universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Ausgewählte Facetten der Studieneignung für technische Studienrichtungen – Konzepte der Gestaltung psychologisch-diagnostischer Verfahren“

verfasst von / submitted by

Mag. rer. nat. Mag. rer. soc. oec. Lisbeth Weitensfelder

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktorin ODER Doktor der Naturwissenschaften (Dr. rer. nat.)
Doktorin OR Doktor der Naturwissenschaften (Dr. rer. nat.)

Wien, 2017 / Vienna 2017

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 091 298
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Psychologie
Betreut von / Supervisor:	Univ.-Prof. i.R. Mag. Dr. Klaus Kubinger

Meiner Tochter Elea
sowie
dem Andenken Erikas gewidmet

Danksagung

In erster Linie gilt mein Dank Univ.-Prof. Mag. Dr. Klaus D. Kubinger für seine Betreuung, aber auch für all die Jahre, in denen ich fachlich wie menschlich unter seiner Führung lernen durfte. Ich hätte mir keinen besseren Mentor wünschen können: Die Art und Weise, wie er Diagnostik lehrt und lebt, hat mich zutiefst inspiriert und in mir eine Zuneigung für den Fachbereich geweckt, die wohl nachhaltig bestehen bleiben wird.

Des Weiteren gilt mein Dank meiner Familie, meinen Eltern und meinen verstorbenen wie lebenden Großeltern, die mir mein Studium ermöglichten und mich – wie es auch meine Geschwister taten – zum Durchhalten ermutigten. Ebenso danke ich meiner geliebten Tochter Elea, die bereits in so jungen Jahren Verständnis für manche Zeiteinschränkungen hatte, und meinem ehemaligen Lebensgefährten Daniel, der diese Dissertation über weite Strecken mitgetragen hat. Meinem Cousin Christoph und meinem Freundeskreis, die mich jahrelang ermunterten, danke ich für die Kraft, die letztlich zum Durchhalten beitrug. Viel Kraft über all die Jahre gab mir auch meine über Ecken verwandte und liebe Freundin Erika, die mit den Höhen und Tiefen der Dissertation mitlebte und in den letzten Tagen der Fertigstellung völlig unerwartet verstarb. Ihrem Andenken sei diese Arbeit mitgewidmet.

Meinen ehemaligen KollegInnen der (damaligen) *Test- und Beratungsstelle* sowie des Arbeitsbereichs Psychologische Diagnostik allgemein danke ich für die Freundschaft, Unterstützung und Inspiration, insbesondere Fabian Becker, Suzana Canji, Martina Frebort, Bettina Hagenmüller, Stefana Holocher-Ertl, Lale Khorramdel, Thomas Scheck, Takuya Yanagida und Sandra Ztrivas.

Diese Dissertation wäre aber ohne Mithilfe von vielen Seiten nicht möglich gewesen, und auch dafür bin ich dankbar: Für die Unterstützung seitens der Kooperationspartner auf der TU Wien, insbesondere Mag. Ilona Herbst, Dr. Franz Reichl und dem damaligen Vizerektor für Lehre, Univ.-Prof. DI Dr. Adalbert Prechtl. Dankbar bin ich auch für zahlreiche Anregungen, Ideen und Hilfestellungen von beteiligten Lehrenden und Studierenden seitens der TU Wien, deren namentliche Aufzählung zu zahlreich wäre; viele von ihnen trugen zur Verbesserung der entwickelten Verfahren bei oder lieferten gar wertvolle Inhalte. Ähnlich verhält es sich mit PraktikantInnen und DiplomandInnen der Psychologie, die im Rahmen von Item- und Testerstellungen und -überprüfungen beteiligt waren, und somit in unterschiedlichen Ausmaßen zur vorliegenden Arbeit beigetragen haben. Letzten Endes gilt mein Dank auch jenen ratsuchenden Studierenden und Studieninteressierten, für welche die in der Dissertation angesprochenen Testbatterien konstruiert sind: Diejenigen, die ich noch zu Zeiten der *Test- und Beratungsstelle* persönlich beraten durfte, dienten als Inspiration, und derjenige größte Teil, der die Testbatterie online und anonym durchlaufen hat, stellt letztlich die Datenquelle und somit die Basis der vorliegenden Arbeit dar.

