöglichkeit wird ein Punkt vergeben) und andererseits nach dem Alles-oder-Nichts-Prinzip (nur wenn alle richtigen und keine falschen Antwortmöglichkeiten markiert werden, wird das Item als „gelöst“ gewertet und mit einem Punkt verrechnet). Letzterer Verrechnungsmodus erhöht die Schwierigkeit der Items deutlich, stößt aber bei Studierenden naturgemäß auf geringere Akzeptanz (vgl. Macher, 2005).

Um die Ratewahrscheinlichkeit zu minimieren und gleichzeitig die Möglichkeit auszuschließen, dass Halbwissen zur Lösung einzelner Items führen kann, wurde bei der Festlegung des Antwortformats der Prüfung zur Ringvorlesung für ein Multiple-Choice mit Mehrfachwahl bei Auswertung nach dem Alles-oder-Nichts-Prinzip entschieden. Die Anzahl richtiger Antwortmöglichkeiten pro Item wurde variiert, wobei es Teil der Aufgabe der Studierenden war, die Anzahl Richtiger je Item herauszufinden. Auf diese Weise wurde sichergestellt, dass nur die Beherrschung des Lehrstoffs zur erfolgreichen Absolvierung der Prüfung führt und Zufallsfaktoren weitestgehend ausgeschlossen werden können. Zur Vereinheitlichung und damit effizienteren Auswertbarkeit wurden fünf Antwortoptionen pro Item festgelegt, wobei entsprechend der Empfehlung von Haladyna, Downing und Rodriguez (2002) niemals keine oder alle Antwortmöglichkeiten richtig sein konnten.

Zur optimalen Anzahl der Distraktoren liegen eine Reihe (teils widersprüchlicher) Ergebnisse vor. Am konklusivsten erscheinen die Metaanalysen von Haladyna und Downing (1989) sowie Rodriguez (2005). Danach sinkt zwar die Itemschwierigkeit bei drei Antwortoptionen im Vergleich zu vier oder fünf Optionen leicht, aber Trennschärfe und Reliabilität steigen, während die Validität davon unberührt bleibt. Wie Haladyna und Downing (1993) im Rahmen einer umfangreichen Reanalyse verschiedener Multiple-Choice-Tests beobachten konnten, sind (unabhängig von der tatsächlich vorgegebenen Anzahl an Distraktoren) bei über zwei Drittel der Items lediglich ein oder zwei Distraktoren auch effektiv wirksam. Nur ein bis acht Prozent aller Items weisen drei wirksame Distraktoren auf. Kubinger, Holocher-Ertl und Frebort (2006, in Druck) konnten allerdings zeigen, dass Items mit weniger als k = 6 Antwortmöglichkeiten (mit lediglich einer richtigen Antwortoption) psychologisch-diagnostisch kaum vertretbar sind. Zusätzlich ergaben ihre Analysen, dass das Antwortformat 2 (aus 5) richtige Antwortoptionen zumindest nicht leichter ist als das freie Antwortformat. Da es aufgrund des Zeitdrucks keine Möglichkeit gab die Güte der Distraktoren vorab zu ermitteln und wir nicht Gefahr laufen wollten, zu einfache Distraktoren anzubieten, wurde demgemäß für ein fünfstufiges Antwortformat entschieden.

Nach Festlegung von Kriterien zur Gestaltung der Multiple-Choice Aufgaben wurden ausgehend von einer umfangreichen Literaturrecherche (vgl. Haladyna & Downing, 1989; Moreno, Martínez & Muñiz, 2006) Richtlinien zur konkreten Itemgestaltung festgelegt. Diese umfassen inhaltliche Empfehlungen (z. B. *Jedes Item muss sich an konkreten inhaltlichen Lehrzielen des Fachgebiets orientieren*), formale Gestaltungshinweise (z. B. *Die Items sollen so kurz wie möglich, aber so lang wie für das Verständnis notwendig sein*) und Formulierungshilfen (z. B. *Die Items sollten positiv formuliert sein, Negationen sind nur in Ausnahmefällen zulässig*). Die Richtlinien sollten den ItemkonstrukteurInnen als wissenschaftliche Grundlage und zur Orientierung an einheitlichen Standards dienen.

