Wie lässt sich der Lernerfolg von Schülerinnen/Schülern im Fach Rechnungswesen messen?

Entwicklung eines Messinstruments

MAG. CHRISTOPH HELM
Universitätsassistent
Johannes Kepler Universität Linz
christoph.helm@jku.at

MAG. BARBARA WIMMER
Senior Lecturer an der Abteilung für Pädagogik und Pädagogische Psychologie
Johannes Kepler Universität Linz
barbara.wimmer@jku.at

Abstract

Im Rahmen einer von den Autor/innen durchgeführten längsschnittlichen Evaluationsstudie zum COoperativen Offenen Lernen (COOL) soll unter anderem der fachliche Lernerfolg der Schüler/innen erfasst werden. Da für die untersuchten Schularten (kaufmännische Schulen, Schulen für wirtschaftliche Berufe) die kaufmännische Kompetenz zentral ist, für diese aber kaum Erhebungsinstrumente vorliegen, wurde im Rahmen eines Pretests ein selbsterstelltes Inventar aus typischen Schulbuchaufgaben des 1. Jahrgangs aus dem Fach Rechnungswesen an Zweitklässler/innen der genannten Schularten erprobt. Der Beitrag stellt erste Ergebnisse dieses Versuchs dar. In den ersten beiden Abschnitten wird auf die Notwendigkeit der Erstellung eines solchen Messinstrumentes, den theoretischen Hintergrund und die „Felderprobung" verwiesen. In Abschnitt drei wird auf psychometrische Qualitätsmerkmale von Tests eingegangen, die im vierten Abschnitt für das entwickelte Messinstrument statistisch überprüft werden. Gleichzeitig wird versucht zu zeigen, inwiefern es gelang, Kompetenzdimensionen abzubilden, die für die Lösung von Aufgaben aus dem Fach Rechnungswesen notwendig erscheinen. Zusammenfassende Erläuterungen zur Güte des Messinstruments mit Schwerpunkt auf die Validität bilden den Abschluss des Beitrags.

Ziel und Hintergrund des Beitrags

Die Abteilung für Pädagogik und Pädagogische Psychologie der JKU Linz führt in Kooperation mit dem COOL-Impulszentrum Steyr ein mehrjähriges Forschungsprojekt durch, das die Analyse der Lernprozesse von Schüler/inne/n kaufmännischer Schulen und Schulen für wirtschaftliche Berufe in kooperativen offenen und traditionellen Unterrichtssettings (LOTUS) zum Gegenstand hat. Die Analyse des fachlichen Lernerfolgs der Schüler/innen beschränkt sich auf die kaufmännische Kompetenz, die für die untersuchten Schularten (kaufmännische Schulen, Schulen für wirtschaftliche Berufe) als zentrales Ausbildungsziel angesehen werden kann, da diese Schularten auf die Ausübung wirtschaftlicher Berufe vorbereiten sollen. Als wichtige wirtschaftliche Fächer werden Rechnungswesen sowie Betriebs- und Volkswirtschaft in den Lehrplänen der oben angeführten Schularten genannt (vgl. BMUKK 2003; BMUKK 2004).
Zur Erhebung entsprechender Kompetenzen gibt es aber keine psychometrisch abgesicherten und für die Stichprobe geeigneten Messinstrumente. Obwohl es im deutschsprachigen wie im angloamerikanischen Raum immer wieder Bemühungen gab und gibt, kaufmännisches Wissen durch psychologisch fundierte Messinstrumente abzubilden (bspw. LEHMANN & SEEBER 2007; WINTHER 2010; BECK & KRUMM 1998; SAGEDER 2003; SCHUMANN et al. 2010; BOTHE et al. 2005), eignen sich die existierenden Messinstrumente für die oben angeführte Längsschnittuntersuchung aus folgenden Gründen nicht:

1. Einige Tests beziehen sich auf das Curriculum für kaufmännisch Auszubildende in der Bundesrepublik Deutschland und decken daher zum Teil andere Inhalte als jene, die in den Curricula der hier untersuchten Schulen vorgesehen sind, ab. Derartige Tests lassen sich im Projekt „ULME III" (LEHMANN & SEEBER 2007) sowie im „ALUSIM" (WINTHER 2010) finden.
2. Tests wie der „WBT" (BECK & KRUMM 1998) oder der „OEKOMA" (SCHUMANN EBERLE, OEPKE, PFLÜGER, GRUBER, STAMM & PEZZOTTA 2010) bilden volks- und betriebswirtschaftliches Allgemeinwissen ab. Zwar sind die dort gemessenen Konstrukte berufs(bildungs)bezogen nicht irrelevant, jedoch erscheinen auch sie zur Erfassung des spezifischen Unterrichtserfolg in den untersuchten Schulen ungeeignet, da die Instrumente nicht aus dem Curriculum abgeleitet wurden (vgl. SCHUMANN et al. 2010, S. 3).
3. Der Betriebswirtschaftswissenstest von SAGEDER (2003) ist zwar curricular abgeleitet, jedoch aus allen drei ersten Jahrgängen der betreffenden Schularten, was dazu führt, dass bestimmte curriculare Teilbereiche einzelner Jahrgänge unterrepräsentiert sind.
4. Manche Tests wurden für Studierende (z. B. „BAKT", BOTHE, WILHELM & BECK 2005) entwickelt, weshalb sie sich für den Einsatz auf Sekundarstufe II wenig eignen.
4. Hinzu kommt, dass einige der vorliegenden Skalen noch im Entwicklungsstadium sind (z. B. „WIWIKOM", ZLATKIN-TROITSCHANSKAIA & KUHN 2010).

Aus diesen Gründen erscheint es notwendig, ein Messinstrument zu entwickeln, das den Anforderungen des Forschungsprojekts bzw. der österreichischen Schulpraxis gerecht wird.