Inhaltsverzeichnis

DANKSAGUNG **5**

INHALTSVERZEICHNIS **7**

EINLEITUNG **9**

1. SELF-ASSESSMENTS IM RAHMEN DER STUDIENWAHLBERATUNG **10**

DIE SELF-ASSESSMENTS VOM „WIENER AUTORENKOLLEKTIV STUDIENBERATUNGSTESTS“ 12

2. ERMITTLUNG VON ANFORDERUNGEN FÜR TECHNISCHE STUDIENRICHTUNGEN **14**

3. KOGNITIVE VS. NICHT-KOGNITIVE PRÄDIKTOREN FÜR DEN STUDIENERFOLG **19**

4. DIE OPERATIONALISIERUNG VON ANFORDERUNGEN **23**

4.1 DIE ERFASSUNG KOGNITIVER FÄHIGKEITEN AM BEISPIEL EINES NEU ENTWICKELTEN RAUMVORSTELLUNGSTESTS 23

4.1.1 Konstruktion des TARV *24*

4.1.2 Weitere Gestaltungsprinzipien des Tests *27*

4.2 DIE ERFASSUNG NICHT-KOGNITIVER ANFORDERUNGEN AM BEISPIEL STUDIENFACHSPEZIFISCHEN INTERESSES 28

4.3 ALLGEMEINE GESTALTUNGSPRINZIPIEN: TESTREIHENFOLGEN 30

5. DISKUSSION UND AUSBLICK **32**

6. LITERATUR **35**

ANHANG A: FACHZEITSCHRIFTENBEITRÄGE (ORIGINALIA) **43**

ANHANG B **133**

ANHANG C: ABSTRACT UND ZUSAMMENFASSUNG **167**

ABSTRACT (ENGLISH) 168

ZUSAMMENFASSUNG (DEUTSCH) 169

LEBENSLAUF **171**

WISSENSCHAFTLICHE PUBLIKATIONEN **172**

Einleitung

An österreichischen (wie auch anderen) Universitäten stellt sich bei begrenzten Ressourcen immer mehr die Frage der optimalen Studienplatzbewirtschaftung – nämlich dahingehend, wie eine oft begrenzte Anzahl an Studienplätzen bestmöglich mit geeigneten Studierenden besetzt werden kann. Während viele Studienrichtungen zur Bestenauswahl mittlerweile auf Aufnahmeprüfugen setzen (müssen), finden sich mit *Self-Assessments* aber auch Bemühungen seitens der Universitäten, Studieninteressierten vorab ein Instrumentarium zur Selbstselektion und/oder zur vertieften Information bereitzustellen.

Sofern es sich bei derartigen Instrumentarien nicht lediglich um Informationen über den Studiengang, den Studienort etc. handelt, sondern die Studieninteressierten auch über die notwendigen Anforderungen für die Studienrichtung und ihre diesbezügliche „Passung" hierfür aufgeklärt werden sollen, stellt sich jedoch die Frage, welche Eigenschaften letztliche benötigt werden, um die jeweilige Studienrichtung auch studieren zu können. Für die Umsetzung seitens der Universitäten stellt sich dann in einem weiteren Schritt die Frage, wie diese Anforderungen bestmöglich erfasst werden können. Dies ist die zugrundeliegende Frage der vorliegenden Dissertation: Nämlich, welche Eigenschaften (hier am Beispiel technischer Studienrichtungen) relevant sind, und wie diese am besten erfasst werden können. Die zu beantwortende Fragestellung lautet demnach: „Welche Eigenschaften spielen insbesondere in technischen Studienrichtungen eine Rolle für den Studienerfolg, und wie kann eine Messung dieser Eigenschaften bestmöglich operationaliert werden?"