3. Maßnahmen zur Qualitätssicherung

Zur Gewährleistung der hohen Qualitätsansprüche eines Auswahlverfahrens wurde ein mehrstufiges Kontrollregulativ zur Prüfung der formalen und inhaltlichen Qualität der Items implementiert. Die Vorgehensweise zur Qualitätssicherung wird im Folgenden vorgestellt.

3.1. Mehrstufiger Prozess der Itemkonstruktion

In Anlehnung an Roid und Haladyna (1982) erfolgte die Entwicklung der Items in einem mehrstufigen Prozess. In einem ersten Schritt wurden alle Vortragenden der Ringvorlesung „Psychologie als Wissenschaft“ gebeten ihre implizit vorhandenen Lehrziele explizit zu formulieren. Parallel dazu wurden vom Projektteam die oben dargestellten Richtlinien für die Itemkonstruktion formuliert. Die konkrete Itementwicklung erfolgte durch eingeschulte ExpertInnen aus den 12 verschiedenen Prüfungsfächern der Ringvorlesung. Diese erste Konstruktionsphase wurde durch das Projektteam supervidiert. In der zweiten Konstruktionsphase wurden alle Fragen vom Projektteam hinsichtlich der Erfüllung der formaler Gestaltungskriterien geprüft und bei Bedarf entsprechend modifiziert. Im nächsten Schritt wurden die Vortragenden der Ringvorlesung gebeten, die Fragen bezogen auf inhaltliche und sprachliche Eindeutigkeit durchzusehen und gegebenenfalls zu korrigieren.

Zur Verwaltung des insgesamt 800-1000 Fragen umfassenden Fragenpools wurde eine Datenbank eingerichtet, die die Auswahl von Fragen nach diversen Kriterien (inhaltliches Lehrziel, Lernziel nach Bloom, Prüfungsfach) erlaubt. Items, die aufgrund ihrer Formulierung einen Hinweis zur Lösung eines anderen Items liefern könnten, wurden in der Datenbank explizit ausgewiesen, sodass diese bei der computerbasierten Itemauswahl nicht in einer Prüfung gemeinsam vorgegeben werden.

Die Erstellung der Prüfungsbögen kann automatisiert über die Datenbank erfolgen, sodass auch eine Zufallsauswahl der Items möglich ist und relativ einfach mehrere Parallelformen mit variierter Itemreihenfolge und Permutation der Antwortoptionen (ausgenommen bei Items deren Antwortoptionen in einer logischen Abfolge angeordnet sind) erstellt werden können. Die Zufallsauswahl der Items ist jedoch an bestimmte Kriterien geknüpft, die eine Ungleichverteilung verhindern soll (z. B. jeweils 7 Items pro Prüfungsfach, Verhältnis reiner Wissensfragen zu Verständnis- und Anwendungsfragen entspricht 4:3, jedes inhaltliche Lehrziel wird nur durch ein Item abgedeckt).

In Zusammenarbeit mit dem Zentralen Informatikdienst der Universität Wien wurde ein Hochgeschwindigkeitsscanner samt entsprechender Software angeschafft, der eine äußerst effiziente Auswertung der Prüfungsarbeiten ermöglicht. Einzige Schwierigkeit dabei war die Einstellung der Schwellenwerte, d. h. ab welchem Prozentsatz ein Antwortkästchen als leer, nur verschmutzt, angekreuzt bzw. vollständig ausgestrichen interpretiert wird. Nach mehreren Testläufen und Verbesserung der Instruktion konnte die Fehlerrate auf ein Minimum gesenkt werden, so dass nur Extremfälle händisch nachgebessert werden mussten. Die weitere Auswertung erfolgt über eine entsprechende SPSS-Syntax-Datei, welche die Dateien mit den Lösungsvektoren (von der Datenbank automatisch erzeugt), sowie die von der Software des Belegscanners erzeugte Textdatei mit den Antwortvektoren einliest und schließlich die Punkte wie auch die Noten automatisiert berechnet.