Ziel des vorliegenden Beitrags ist es daher, ein Verfahren zu entwickeln (im Folgenden mit WBB „Wissensüberprüfung von Basiskenntnissen der Buchhaltung" bezeichnet), das den Lernerfolg von Schüler/inne/n im Fach Rechnungswesen des 1. Jahrgangs erhebt. Im Zentrum steht eine statistische Analyse der Items des WBB vor dem Hintergrund der probabilistischen Testtheorie. Sie soll zeigen, dass die Qualitätsmerkmale psychometrischer Tests erfüllt werden und der Test sich daher zum Einsatz im angeführten Forschungsprojekt eignet.

Testzusammenstellung und -durchführung

Definition des Konstrukts und Auswahl der Testitems

Im LOTUS-Projekt wird der Frage nachgegangen, inwiefern das Lehrer/innen/handeln in selbstgesteuerten Lernphasen den Lernerfolg der Schüler/innen beeinflusst bzw. in Zusammenhang mit dem Unter-

richtserfolg steht. Im Fach Rechnungswesen ist der Unterricht traditionell stark an Schulbüchern orientiert, weshalb sich, nach Ansicht des Autors und der Autorin, der Unterrichts- bzw. Lernerfolg vor allem im Ausmaß der Bewältigung schulbuchtypischer Aufgabenstellungen zeigt. Vor diesem Hintergrund handelt es sich beim WBB um eine anforderungsorientierte Testkonstruktion (auch Prototypenansatz, vgl. BÜHNER 2008, S. 46ff.). Es wurde eine Analyse zweier Schulbücher (Trauner und Manz Verlag) durchgeführt: Für nahezu alle Themenbereiche, die das unten beschriebene Konstrukt betreffen, wurden repräsentative Aufgaben ausgewählt und fünf Schwierigkeitsstufen (Einschätzung durch die Autor/innen) zugewiesen. Aus diesen Kategorien wurden dann Items für die hier vorgestellte Version des WBB ausgewählt. Somit stellt der WBB einerseits eine *rationale Testentwicklung* dar (vgl. BÜHNER 2008, S. 47), da sie auf rein inhaltlichen Gesichtspunkten basiert, nämlich den Schulbuchinhalten. Andererseits wurden nur jene Schulbuchaufgaben ausgewählt, deren Bewältigung eine bestimmte Kompetenz erfordert. Diese Kompetenz kann als ein latentes Konstrukt beschrieben werden, *das die Fähigkeit darstellt, das System der Doppelten Buchhaltung (GROHMANN-STEIGER, SCHNEIDER, & EBERHARTINGER 2008, S. 52f.) anzuwenden, um bspw. Geschäftsfälle korrekt zu verbuchen und um deren Auswirkungen auf den Unternehmenserfolg zu beschreiben. Mit anderen Worten: Der Test misst die Fähigkeit, die Systematik der Verbuchung im Hauptbuch anzuwenden (vgl. BMUKK 2004).* Diese Arbeitsdefinition stellt die Basis für die deduktive, theorieorientierte Auswahl der Testitems dar. Des Weiteren ist anzumerken, dass die vorliegenden Analysen keine spezifische Kompetenzstruktur etwa im Sinne von Kompetenzstufen zu prüfen versucht. Dazu fehlen detaillierte Überlegungen, die in weiteren Forschungsarbeiten durchzuführen sind. Die oben angeführte Arbeitsdefinition kann jedoch als Komponentenstruktur eines zu prüfenden „Kompetenzmodells" angesehen werden.

Pilottestung

Der Einsatz des WBB erfolgte in den Monaten Februar und März 2012, weshalb eine Testung von Schüler/inne/n des 1. Jahrgangs nicht zielführend gewesen wäre, da wichtige Stoffgebiete im Unterricht noch nicht behandelt wurden. 424 Schüler/innen überwiegend des 2. Jahrgangs bilden die Stichprobe (zwei Klassen befinden sich bereits im 3. Jahrgang):

Stichprobe (n = 424 Schüler/innen, 21 Klassen)		
Schulart	76 % Schulen f. wirt. Berufe	24 % kaufmännische Schulen
Geschlecht	79 % weiblich	21 % männlich
zuhause gesprochene Sprache	89 % Deutsch	11 % andere Sprache
Schulstandorte: Biedermannsdorf, Bregenz, Graz, Haag, Linz, Saalfelden, Wien		

Tabelle 1: Beschreibung der Stichprobe

Die Bearbeitungszeit des WBB wurde mit einer Unterrichtseinheit limitiert. Bei Testdurchführung konnte kein Zeitdruck für die Schüler/innen erkannt werden. Zur Lösung der Testaufgaben wurde den Schüler/inne/n Kontenplan und Taschenrechner zur Verfügung gestellt. Dem WBB wurde ein kurzer Fragebogen vorangestellt, der zusätzliche Informationen, wie die Zeugnis- und Schularbeitsnoten in den Fächern Rechnungswesen, Mathematik und Deutsch, biografische Daten, den Beruf der Eltern und Informationen zu einem etwaigen Migrationshintergrund erfragte. Im Schlussteil des Beitrags erfolgt eine Auswertung dieser Zusatzvariablen. Zuvor wird auf die Testgüte des WBB eingegangen. Dazu werden im ersten Schritt Anforderungen an psychometrische Tests allgemein erläutert. Im zweiten Schritt wird mittels Analysen probabilistischer Testtheorie danach gefragt, ob der WBB diese Anforderungen erfüllt.