Beispielhaft werden zur Beantwortung die *Self-Assessments* des „Wiener Autorenkollektivs Studienberatungstests" (kurz: *Wiener Self-Assessments*) herangezogen, und zwar jene Testbatterien für technische Studienrichtungen, die von der Technischen Universität Wien für die Studienrichtungen Architektur, Maschinenbau, Elektrotechnik und Informatik bereitgestellt werden (vgl. hierzu auch Kubinger, Frebort, Khorramdel & Weitensfelder, 2012, 2013). Auf Grundlage von Daten aus den *Wiener Self-Assessments* wurden drei Fachartikel für „Peer-Reviewed Journals" verfasst: Zwei zur ausgewählten Anforderung Interesse und deren Operationalisierung (Weitensfelder, subm.a, subm.b), sowie ein Artikel zur Reihenfolgengestaltung einer derartigen Testbatterie in ihrer Gesamtheit (Weitensfelder, in press). Darüber hinaus finden sich aber noch weitere Publikationen, die mit der Dissertation in Verbindung stehen, insbesondere weitere Testentwicklungen zum Thema (siehe im Herausgeberwerk von Kubinger, Frebort, Khorramdel & Weitensfelder, 2012, insbesondere Weitensfelder, 2012; Weitensfelder & Becker, 2012; Weitensfelder & Hofer, 2012; Weitensfelder, Undeutsch, Khorramdel & Useini, 2012) sowie ein Vortrag, zu dem der Fachartikel in einem separaten Anhang wiedergegen ist.

1. Self-Assessments im Rahmen der Studienwahlberatung

Unter Self-Assessments versteht man allgemein psychologische Untersuchungssiutationen, bei denen „sich eine Person (zumeist über das Internet) selbst [testet] und … folglich den diagnostischen Prozess eigenerantwortlich und fachpsychologisch unkontrolliert [absolviert]“ (Kubinger, 2009, S. 27). Self-Assessments können dabei unterschiedliche Ziele verfolgen und in ausbildungsübergreifende (z.B. zur Entscheidung für vs. gegen ein Studium), fächerübergreifende (z.B. für mögliche Studienangebote), studienfeldspezifische (z.B. für naturwissenschaftliche Studienrichtungen) und studienfachspezfische (für eine konkrete Studienrichtung) Self-Assessments unterschieden werden (vgl. Milbradt, Zettler, Putz, Heukamp & Hornke, 2008).

Während Self-Assessment ursprünglich zur Selektionsdiagnostik im Sinn eines *Screenings* gedacht waren, haben diese nun vermutlich im Rahmen der Studienberatung größere Bedeutung erlangt (z.B. Kubinger, Frebort & Müller, 2012) und werden mittlerweile von vielen deutschsprachigen Universitäten angeboten. Ein frühes Beispiel für ein studienfachspezifisches Self-Assessment findet sich an der RWTH Aachen, an der Self-Assessments schon seit der frühen Jahrtausendwende angeboten werden (z.B. Hornke, Wosnitza & Bürger, 2013). Aber auch an österreichischen Universitäten sind Self-Assessments zunehmend im Kommen, an den größeren Universitäten in Wien beispielsweise werden studienfachspezifische Self-Assessments für zumindest einige Studienrichtungen von der TU Wien (studienwahl.tuwien.ac.at) und der Universität Wien (self-assessment.univie.ac.at) angeboten, ein etwas breiter gefasstes Self-Assessment bietet auch die Wirtschaftsuniversität Wien an (https://www.wu.ac.at/studium/services-fuer-studieninteressierte/self-assessment-tests/), und an der Universität für Bodenkultur wird ein Self-Assessment im Rahmen eines Aufnahmeverfahrens eingesetzt (http://www.boku.ac.at/studienservices/themen/zulassung/aufnahmeverfahren-bachelor-lebensmittel-und-biotechnologie/). Dabei können Self-Assessments unterschiedliche Zugänge verfolgen, sei es, um Studieninteressierte primär über Rahmenbedingungen zu informieren und Erwartungen zurechtzurücken, oder auch um im testdiagnostischen Sinn Studieninteressierten eine „Passung“ zur intendierten Studienrichtung rückzumelden – hier könnte man auch in „umweltzentrierte“ und „personenzentrierte“ Self-Assessments unterscheiden (Hasenberg & Schmidt-Atzert, 2014); selbstverständlich sind auch Kombinationen aus diesen Angeboten möglich. Geht es um Self-Assessments zur Eignungsbeurteilung basierend auf psychologisch-diagnostischen Verfahren, ist die prognostische Validität der Verfahren in Hinblick auf die wohl interessierenden Studienerfolgsdaten allerdings schwierig (s. z.B. Kubinger, Frebort & Müller, 2012; Kubinger et al., 2013): Abgesehen von der Problematik, an Noten zu kommen, und dass Noten als Erfolgskriterium so einige Kritikpunkte aufweisen (zusammenfassend am Beispiel Schulnoten z.B. Spiel, Litzenberger & Haiden, 2007), ist eine prognostische Validität für relativ weit in der Zukunft liegende Ereignisse besonders schwierig (Kubinger, 2009). Geht es um Studiennoten, haben sich zudem Schulnoten als der beste Einzelprädiktor erwiesen (Hell, Trapman & Schuler, 2008), wobei die (deutsche) Durchschnittsnote ein besserer Prädiktor ist als Einzelnoten (Trapman, Hell, Weigand & Schuler, 2007). Studierfähigkeitstests weisen (mit Unterschieden je Studienrichtung) zwar ebenfalls recht gute Validitäten auf, sind jedoch