Zur Bewältigung der zahlenmäßig umfangreichen Prüfungseinsicht wurde unter den Gesichtspunkten der schnellen Umsetzbarkeit und weitgehenden Transparenz eine Online-Prüfungseinsicht realisiert. Dabei werden den Studierenden die eigenen Antwortbelege nach dem Scan durch den Belegleser zur Verfügung gestellt sowie ein Überblick der automatisch eingelesenen Antworten und die Lösungen rückgemeldet.

3.2. Itemanalyse

Die Prüfungen zu „Psychologie als Wissenschaft I“ (1080 Prüfungsantritte im Studienjahr 2005/06) und „Psychologie als Wissenschaft II“ (942 Prüfungsantritte im Studienjahr 2005/06) haben zu verschiedenen Terminen stattgefunden. Es wurden jeweils 42 Items in drei Parallelgruppen vorgegeben. Mit einem Cronbach-Alpha von .85 bzw. .88 weist die interne Konsistenz beider Prüfungsteile eine als gut zu beurteilende Reliabilität auf.

Zur Identifikation von Items, die unterschiedliche Lösungswahrscheinlichkeiten für verschiedene Gruppen aufweisen, wurde auf Analysen der Odds Ratio nach Mantel-Haenszel (vgl. Dorans & Holland, 1993) zurückgegriffen. Die Prüfungsfragen können bezüglich der Testgruppen (variierte Darbietungsreihenfolge der Items) großteils als fair angesehen werden. Von den 84 vorgegebenen Items mussten lediglich drei als stark und weitere sieben als leicht biased (nach Dorans & Hollands, 1993) eingestuft werden (d. h. die Zugehörigkeit zu einer Parallelgruppe begünstigte deren Lösung).

Auch bezüglich des Geschlechts wurden nur wenige auffällige Items identifizieren. So finden sich nur drei Items mit starkem Bias, die primär Männer begünstigten. Inhaltlich weisen zwei dieser Items einen stark technischen Schwerpunkt auf. Einmal sollte das Gedächtnis in Form einer Analogie mit Komponenten eines Computers verglichen werden und ein anderes Mal waren (u. a. technische) Probleme bei der Ableitung von DC-Potentialen zu identifizieren.

Eine weitere Analyse sollte prüfen, ob Items, die höhere Lernziele entsprechend der Bloomschen Lehrzieltaxonomie (1956) prüfen, schwieriger sind als reine Wissensitems. Der Vergleich der Itemschwierigkeiten ließ mit MW_{Wissen}=.53 bzw. $MW_{Höhere}$=.46 zunächst nur marginale Unterschiede hinsichtlich der Itemgruppen (U=665, p=.07) erkennen (vgl. Tabelle 1). Entgegen der Hypothese scheinen Wissensitems schwieriger zu sein als Items, die höhere Lernziele verfolgen.

Tabelle 1: Mittlere Itemschwierigkeiten nach Lernziel und Fachbereich

	Wissen	Höhere Lernziele
Allgemeine Psychologie	.71	.46
Methodenlehre	.35	.63
Entwicklungspsychologie	.65	.43
Differentielle Psychologie	.48	.58
Biologische Psychologie	.41	.35
Sozialpsychologie	.78	.65
Einführung in die Psychologie	.45	.34
Psychologische Diagnostik	.34	.45
Klinische Psychologie	.61	.34
Wirtschaftspsychologie	.52	.42
Bildungspsychologie	.60	.38
Evaluation	.46	.33
Gesamt	.53	.46

Berücksicht man darüber hinaus die jeweiligen Fachbereiche ergibt sich eine interessante Wechselwirkung (vgl. Tabelle 2). Während bei den meisten Prüfungsfächern wie erwartet Items, die reines Faktenwissen prüfen, leichter sind als Items höherer Lernziele (MW_{Wissen}=.58, $MW_{Höhere}$=.42), verhält es sich bei statistisch-methodische Prüfungsfächern (Methodenlehre, Differentielle Psychologie und Psychologische Diagnostik) genau umgekehrt (MW_{Wissen}=.39, $MW_{Höhere}$=.55). In diesen Fächern sind anwendungsorientierte Fragestellungen offensichtlich einfacher zu beantworten als reine Wis-

sensfragen. Die Resultate einer parameterfreien Rang-Varianzanalyse bestätigen dieses Ergebnis (vgl. Tabelle 2 und Abbildung 1).