Anforderungen an Tests

Damit Messungen von Persönlichkeitseigenschaften und Fähigkeiten aussagekräftig sind, müssen sie bestimmten Anforderungen genügen (vgl. BÜHNER 2008, S. 301ff.):

1. Messungen sollten *voneinander unabhängig* sein. Man spricht *von lokaler stochastischer Unabhängigkeit* und *Eindimensionalität* eines Tests, d. h. der Test misst nur *ein* latentes Konstrukt, z. B. ausschließlich einen Bereich der Rechnungswesenkompetenz, wie er weiter oben definiert ist, und nicht auch etwa eine Mathematikteilkompetenz. Diese Forderung hat auch einen mathematischen Hintergrund: „nur wenn Ereignisse unabhängig voneinander sind, ist eine Multiplikation ihrer Wahrscheinlichkeiten erlaubt (siehe Multiplikationstheorem für unabhängige Ergebnisse)." (BÜHNER 2008, S. 303). Zur Schätzung der Itemschwierigkeiten und der Personenfähigkeiten müssen im unten dargestellten Rasch-Modell Wahrscheinlichkeiten aufmultipliziert werden, was — dem Multiplikationstheorem entsprechend — eben nur erlaubt ist, wenn sich die Testantworten nicht gegenseitig beeinflussen. Daraus ist für die praktische Testdurchführung einerseits abzuleiten, dass die Wahrscheinlichkeit der Lösung eines Items nicht vom Lösen des vorangegangenen Items beeinflusst sein darf. Daher beinhaltet der WBB keine derartigen Items. Andererseits muss gewährleistet sein, dass die Lösungswahrscheinlichkeit einer Aufgabe nicht davon beeinflusst ist, ob eine andere Person das Item löst, weshalb bei Testdurchführung drauf geachtet wurde, dass einzelne Schüler/innen nicht vom Sitznachbarn oder der Sitznachbarin abschreiben.
2. Die Ermittlung der Itemschwierigkeit sollte unabhängig von der getesteten Personengruppe und die Ermittlung der Personenfähigkeit sollte unabhängig von den eingesetzten Items erfolgen. Man spricht von *Invarianz* bzw. *spezifischer Objektivität.* Items besitzen nach der probabilistischen Testtheorie immer dieselbe Schwierigkeit, egal von welcher Person sie gelöst werden. Das steht nicht im Widerspruch dazu, dass sich Personen mit einer höheren Fähigkeitsausprägung leichter beim Lösen des Items tun, als Personen mit geringerer Fähigkeitsausprägung. *Ein Test soll immer dieselbe Fähigkeit und denselben Unterschied zwischen Personen messen, egal welche Personen (Personenuntergruppen, z. B. Männer oder Frauen) getestet und welche Item (Itemuntergruppen, z. B. nur jedes zweite Item) eingesetzt werden* (vgl. BÜHNER 2008, S. 308).
3. Die Skalenqualität der Messwerte sollte mindestens auf dem Niveau der Intervallskalen[1] liegen, damit die Höhe des Unterschieds im Testwert zweier Personen interpretiert werden kann. Um das zu gewährleisten, muss der Testung ein Messmodell zugrunde gelegt werden — man spricht von einer *Verrechnungsvorschrift.* „Unter einem Messmodell für Tests versteht man vereinfacht eine Funktion, mit der man prognostizieren kann, welche Antwort eine Person auf ein Item gibt." (BÜHNER 2008, S. 310). Ein in der Psychometrie viel verwendetes Messmodell ist das Rasch-Modell (RASCH 1960), das auch der Fähigkeitsschätzung im Rahmen der PISA-Studie zugrunde liegt. Das Rasch-Modell nimmt eine probabilistische Beziehung zwischen Personen-

1 Das Intervallskalenniveau (z. B. beim IQ) beschreibt ein Messniveau, das im Gegensatz zu Ordinal- (z. B. bei den Schulnoten) und Nominalskalen (z. B. beim Geschlecht) erlaubt den Abstand zwischen zwei Werten zu messen bzw. zu interpretieren.

fähigkeit und Itemantwort an (BÜHNER 2008, S. 312). D. h. es berücksichtigt (im Gegensatz zur klassischen Testtheorie), „dass es — wenn auch mit geringer Wahrscheinlichkeit — möglich ist, dass eine fähigere Person ein im Verhältnis leichteres Item nicht löst und eine weniger fähige Person ein im Verhältnis schwereres Item lösen kann" (ebd.), sodass „für jedes Item unabhängig von der Itemschwierigkeit für jede Personenfähigkeit auf der x-Achse eine Lösungswahrscheinlichkeit ermittelt werden [kann]" (ebd., S. 313). Die Lösungswahrscheinlichkeit lässt sich durch folgende Formel ermitteln (Herleitung bei ROST 2004, S. 115ff.):

$$p\,[X_{vi} = x] = \frac{\exp\,[x_{vi}\,(\theta_v - \sigma_i)]}{1 + \exp\,(\theta_v - \sigma_i)},\ x = 0,1$$

θ = Personenfähigkeit, σ = Itemschwierigkeit

Im Rasch-Modell sind konstante Itemtrennschärfen festgelegt, d. h. alle Testaufgaben trennen in gleicher Weise zwischen fähigen und weniger fähigen Personen. In diesem Fall ist der Fähigkeitswert einer Person nicht länger davon abhängig, welche Items gelöst wurden, sondern nur wie viele, da jedes Item „gleich viel wert ist" (BÜHNER 2008, S. 344), was zum angestrebten Intervallskalenniveau der Daten führt.

Sind die beobachteten Testdaten „Rasch-Modell-konform", so erfüllen die Items des Tests alle oben angeführten Anforderungen an gute psychologische Tests (Eindimensionalität und lokale stochastische Unabhängigkeit, spezifische Objektivität bzw. Testinvarianz über Personen- und Itemsubgruppen, erschöpfende Statistiken bzw. gleiche Trennschärfen). Zusammenfassend kann festgehalten werden: wenn die Testdaten dem Rasch-Modell entsprechen, dann „sagt der Summenwert der Itemantworten auch wirklich etwas über den Ausprägungsgrad einer Person auf der latenten Variable (Fähigkeit) aus." (ebd., S. 33).