nicht ganz so valide wie Noten (Hell, Trapman & Schuler, 2008); demzufolge kann darauf geschlossen werden, dass die Validität von Self-Assessments, bei denen Studieninteressierte zudem eine eher „lockere“ und unkontrollierte Testsituation vorfinden, keinesfalls an die Validität von Schulnoten heranreichen. Denn insbesondere bei Self-Assessments gibt es Probleme die Stichprobe betreffend, z.B. durch Selbstselektion, denn der Schluss liegt nahe, dass bei einer frewilligen Teilnahme gerade die Motivierteren die Testbatterie durchlaufen (zur „Freiwilligenproblematik“, dass Freiwillige andere Testwerte erzielen können als jene im Rahmen einer Auswahlsituation, siehe zudem Karner, 2002). Aber auch die Unkontrolliertheit der Bearbeitung kann zu Problemen führen. Hier eröffnen sich z.B. Rahmenbedingungen, die bei üblichen Self-Assessment-Studien selten genannt werden, die sich aber ebenfalls hinderlich für Validierungsstudien erweisen: So erfolgt nämlich der Zugang zu den Self-Assessments rund um die Uhr, im Schnitt für das *Wiener Self-Assessment Psychologie* gegen 17 Uhr, und in mehr als einem Viertel der Fälle erst nach 20 Uhr (Scheck, Becker, Weitensfelder & Frebort, 2012).

Nichtsdestotrotz gibt es Studien, die die Ergebnisse von Self-Assessments mit Erfolgskriterien wie Noten in Verbindung bringen (z.B. Hornke et al., 2013), mit – trotz der oben genannten Hindernisse – berichtenswerten Erfolgen: So korrelieren beispielsweise Skalen des Self-Assessments der RWTH Aachen mit der Durchschnittsnote nach zwei Semestern Maschinenbaustudium um $r = 0{,}4$ und höher (ebd.), wobei Leistungsskalen deskriptiv etwas niedrigere Korrelationen aufzuweisen scheinen als Interesse und Motivation (Leistungsskalen: $r = 0{,}406$ mit der realen bzw. $r = 0{,}413$ mit einer so genannten „impliziten Durchschnittsnote“, bei der auch negative und später ev. ersetzte Prüfungsergebnisse berücksichtigt werden; Interesse: $r = 0{,}431$ mit der realen und $r = 0{,}445$ mit der impliziten Durchschnittsnote; Motivation: $r = 0{,}437$ mit der realen und $r = 0{,}496$ mit der impliziten Durchschnittsnote).