Tabelle 2: Rang-Varianzanalyse nach Kubinger (1986) der Itemschwierigkeiten nach Fachgebiet und Lernziel

	Chi2	df	p
Fachbereich	21.18	11	.03
Lernziel	4.94	1	.02
Fachbereich * Lernziel	21.76	11	.03

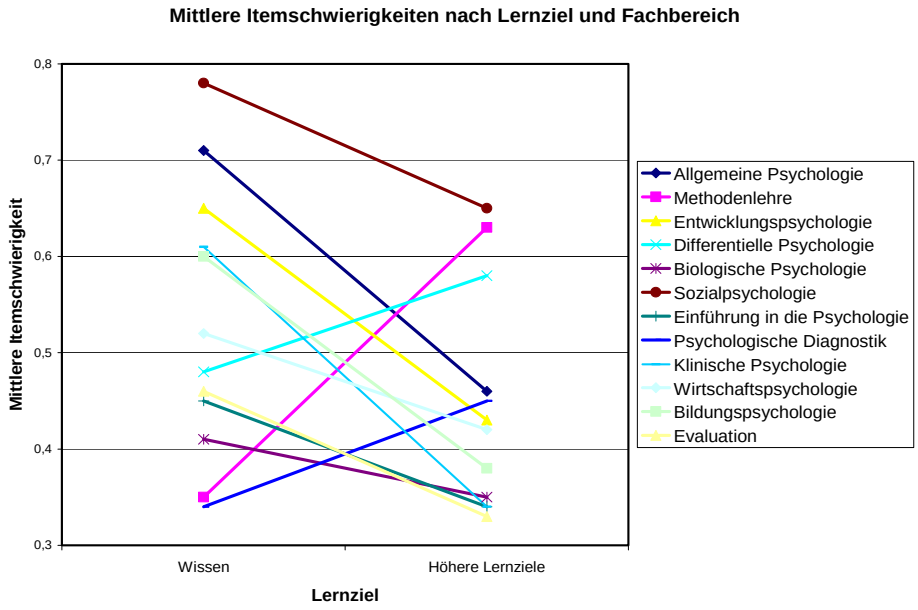

Abbildung 1: Mittlere Itemschwierigkeiten nach Lernzielen und Fachbereich

Um Aufschluss über die Qualität der einzelnen Items bzw. der Distraktoren – besonders in den statistisch-methodischen Prüfungsfächern – zu erhalten, wurden die relativen Häufigkeiten der gewählten Antwortmöglichkeiten pro Item getrennt für drei verschiedene Leistungsgruppen berechnet. Die Bildung der Leistungsgruppen (hoch, mittel, schwach) erfolgte in Abhängigkeit des Gesamtscore pro Prüfung, wobei in Anlehnung an die Empfehlung von Kelley (1939, zitiert nach Ebel, 1954) die beiden Extremgruppen einen Stichprobenumfang von jeweils 27% der Gesamtstichprobe umfassten. Diese Analysemethode ermöglicht die Identifikation der Güte der Antwortmöglichkeiten. Wenn beispielsweise richtige Antwortoptionen von sehr leistungsfähigen Studierenden (Personen, die insgesamt einen sehr hohen Score erreicht haben) weniger häufig gewählt werden als von leistungsschwachen Studierenden, so ergibt sich daraus ein Hinweis auf eine mögliche Uneindeutigkeit der Antwortoption.

Ein Beispiel (Psychologische Diagnostik)**:**
Auf welche(n) Aspekt(e) der Intelligenz zielt der Begriff der Intelligenz bei Binet ab?
a) Problemlösefähigkeit des täglichen kindlichen Lebens. (*richtig*)
b) Erfassung komplexer geistiger Prozesse. (*richtig*)
c) Quotient zwischen Intelligenz- und Lebensalter.
d) "Primary mental abilities".
e) Erhebung verschiedener, voneinander unabhängiger Fähigkeiten.