Ergebnisse testtheoretischer Analysen zum WBB

Der WBB setzt sich aus folgenden Aufgabenbereichen zusammen:

1. *Bilanzerstellung:* Bilanzposten müssen dem Anlage-, Umlaufvermögen, Eigen- und Fremdkapital zugeordnet werden. (8 Items)
2. *Kontenarten:* Die Verbuchung bestimmter Geschäftsfälle muss den Kontenarten aktives, passives Bestandskonto, Aufwand und Ertrag zugeordnet werden. (8 Items)
3. *Kontenseite:* Es muss erkannt werden, ob der Sachverhalt auf Soll- oder Habenseite verbucht wird. (8 Items)
4. *Verbuchung laufender Geschäftsfälle ohne Belege* (13 Items)
5. *Inventur:* Berechnung des Aufwands und Verbuchung (2 Items)
6. *Privatentnahme:* Darstellung der Konten Eigenkapital und Privat (5 Items)
7. *Verbuchung laufender Geschäftsfälle mit Belegen* (3 Items)
8. Bei allen Buchungssätzen musste die *Gewinnauswirkung* bestimmt werden. (16 Items)

Um zu zeigen, ob die Antwortmuster auf diese Items, die im Rahmen der Pilottestung des WBB erhoben wurden, dem Rasch-Modell entsprechen und die Testaufgaben somit alle genannten Testanforderungen erfüllen, stehen mehrere Verfahren zur Verfügung, von denen die drei gängigsten die grafische Modellkontrolle, Signifikanztests (z. B. Pearson-Chi²-Test) und der Vergleich von Modellen sind.[2] Die Überprüfung der Testitems zeigt, dass die Aufgabenbereiche 1 „Bilanzerstellung", 2 „Kontenarten", 3 „Kontenseite" und 6 „Privatentnahme" nicht Rasch-skalierbar sind. Auch die Berücksichtigung von Ratewahrscheinlichkeiten bei den Multiple-Choice-Formaten in den Aufgabenbereichen 2 und 3 führten zu keiner signifikant besseren Passung der Daten an das Modell. Zudem wurde die Verbuchung des Bankauszugs auch in den meisten Schulklassen des zweiten Jahrgangs noch nicht durchgenommen, weshalb sie ebenfalls von der Analyse ausgenommen wurde. Aus diesen Gründen beschränken sich die folgenden Analysen auf die Items aus den Bereichen 4, 5, 7 und 8, wobei die ersten drei Bereiche zur Kategorie „Buchungssätze" (16 Items) zusammengefasst werden und der letzte Bereich die Kategorie „Gewinnauswirkung" (13 Items[3]) bildet.

1. Grafische Modellgeltungskontrolle

„[Zur grafischen Modellprüfung, Anm.] wird die Stichprobe zuerst in zwei Teilstichproben geteilt (z. B. Intelligenz am Median). Für beide Stichproben werden nun die Itemparameter [= Itemschwierigkeiten, Anm.] bestimmt [= mit der (bedingten) Maximum-Likelihood-Funktion geschätzt, Anm.]. Die Itemparameter werden dann in einem Streudiagramm dargestellt. Liegen diese auf der Diagonalen des Streudiagramms, fallen die Itemparameter in beiden Stichproben gleich aus, und man kann von Modellgeltung ausgehen." (BÜHNER 2008, S. 343). Der Nachteil dieser grafischen Methode liegt darin, dass es viele Möglichkeiten gibt, die Stichproben aufzuteilen.

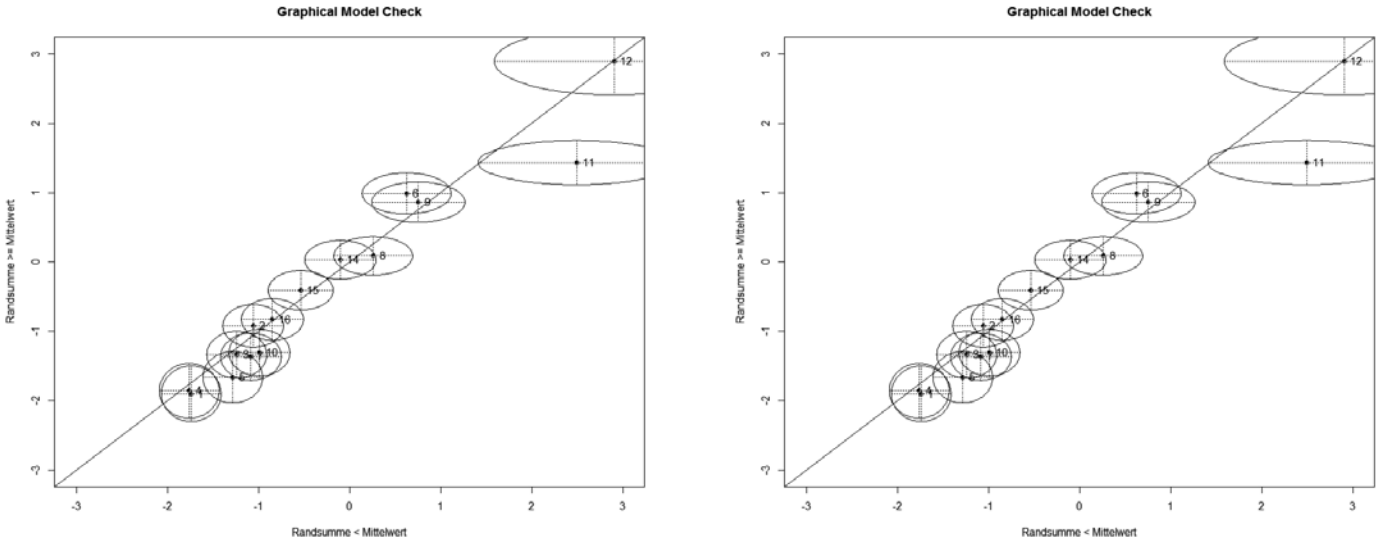


Abbildung 1 und 2: Grafische Modellgeltungskontrolle für „Buchungssätze" und „Gewinnauswirkung"

Konventionell wird die Stichprobe (1) am Median oder Mittelwert in die Gruppe der fähigeren und weniger fähigen Proband/inn/en unterteilt sowie (2) am Kriterium Geschlecht in männlich und weiblich. Für den WBB schien zudem die Unterteilung in Hinblick auf die Schularten zielführend. Allerdings stellte sich diese aufgrund der zu geringen Anzahl von Schüler/inne/n aus kaufmännischen Schulen in der Stichprobe als nicht aussagekräftig heraus.