Vielfach werden für Studienerfolg jedoch auch andere Kriterien als die erreichten Studiennoten von Interesse sein, wie z.B. Studienzufriedenheit, Abbrecherquoten etc. Insbesondere Abbrecherquoten stellen ein Problem dar, und zwar nicht nur auf individueller Ebene, sondern auch auf Hochschulebene, wo Studienabbrüche eine „Verschwendung“ knapper Ressourcen bedeuten (vgl. z.B. Heublein & Wolter, 2011). Zwar gibt es einige Unschärfeprobleme bei der Bestimmung, welche Studierenden letztlich als AbbrecherInnen gewertet werden können, z.B., weil einige vielleicht im späteren Verlauf ihr Studium wieder aufnehmen (ebd.), insgesamt sind die Anteile allerdings recht hoch – bei deutschen Universitäten lagen die Abbruchquoten für das Bachelorstudium bei den StudienanfängerInnen 2006/2007 bei 35%, bei den AnfängerInnen 2008/2009 bei 33% (Heublein, Richter, Schmelzer & Sommer, 2014). Als gewichtigste Abbruchgründe einer deutschen groß angelegten Exmatrikulationsbefragung werden Leistungsdefizite (20% gaben Leistungsprobleme in Bezug auf die Anforderungen an, zusätzlich 11% Prüfungsversagen), finanzielle Probleme (19%) und Motivationsprobleme (18%) angegeben (Heublein, Hutzsch, Schreiber, Sommer & Besuch, 2010). Insbesondere in Bezug auf Leistungs- und Motivationsprobleme kann ein Self-Assessment zur Studienwahlberatung allerdings zur Frühaufklärung beitragen. Freilich werden einige dieser Personen ein Self-Assessment oft gar nicht durchlaufen, weil sie vielleicht nicht davon informiert sind (ebenso, wie man im umgekehrten Fall davon ausgehen muss, dass nicht alle Personen, die ein Self-Assessment

durchlaufen, das Studium tatsächlich ernsthaft in Erwägung ziehen). Dennoch gaben in einer Nachbefragung 36% der TeilnehmerInnen aus dem ersten *Wiener Self-Assessment Psychologie* an, dass das Self-Assessment sie in ihrer Studienentscheidung beeinflusst habe, 76% der Befragten würden das Self-Assessment auch gern weiterempfehlen (Sonnleitner, Kubinger & Frebort, 2009). Auch andere Studien zu Self-Assessments zeigen auf, dass diese durchaus als Entscheidungshilfe für das Studium gesehen werden können, wenngleich diese nur von einer Minderheit als zentrale Entscheidungshilfe gesehen werden (Pietrangeli, Sindern & Rudinger, 2013): In einer ersten Nachbefragung zu einem fachspezifischen Self-Assessment der Universität Bonn gaben 17% an, sich aufgrund der Ergebnisse für oder gegen dieses Studium entschieden zu haben; in einer noch später folgenden Nachbefragungsphase waren es (bei allerdings deutlich weniger Teilnehmenden) immerhin 48%. Jedenfalls scheinen aber Self-Assessments zumindest eine gute Grundlage für eine fundierte Studienfachentscheidung zu bieten: So gaben 53-86% der TeilnehmerInnen nach explizier Nachfrage an, dass sie nach Durchlaufen eines Self-Assessments ihre Studieneignung besser hätten einschätzen können (Milbradt, Zettler, Putz, Heukamp & Hornke, 2008). Ein ähnlich hoher Prozentsatz (50-81%) berichtete, die Anforderungen für das Studium besser zu kennen (ebd.).

Aber auch kritische Belange im Hinblick von Self-Assessments sollten nicht außer Acht bleiben (z.B. Kubinger, 2015): So ist der Einsatz von Self-Assessments besonders dann kritisch, wenn gewisse Grundsätze, wie jener, dass auch ein Self-Assessment eine berufsbezogene Eignungsbeurteilung darstellt und somit auf den Prinzipien der DIN 33430 (DIN Deutsches Institut für Normung e.V., 2002) beruhen sollte, übergangen werden. Eine Vorgangsweise anhand der DIN 33430 erfordert aber auch, dass Self-Assessments auf einer *state-of-the-art* gewonnenen Anforderungsanalyse beruhen, denn nur mittels Anforderungsanalyse ist sicherzustellen, dass die tatsächlich für die Studienrichtung notwendigen Fähigkeiten und Persönlichkeitseigenschaften berücksichtigt werden. Kritisch ist zudem, wenn die eingesetzten psychologisch-diagnostischen Verfahren nicht den einschlägigen Gütekriterien entsprechen, was für seriöse Self-Assessments letztlich aber zu Durchlaufzeiten von etwa zwei Stunden führt – denn anderenfalls lässt sich bei vielseitigen Anforderungen nicht hinreichend messgenau diagnostizieren (Kubinger, 2015). Explizit empfohlen (ebd.) werden für Self-Assessments letztlich auch Verfahren der sogenannten „experimentalpsychologischen Verhaltensdiagnostik“ – also solche Verfahren, in denen Persönlichkeitseigenschaften nicht lediglich erfragt, sondern vom Computer objektiv „beobachtet“ werden – dieser Umstand kommt derzeit in Self-Assessments noch wenig zu tragen, wird allerdings in den *Wiener Self-Assessments* (s.u.) bereits umgesetzt.