Tabelle 3: Relative Häufigkeiten der gewählten Antwortmöglichkeiten für drei Leistungsgruppen

		Item gesamt	a	b	c	d	e
Leistungsgruppe	3 hoch	,3882	,9882	,5922	,3725	,0118	,0510
	2 mittel	,1319	,8310	,3356	,5810	,1435	,1644
	1 niedrig	,0314	,5765	,2706	,6392	,2902	,2431
	Gesamt	,1741	,8047	,3875	,5403	,1476	,1550

Bei dem dargestellten Item handelt es sich um ein als problematisch identifizierterbares Wissensitem: Die relativen Häufigkeiten (vgl. Tabelle 3) weisen auf eine hohe Schwierigkeit hin, die vermutlich auf Distraktor „c“ zurückzuführen ist, der relativ oft – auch von eigentlich leistungsfähigen Studierenden – fälschlich als richtig angenommen wird. Eine mögliche Erklärung könnte sein, dass die gegebene Definition der richtigen Definition („Differenz zwischen Intelligenz- und Lebensalter“) zu nahe kommt und daher sehr leicht Verwechslungen möglich sind. Diese suboptimale Antwortmöglichkeit wurde daher umformuliert (neu: „Intelligenz ist, was der Intelligenztest misst“) und bei einem späteren Prüfungstermin erneut vorgegeben, um so die Verbesserung auch empirisch zu prüfen. Dabei zeigte sich, dass die zuvor kritische Antwortmöglichkeit nun für Leistungsfähige mehr oder weniger auszuschließen ist, für leistungsschwächere Studierende hingegen noch immer als mögliche Lösung in Frage kommt (vgl. Tabelle 4).

Tabelle 4: Relative Häufigkeiten der gewählten Antwortmöglichkeiten für drei Leistungsgruppen nach Korrektur des Items

		Item gesamt	a	b	c	d	e
Leistungsgruppe	3 hoch	,8595	,9669	,7190	,0579	,0248	,0579
	2 mittel	,6557	,9290	,4918	,1421	,0929	,1913
	1 niedrig	,3070	,6667	,3333	,1667	,3070	,2281
	Gesamt	,6196	,8684	,5144	,1244	,1316	,1627

4. Diskussion und Ausblick

Der vorliegende Beitrag dokumentiert den aufwendigen Prozess der Umstellung einer Prüfung auf Multiple-Choice-Antwortformat und veranschaulicht die Praktikabilität herkömmlicher Taxonomien bei der Itemkonstruktion. Aufgrund der Brisanz des Themas „Studierendenauswahl“, welches in Österreich vor dem Hintergrund des „freien Hochschulzugangs“ durchaus noch kritisch gesehen wird, wurde auf Qualitätssicherung besonderer Wert gelegt. Die Orientierung an Standards der Test- und Itementwicklung gewährleistete schließlich die Implementierung einer fundierten Wissensprüfung, die sowohl von Seiten der Studierenden weitgehend beschwerdefrei blieb, als auch den Medien nicht angreifbar erschien.

Aus der Darstellung der Itemanalyse wird aber auch klar, dass selbst aufwendig konstruierte Items Gefahr laufen der empirischen Prüfung nicht stand zu halten. So konnten aus der Itemanalyse noch zahlreiche Hinweise für die sukzessive Verbesserung des Itemmaterials gewonnen werden. Dabei hat sich auch gezeigt, dass die sechs-stufige Klassifikation nach Bloom (1956) nicht praktikabel ist, da die Grenzen zwischen den verschiedenen „höheren" Lernzielen (z. B. zwischen *Analyse* und *Evaluation*) eher fließend sind und im Einzelfall die Klassifikation schwer fällt. Als gangbare Methode hat sich in der Praxis vielmehr die Unterscheidung zwischen reinen „Wissensitems" und „Items höherer Lernziele" herausgestellt. Auch der hierarchische Ansatz von Bloom konnte trotz des mehrstufigen Kontrollregulativs im Rahmen der Itemkonstruktion anhand des Itemmaterials nicht belegt werden. Wissensitems haben sich gegenüber Items höherer Lernziele nicht durchgängig als einfacher herausgestellt. Allerdings könnte der „Methodenanteil" der Prüfungsfächer eine Moderatorvariable sein, die jedoch sowohl theoretisch als auch empirisch zu prüfen sein wird.