Abbildung 1 und 2 zeigen die geschätzten Aufgabenschwierigkeiten für die Personen mit niedrigen und hohen Testrohwerten, die Kreise/Ellipsen bilden die Konfidenzintervalle. Es zeigt sich, dass alle Testitems der Kategorie „Buchungssätze" (linke Abbildung) der grafischen Modellkontrolle standhalten und somit für beide Personengruppen gleich leicht/schwer zu lösen sind bzw. hier keine signifikanten Unterschiede bestehen. Für die Testitems der Kategorie „Gewinnauswirkung" (rechte Abbildung) gilt dasselbe. Wählt man als Teilungskriterium das Geschlecht, so sind die Items „Bareinlage auf das Bankkonto" und „Zinserträge des Bankkontos" auszuscheiden. Sie liegen (inkl. Konfidenzintervall) jeweils oberhalb der Winkelhalbierenden und sind daher für Männer statistisch signifikant leichter zu lösen als für Frauen. Der Andersens Likelihood-Quotienten-Test stellt das mathematische Pendant der grafischen Modellgeltungskontrolle dar. P-Werte kleiner .05 zeigen signifikante Modellverletzungen an. Für die beiden Itemkategorien liegen p-Werte von .349 (Chi2-Wert: 17.573, df: 16) und .275 (Chi2-Wert: 14.417, df: 12) vor und somit Personenhomogenität.

2. Der Pearson-Chi²-Test

Die Modellprüfung nach dem Pearson-Chi²-Test erfolgt in drei Schritten:

1. „Der Pearson-Chi²-Test prüft die Abweichungen der Antwortmuster, die unter dem Rasch-Modell zu erwarten sind, von den tatsächlich beobachteten Antwortmustern." (BÜHNER 2008, S. 346). Beispielsweise ist die Wahrscheinlichkeit des Auftretens des Antwortmusters (0001) sehr gering. D. h., dass jemand alle einfachen Items falsch, aber das schwierigste Item richtig löst, ist sehr unwahrscheinlich. Tritt ein solches Antwortmuster in den Daten sehr oft auf, dann wird die *Prüfgröße* erhöht.
2. Aus den erhobenen Daten werden weitere *Stichproben simuliert, für die das Rasch-Modell gilt;* dies geschieht im sogenannten Bootstrap-Verfahren. In der Regel werden etwa 200 Stichproben simuliert, für die ebenfalls wie in Schritt 1 eine Prüfgröße errechnet wird.
3. Die Prüfgröße aus Schritt 1 wird nun mit den Prüfgrößen aus Schritt 2 verglichen. Ist die Prüfgröße aus den *beobachteten Daten niedriger* als die Prüfgrößen, die zu den 5%[4] der *größten* Prüfgrößen aus den 200 geschätzten Stichproben zählen, so liegt Rasch-Modellgültigkeit vor. Mit anderen Worten, bei einem p-Wert kleiner als 5% wird die Nullhypothese, dass Modellgeltung vorliegt, verworfen. Für die Itemkategorien „Buchungssätze" und „Gewinnauswirkungen" ergeben sich folgende *Goodness of Fit*-Statistiken bzw. p-Werte: .450 und .245, was für die Geltung der Rasch-Modelle bei den vorliegenden Daten spricht.

3. Entscheidung zugunsten des besseren Modells

Das Rasch-Modell stellt ein strenges logistisches Modell dar, da es von gleicher Itemtrennschärfe für jedes Item ausgeht. Lockert man diese Annahme und lässt unterschiedliche Trennschärfen zu, so spricht man von einem 2-parametrischen Modell. Wird zusätzlich, wie bei den „Gewinnauswirkungsfragen" im WBB nötig, ein Rateparameter berücksichtigt, so spricht man von einem 3-parametrischen Modell. Wenn die Daten besser zum 2- bzw. 3-parametrischen Modell passen, dann ist die Annahme des Rasch-Modells zu verwerfen. Ein *Analysis of Variance*- bzw. Likelihood-Ratio Test (p-Wert: .064) zeigt, dass die Antwortmuster der Testitems aus der Kategorie „Gewinnauswirkung" nicht signifikant besser durch ein 3-parametrisches Modell — in dem Rateparameter und Trennschärfe frei geschätzt werden — vorhergesagt werden können als durch das Rasch-Modell.

Itemanalysen

Die Abbildungen 3 und 4 zeigen den probabilistischen Zusammenhang zwischen Personenfähigkeit und der Lösungswahrscheinlichkeit eines Items: je höher die Personenfähigkeit desto höher die Wahrscheinlichkeit den Geschäftsfall richtig zu verbuchen, bzw. die Auswirkung auf den Unternehmensgewinn korrekt einzuschätzen, wobei sich die Wahrscheinlichkeiten asymptotisch 0 und 1 annähern. Die