Die Self-Assessments vom „Wiener Autorenkollektiv Studienberatungstests“

Bezugnehmend auf die oben gebrachten Aspekte, wie denn Self-Assessments zu gestalten seien, wurden vom „Wiener Autorenkollektiv Studienberatungstests“ Self-Assessments für das Studium der Psychologie sowie für vier technische Studienrichtungen (Architektur, Maschinenbau, Elektrotechnik und Informatik) erstellt (s. hierzu z.B. Kubinger, Frebort, Khorramdel & Weitensfelder, 2012, 2013), im Folgenden kurz als *Wiener Self-Assessments* bezeichnet. Sämtliche dieser Self-Assessments basieren auf Anforderungsanalysen (s. unten)

für die jeweilige Studienrichtung; anhand der Anforderungen wurden bestehende psychologisch-diagnostische Verfahren ausgewählt (und teilweise adaptiert) oder gar neu entwickelt. Die Verfahren beinhalten Leistungstests, Persönlichkeitsfragebogen sowie Verfahren der experimentalpsychologischen Verhaltensdiagnostik (auch bekannt als Objektive Persönlichkeitstests sensu R. B. Catell).

Im Unterschied zu einer herkömmlichen psychologisch-diagnostischen Untersuchung erfolgt die Rückmeldung der Ergebnisse nicht norm-, sondern kriteriumsorientiert, d.h. anhand bestimmter, vorher festgesetzter Cut-Off-Werte: Teilweise an Normen angelehnt, teilweise aber auch (nur) fachautoritär, wurde je Verfahren festgelegt, welche Testkennwerte von TeilnehmerInnen zumindest erzielt werden sollen, um eine Stärke in Hinblick auf die geforderte Anforderung aufzuweisen.

Die Rückmeldung an die TeilnehmerInnen erfolgt (jeweils nach einer Erläuterung, welche Anforderungen gemessen wurden und weshalb diese im Studium relevant sind) im Anschluss an das Durchlaufen der Testbatterie jeweils als „Stärke“ oder „Schwäche“ in Hinblick auf die Anforderung. Dabei folgt das sehr ausführliche Feedback jedenfalls dem Prinzip der förderorientierten Diagnostik: Denn abgesehen von der Rückmeldung als „Stärken“ oder „Schwächen“ in den jeweiligen Anforderungen erhalten Testpersonen im Fall von Schwächen noch weiterführende Tipps und Hinweise; sofern die Schwächen mit Verhaltensstrategien (teilweise) kompensierbar oder zumindest teilweise trainierbar sind, werden die TeilnehmerInnen darauf hingewiesen. Im schlimmsten Fall, nämlich dass es für eine spezifische Schwäche keine Trainingsmöglichkeiten gibt, werden die Testpersonen zumindest darüber informiert, welche Bereiche des Studiums für sie voraussichtlich schwieriger sein werden und wo sie mit Mehraufwand rechnen müssen. (Zur Gestaltung der Rückmeldungen siehe ausführlicher Scheck et al., 2012).