Kontakt:

Univ.-Ass. Mag. Dr. Margarete Litzenberger
Arbeitsbereich Psychologische Diagnostik
Institut für Entwicklungspsychologie und Psychologische Diagnostik
Fakultät für Psychologie
Universität Wien
Liebiggasse 5
A-1010 Wien
Tel.: 0043/1-4277-47855
Fax.:0043/1-4277-47905
e-mail: margarete.litzenberger@univie.ac.at
Homepage: http://www.univie.ac.at/Psychologie/diagnostik

Literatur

Bloom, B. S. (1956). *Taxonomy of educational objectives*. New York: McKay.

Candida, C., Petersen, C. C. & Petersen, J. L. (1976). Linguistic determinants of the difficulty of true-false items. *Educational and Psychological Measurement, 36*, 161-164.

Crites, J. O. (1974). Career development processes. A model of vocational maturity. In E. L. Herr (Ed.), *Vocational guidance and human development* (pp. 296-320). Boston: Houghton Mifflin.

Dorans, N. J. & Holland, P. W. (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In P. W. Holland & H. Wainer. (Eds.), *Differential Item Functioning*. (pp. 35-66). Hillsdale: Erlbaum.

Downing, S. M., Baranowsk, R. A., Grosso, L. J. & Norcini, J. R. (1995). Item type and cognitive ability measured: The validity evidence for multiple true-false items in medical speciality certifications. *Applied measurement in Education, 8*, 189-199.

Ebel, (1954). Procedures for the analysis of classroom tests. *Educational and Psychological Measurement, 14*, 352-364.

Haladyna, T. M. (1992). The Effectiveness of Several Multiple-Choice Formats. *Applied Measurement in Education, 5*, 73-88.

Haladyna, T. M. & Downing, S. M. (1989). Validity of a Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement in Education, 2*, 51-78.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice item? *Educational and Psychological Measurement, 53*, 999–1010.

Haladyna, T. M.; Downing, S. M. & Rodriguez, M.C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education, 15*, 309-334.

Kubinger, K. (1986). A Note on Non-Parametric Tests for the Interaction in Two-Way Layouts. *Biometrical Journal, 28*, 67-72.

Kubinger, K.D., Holocher-Ertl, S. & Frebort, M. (2006, in Druck). Zur testtheoretischen Qualität von Multiple-Choice-Items: 2 aus 5 vs. 1 aus 6 richtige Antwortmöglichkeiten. In B. Gula, R. Alexandrowicz, S. Strauß, E. Brunner, B. Jenull-Schiefer, & O. Vitouch (Hrsg.), *Proceedings der 7. Tagung der Österreichischen Gesellschaft für Psychologie in Klagenfurt* (S. 119-124). Lengerich: Pabst.

Legg, S. M. (1991). *Handbook on Testing and Grading*. Gainsville: University of Florida. Online im Internet: http://www.at.ufl.edu/testing/handbook.pdf (2005-07-27).

Macher, S. (2005). *Standardisierte Prüfungsmethoden in der medizinischen Ausbildung – Handbuch zur Konstruktion von Prüfungsaufgaben.* Graz: Medizinische Universität. Online im Internet: http://thor.meduni-graz.at/qm/documents/QM_SM_HandbuchPruefungsmethoden _20050404_01.pdf (2006-08-08).

Moreno, R., Martínez, R. J. & Muñiz, J. (2006). New Guidelines for Developing Multiple-Choice Items. *Methodology, 2*, 65-72.

Rodriguez, M. C. (2005). Multiple-Choice Items: A Meta-Anaysis of 80 Years of Research. *Educational Measurement: Issues and Practices, 24*, 3-13.

Roid, G. H. & Haladyna, T. M. (1982). *A technology for test-item writing*. New York: Academic Press.

Seifert, K. H. (1984). Berufswahlreife. In Bundesanstalt für Arbeit (Hrsg.), *Handbuch der Berufswahlvorbereitung* (S. 186–197). Mannheim: TransMedia.