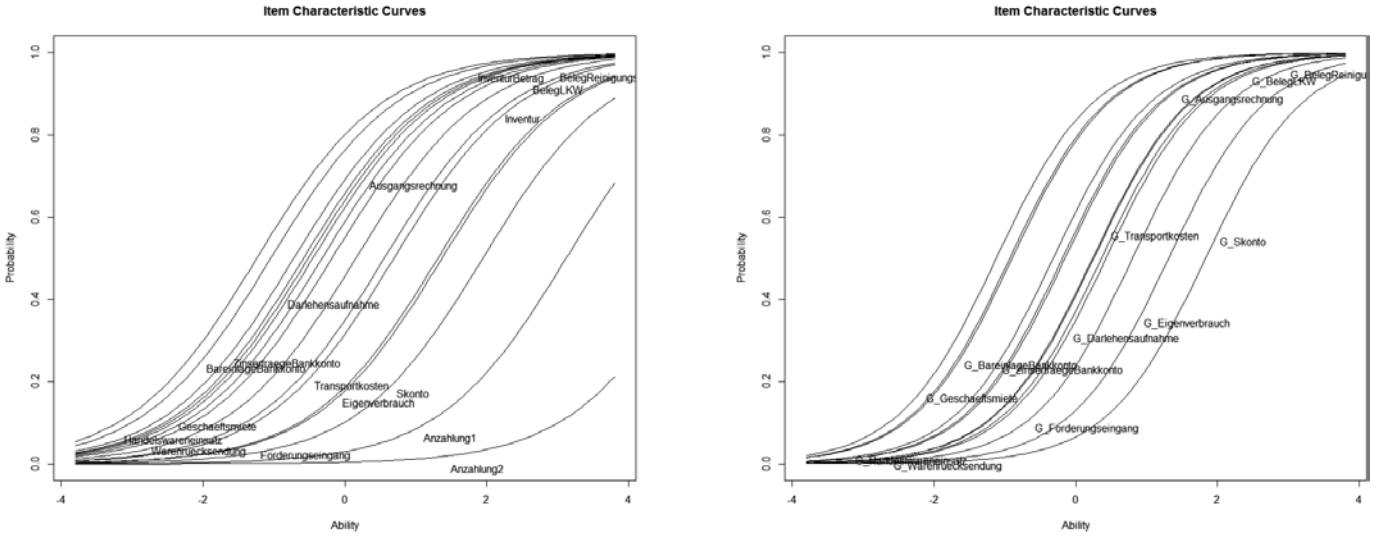


Abbildung 3 und 4: Item Characteristic Curves für „Buchungssätze" (li.) und „Gewinnauswirkung" (re.)

2 Alle psychometrischen Analysen wurden mit dem Statistikprogramm R und den Paketen „ltm" (RIZOPOULOS 2011) und „eRm" (MAIR, HATZINGER & MAIER 2011) durchgeführt.

3 Die Gewinnauswirkung beim Buchungssatz zur Inventur wurde nicht erfragt; zwei Items (zur Anzahlung) wurden von zu wenigen Proband/inn/en gelöst.

4 Alternativ wird oft auch 10% oder 20% empfohlen, um die Nullhypothese der Modellkonformität strenger zu prüfen.

Lage der Item Characteristic Curve auf der horizontalen Achse zeigt die Itemschwierigkeit an, die auch in Tabelle 2 abzulesen ist; negative Werte verweisen auf schwierige und positive Werte auf leichte Items:

Buchungssätze	Item-schwierigkeit	Gewinn-auswirkung	Itemschwie-rigkeit
Anzahlung 2. Teil	−4.961		
Anzahlung 1. Teil	−2.889		
Skonto	−1.547	Skonto	−2.414
Forderungs-eingang	−0.895	Eigenverbrauch	−1.685
Eigenverbrauch	−0.830	Forderungs-eingang	−0.977
Transportkosten	−0.126	Darlehens-aufnahme	−0.453
Inventur	0.029	Beleg Reinigungs-material	−0.363
Beleg LKW	0.482	Transportkosten	−0.210
Beleg Reinigungs-material	0.863	Beleg LKW	−0.184
Warenrück-sendung	1.019	Warenrück-sendung	0.370
Ausgangsrechnung	1.165	Ausgangsrechnung	0.423
Darlehens-aufnahme	1.239	Zinserträge Bankkonto	0.613
Geschäftsmiete	1.314	Bareinlage Bankkonto	1.490
Zinserträge Bankkonto	1.467	Handelswaren-einsatz	1.543
Handelswaren-einsatz	1.834	Geschäftsmiete	1.847
Bareinlage Bankkonto	1.834		

Tabelle 2: Itemschwierigkeiten

Wie erwartet erweisen sich die Verbuchung der Transportkosten, des Eigenverbrauchs und vor allem des Skontos sowie der Anzahlung als besondere Herausforderung für die Schüler/innen. Überraschend erscheint jedoch, dass auch der Ausgleich einer Forderung mittels Banküberweisung für die Schüler/innen eine Schwierigkeit darstellt — hier wurde des Öfteren fälschlicherweise die Umsatzsteuer mitverbucht. Ein ähnliches Bild zeigt sich für die Einschätzung der Gewinnauswirkungen. In Summe erscheint der Test aber eher als zu leicht für die vorliegende Stichprobe. Vor allem Items im mittleren und höheren Schwierigkeitsbereich sollten ergänzt werden, um die Aussagekraft des Tests zu erhöhen.

Objektivität, Reliabilität und Validität des WBB

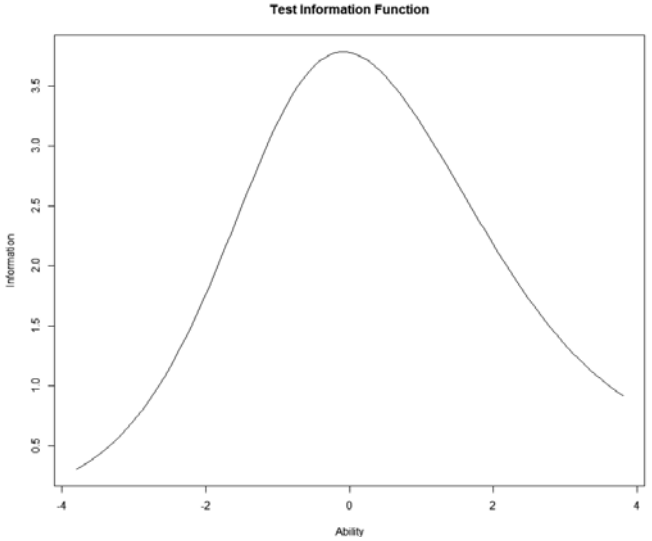
Anhand der grafischen Modellgeltungskontrolle konnte gezeigt werden, dass der WBB über Objektivität im Sinne von Item- und Personenhomogenität verfügt. Die in Abbildung 5 und 6 dargestellten Testinformationsfunktionen zeigen zudem an, wie die Messgenauigkeit des Tests in Abhängigkeit der Personenfähigkeit variiert.