Um zusätzlich zur testdiagnostischen Vorgehensweise noch spezifisch auf Studienerwartungen eingehen (und diese zurechtrücken) zu können, existieren für einige der Studienrichtugnen zudem sogenannte „Erwartungsfragebögen“ (s. z.B. bei Weitensfelder, Undeutsch, Khorramdel & Useini, 2012), in denen in der Regel zutreffende sowie fälschliche Erwartungen an das Studium in Itemform präsentiert werden; die Studieninteressierten sollen für sich beantworten, ob sie die erfragten Gegebenheiten erwarten oder nicht und erhalten im Anschluss an das Self-Assessment eine Rückmeldung darüber, wie ExpertInnen für die Studienrichtung (Lehrende bzw. Lehrende und Studierende) diese Gegebenheiten beurteilen. Gegen Ende des Self-Assessments werden die TeilnehmerInnen zudem aufgefordert, noch einige Fragen zu beanworten, die – abgesehen davon, dass sie auch für Evaluationsstudien von Nutzen sein können – sie dazu anhalten sollen, auch an Rahmenbedingungen zu denken, die über die eingesetzten Verfahren und meist Testdiagnostik im Allgemeinen hinausgehen, die sich aber ebenfalls auf Studienwahl oder Studienerfolg auswirken können (z.B., ob eine Berufstätigkeit neben dem Studium geplant ist, ob es Alternativen zur Studienwahl gibt etc.).

2. Ermittlung von Anforderungen für technische Studienrichtungen

Eine bewährte Methode zur Ermittlung von Anforderungen besteht bereits seit den 50er Jahren des vorigen Jahrunderts (Flanagan, 1954): Der Vorteil dieser als *Critical Incident Technique* bekannten Methode besteht darin, dass die Anforderungen verhaltensnah erhoben werden, das heißt, dass die Anforderungen durch Verhaltensbeschreibungen gewonnen werden, und nicht nach Eigenschaften gefragt wird, die vielleicht aus Gründen der sozialen Erwünschtheit genannt oder vermutet werden, aber mit den tatsächlichen Anforderungen wenig zu tun haben. Im Zuge der *Critical Incident Technique* werden erfolgskritische (im Sinn von erfolgsrelevanten) Situationen erfragt, und es wird jeweils in Erfahrung gebracht, wie sich Personen in den genannten Situationen verhalten haben. Die Verhaltensweisen werden in weiterer Folge (qualitativ inhaltsanalytisch, z.B. nach Mayring, 2008) in Eigenschaften „übersetzt", aus denen dann die Anforderungen resultieren. Ein Anwendungsbeispiel für den Studienkontext findet sich ausführlich bei Khorramdel, Maurer, Frebort & Kubinger (2012). Konkret werden dabei von ExpertInnen (das sind im Fall der Studieneignung z.B. Lehrende oder fortgeschrittene Studierende) Situationen erfragt, in denen sich besonders gezeigt hat, dass jemand im Studium erfolgreich (oder im umgekehrten Fall eben nicht erfolgreich) war. Nach einer Beschreibung der Situation werden die Verhaltensweisen erfragt, die in der jeweiligen Situation gezeigt wurden. Erst diese werden dann in die zugrundeliegenden Eigenschaften „übersetzt". So könnte beispielsweise ein Lehrender als Situation eine Besprechung der laufenden Bachelorarbeit nennen. Als beobachtetes Verhalten könnte der Lehrende erzählen, dass sich der Studierende neben der empfohlenen Literatur noch viel Zusatzliteratur besorgt habe, sich in das Thema gut eingearbeitet habe, über die bestehende Fragestellung des Professors nachgedacht und diese nicht ganz passend gefunden habe, dafür aber einen eigenen lösungsorientierten Vorschlag zur Abänderung der Fragestellung eingebracht habe, der auf tiefgehendes Verständnis hingedeutet habe etc. Diese genannten Verhaltensweisen sind es letztlich, die nach einem bestimmten Schlüssel in die resultierenden Anforderungen (hier könnte dies z.B. sein: Engagement, Gewissenhaftigkeit, kritisches Denken, Eigenständigkeit, Problemlösefähigkeit/schlussfolgerndes Denken etc.) „übersetzt" werden.

Derart wurde für die Studienrichtungen Architektur, Maschinenbau, Elektrotechnik und Informatik vorgegangen, nachfolgend (Tab. 1) ist das so resultierende Anforderungsprofil für die Studienrichtung Architektur dargestellt (Khorramdel et al., 2012).