Die inhaltliche Validität des Messinstruments sollte bereits im Rahmen der Testkonstruktion durch Orientierung am Prototypenansatz gewährleisten werden. Die durchgeführten Rasch-Modell-Geltungskontrollen belegen zudem die Eindimensionalität der gemessenen Personenfähigkeit. Aus inhaltlichen Gesichtspunkten kann jedoch argumentiert werden, dass auch eine mathematische Teilkompetenz miterhoben wird, da beim Berechnen der Umsatzsteuer das Prozentrechnen für korrekte Lösungen nötig ist. Die mittelstarke und hoch signifikante Korrelation von 0.56** zwischen den beiden gemessenen „Kompetenzbereichen" verweist darauf, dass es sich bei um sehr ähnliche, nicht aber gleiche Konstrukte handelt.

Tabelle 3 zeigt den Zusammenhang der beiden gemessenen Kompetenzen mit dem Außenkriterium „Schulnote". Wie erwartet korrelieren die Testrohwerte der Rasch-konformen Items des Subtests „Buchungssätze" einerseits *(in Tabelle 3 vor dem Schrägstrich)* und des Subtests „Gewinnauswirkung" andererseits *(in Tabelle 3 nach dem Schrägstrich)* vor allem mit den Zeugnis- und Schularbeitsnoten aus den Fächern Rechnungswesen und Mathematik. Interessanterweise korreliert die letzte Schularbeitsnote aus dem Fach Rechnungswesen nicht signifikant mit der Fähigkeit, Geschäftsfälle zu verbuchen. Ein Grund könnte darin liegen, dass die erhobene Schularbeitsnote im 2. Jahrgang bereits andere Stoffinhalte bzw. Fähigkeiten zum Jahresabschluss und nicht zu den laufenden Buchungen widerspiegelt.

Weitere Analysen zeigen, dass bezüglich des *Geschlechts* und der *zuhause gesprochenen Sprache* keine signifikanten Mittelwertunterschiede in beiden Fähigkeitsdimensionen vorliegen.

Dagegen klärt der Schulstandort als alleiniger Prädiktor bereits 18% bzw. 6% der Varianz in den beiden Fähigkeitsdimensionen auf. In Regressionsanalysen simultan betrachtet erweisen sich für die Kategorie „Buchungssätze" der *Schulstandort* und die *Zeugnisnote im Fach Rechnungswesen* als einzig signifikante Prädiktoren. Die Er-



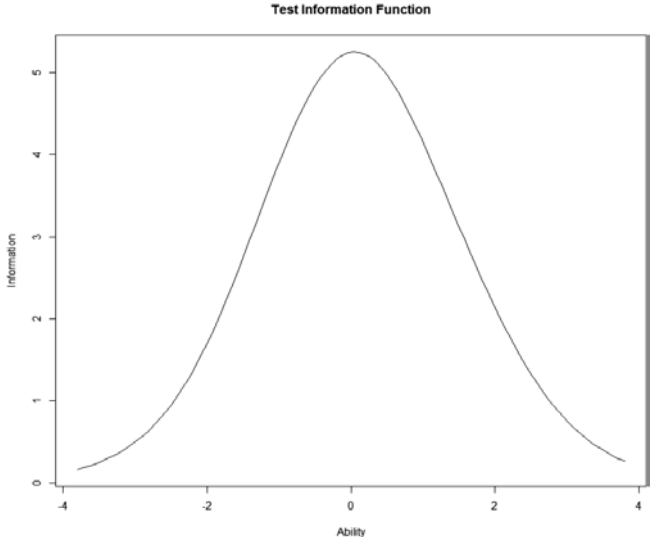


Abbildung 5 und 6: Testinformationsfunktion für „Buchungssätze" und „Gewinnauswirkung"

	letzte Zeugnisnote			letzte Schularbeitsnote (2. Jg.)		
Schulstandort	**RW**	**M**	**D**	**RW**	**M**	**D**
1 (n = 82)	–,28**/-,23*	–,25*/-,22*	n.s. / n.s.	n.s. / n.s.	n.s. / n.s.	n.s. / n.s.
2 (n = 16)	–,62*/ n.s.	–,56*/ n.s.	n.s. / n.s.	n.s. / n.s.	n.s. / n.s.	n.s. / n.s.
3 (n = 88)	–,39**/-,36**	–,43**/-,35**	–,44**/–,42**	–,52**/–,42**	–,46**/–,35**	–,44**/–,41**
4 (n = 103)	–,49**/-,33**	–,39**/–,28**	n.s. / n.s.	–,31**/–,25*	–,22*/–,27**	n.s. / n.s.
5 (n = 25)	–,57**/ n.s.	n.s. / n.s.	–,49*/ n.s.	n.s. / n.s.	–,56**/ n.s.	–,43*/n.s.
6 (n = 21)	n.s. / n.s.	n.s. / n.s.	n.s. / n.s.	n.s. / n.s.	n.s. / n.s.	n.s. / n.s.
ALLE	**–,23**/-,22****	**–,24**/–,26****	**–,18**/–,14****	**n.s./–,16****	**–,22**/–,23****	**–,12*/–,16****

Tabelle 3: Korrelationen zwischen den Fähigkeitswerten und Schulnoten je Schulstandort
Anmerkung: RW = Rechnungswesen, M = Mathematik, D = Deutsch, **/* = Korrelation ist auf dem Niveau 0.01/0.05 (2-seitig) signifikant.

gebnisse liefern Indizien für die heterogene Zusammensetzung der Schulnoten, die klassen- und schulstandortspezifisch sehr unterschiedlich ausfallen können und so die Schulnote in ihrer Funktion als Vergleichsmaßstab schwächen.

Zusammenfassende Diskussion

Ziel der vorliegenden Untersuchung ist die Erstellung eines Instrumentes zur Messung des Unterrichts- bzw. Lernerfolgs von Schüler/inne/n des 1. Jahrgangs aus berufsbildenden mittleren und höheren Schulen im Fach Rechnungswesen. Der Einsatz schulbuchtypischer Aufgaben, wie der Verbuchung von Geschäftsfällen und die Einschätzung über deren Gewinnauswirkung, zeigt sich im Gegensatz zum Einsatz von Multiple-Choice- oder Zuordnungsaufgaben aus psychometrischen Gesichtspunkten sinnvoller. Inhaltlich kann dagegen argumentiert werden, dass sich auch oder gerade in Multiple-Choice-Aufgaben das Verständnis über die Systematik der Doppik zeigen müsste, da diese Aufgabenstellungen im Unterricht nicht repetitiv eingeübt werden. Für die Erstellung eines Testinstrumentes, das den Anforderungen psychometrischer Tests gerecht wird, sollten allerdings nur Rasch-skalierbare Aufgaben ausgewählt werden. Diese Eigenschaft — so zeigt die Untersuchung — liegt vor allem bei Testitems vor, die den Schüler/inne/n nicht nur inhaltlich, sondern auch in der formalen Darbietung vertraut sind. So mussten einerseits die im Unterricht offenbar selten verwendeten Multiple-Choice-Aufgaben ausgeschieden werden, andererseits hatten Schüler/innen Probleme, Buchungssätze in ungewohnte Rastervorlagen einzutragen. Vor diesem Hintergrund erscheint es für die schulische Praxis empfehlenswert, auch andere Aufgabenformate einzusetzen, um das inhaltliche Verständnis der Schüler/innen nicht von einem Format abhängig zu machen und so den Wissenstransfer zu fördern. Einen ebenfalls zu diskutierenden Aspekt stellt die Frage dar, inwiefern zur Bewältigung der eingesetzten Testaufgaben überhaupt eine Kompetenz im Weinertschen Sinne nötig ist. So verweisen KLIEME et al. (2007, S. 74) darauf, dass „eine eng gefasste Leistungserfassung" — wie möglicherweise im vorliegenden Fall — „dem Anspruch von Kompetenzmodellen nicht gerecht werden" kann. ☒

LITERATUR

» *BECK, K & KRUMM, V. (1998). Wirtschaftskundlicher Bildungs-Test. Handanweisung. Göttingen: Hogrefe.*

» *BERLINGER, R., ACKERLAUER, I., ELLMER, M. & FREI, J. (2012). Rechnungswesen heute I/1 HAK/HAS. Linz: Trauner.*

» *BMUKK (2003). Lehrplan für die Höhere Lehranstalt für wirtschaftliche Berufe. http://www.abc.berufsbildendeschulen.at/upload/1144_FSwB%20und%20HLW.pdf [Zugriff am 6. Juni 2012].*

» *BMUKK (2004). Lehrplan für die Handelsakademie. http://www.abc.berufsbildendeschulen.at/upload/598_HAK%20LP%202004%20-%20Anlage%201.pdf [Zugriff am 6. Juni 2012].*

» *BOTHE, T., WILHELM, O. & BECK, K. (2005). Assessment of declarative business administration knowledge: Measurement development and validation. Berlin: Humboldt-Universität zu Berlin. Institut zur Qualitätsentwicklung im Bildungswesen (IQB).*

» *BÜHNER, M. (2008). Einführung in die Test- und Fragebogenkonstruktion. München: Pearson.*

» *GROHMANN-STEIGER, C., SCHNEIDER, W. & EBERHARTINGER, E. (2008). Einführung in die Buchhaltung im Selbststudium. Band I. Wien: Facultas.*

» *HABERL, K.-P., LECHNER, R., BAUER, H. & VEIDL, G. (2012). Rechnungswesen & Controlling HAK I. Wien: Manz.*

» *KLIEME, E., AVENARIUS, H., BLUM, W., DÖBRICH, P., GRUBER, H., PRENZEL, M., REISS, K., RIQUARTS, K., ROST, J., TENORT, H-E. & VOLLMER, J. H. (2007). Zur Entwicklung nationaler Bildungsstandards. Expertise. Berlin: BMBF.*

» *LEHMANN, R. H. & SEEBER, S. (Hrsg.) (2007). ULME III. Untersuchung der Leistungen, Motivation und Einstellungen der Schülerinnen und Schüler in den Abschlussklassen der Berufsschulen. Behörde für Bildung und Sport der Freien und Hansestadt Hamburg (Hrsg.). Berlin. Polyprint.*

» *MAIR, P., HATZINGER, R. & MAIER, M. (2011). eRm. Extended Rasch Modeling. 0.14-0.*

» *RASCH, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago. The University of Chicago Press.*

» *RIZOPOULOS, D. (2011). ltm. Latent Trait Models under IRT. 09-7.*

» *ROST, J. (2004). Lehrbuch Testtheorie – Testkonstruktion. Bern. Huber.*

» *SAGEDER, J. (2003). Die Wirtschaft geht uns alle an! Unveröffentlichte Testversion. Linz. Johannes Kepler Universität. Abteilung für Pädagogik und Pädagogische Psychologie.*

» *SCHUMANN, S., EBERLE, F., OEPKE, M., PFLÜGER, M., GRUBER, C., STAMM, P. & PEZZOTTA, D. (2010). Inhaltsauswahl für den Test zur Erfassung ökonomischen Wissens und Könnens im Projekt „Ökonomische Kompetenzen von Maturandinnen und Maturanden (OEKOMA)". Forschungsbericht. Zürich. Universität Zürich. Institut für Gymnasial- und Berufspädagogik. http://www.ife.uzh.ch/igb/forschungsprojekte/oekonomiekompetenz/ergebnisse/Bericht_Inhaltsauswahl_OEKOMA.pdf [Zugriff am 6. Juni 2012].*

» *WINTHER, E. (2010). Kompetenzmessung in der beruflichen Bildung. Bielefeld. Bertelsmann.*

» *ZLATKIN-TROITSCHANSKAIA, O. & KUHN, C. (2010). Messung akademisch vermittelter Fertigkeiten und Kenntnisse von Studierenden bzw. Hochschulabsolventen: Analyse zum Forschungsstand (Arbeitspapiere Wirtschaftspädagogik Nr. 56). Mainz: Lehrstuhl für Wirtschaftspädagogik, Johannes Guttenberg-Universität.*