Kategorien	Anforderungen Architektur	Häufig-keit
Motivation	Engagement, Lernbereitschaft, Durchhaltevermögen, Interesse am Fach, Erfolgszuversicht, Zielorientierung, Zielsetzung	218
Intelligenz	Kreativität, Kritisches Denken, Schlussfolgerndes Denken, Verbale Fähigkeiten, Räumliches Vorstellungsvermögen, Figurale Fähigkeiten, Numerische Fähigkeiten, Flexibilität im Denken	178
Soziale Kompetenz	Kommunikationsfähigkeit, Kontaktfähigkeit, Überzeugungsstärke, Teamfähigkeit	109
Organisation	Organisationstalent, Informiertheit, Zeitmanagement, Prioritätensetzung, Planen	108
Emotionale Kompetenz	Distanzierungsfähigkeit, Standhaftigkeit, Offenheit	61
Arbeitsstil und Arbeitsverhalten	Gewissenhaftigkeit, Eigeninitiative, Selbständiges Arbeiten, Realistische Selbsteinschätzung	59
Belastbarkeit	Frustrationstoleranz, Arbeiten unter Zeitdruck	55
Lernverhalten	Lernfähigkeit, Lernstil	24
Wissen	Fachspezifisches Wissen	22
Selbstbild	Selbstbewusstsein	22

Tab. 1: Anforderungsprofil für das Architekturstudium (vgl. Khorramdel, Maurer, Frebort & Kubinger, 2012), basierend auf Fragebögen bzw. Interviews von 33 Studierenden und 11 Lehrenden

Betrachtet man dieses Anforderungsprofil und vergleicht es mit anderen technischen Studienrichtungen, beispielsweise Elektrotechnik und Informatik, fallen viele Überlappungen ins Auge: Tabelle 2 stellt die Anforderungsprofile der *Test- und Beratungsstelle* für die Studienrichtungen Informatik und Elektrotechnik dar (Khorramdel, Weitensfelder & Frebort, unpubl.), Tabelle 3 stellt das Anforderungsprofil für das Maschinenbaustudium dar (Khorramdel & Maurer, unpubl.). Unterschiede zwischen diesen oder gar zu den Anforderungen für Architektur scheinen sich lediglich in der Gewichtung mancher Anforderungen sowie in einigen „Spezialfähigkeiten" zu ergeben, wie in räumlichen Fähigkeiten (Architektur) oder der – nur in Elektrotechnik relevanten – Anforderung der Fähigkeit, Details im Ganzen erkennen zu können und das Ganze aus Details zusammensetzen zu können, die in der Tabelle als „figurale Intelligenz" ausgewiesen ist. Die Überlappungen beschränken sich dabei nicht nur auf technische Studienrichtungen, sondern bestehen auch mit einem allgemeinen Anforderungsprofil zur Studieneignung (Khorramdel & Maurer, unpubl., in Maurer, 2009; s. Tabelle 4) – was darauf schließen lässt, dass bis auf wenige spezielle Fähigkeiten in allen Studienrichtungen großteils die selben Fähigkeiten und Persönlichkeitseigenschaften relevant sind.

Folgende Anforderungen scheinen im Verlgeich hingegen spezifisch bzw. sind nicht im allgemeinen Anforderungsprofil zur Studieneignung enthalten:

- Für Architektur: Räumliches Vorstellungsvermögen und figurale Fähigkeiten
- Für Maschinenbau: Technisches Verständnis
- Für Elektrotechnik: Figurale Intelligenz

- Sowie allgemein: Interessen (da diese je Fach unterschiedlich sind) und fachspezifisches Vorwissen.

Kategorien	**Anforderungen Informatik (…): eher selten genannt**	**Anforderungen Elektrotechnik (…): eher seltener genannt**	**Häufig-keiten (Inf/Elek)**
Motivation	Engagement, Interesse am Fach, Lernbereitschaft, Durchhaltevermögen, Erfolgszuversicht, Zielorientierung, (Wissensdurst/Neugier)	Engagement, Lernbereitschaft, Interesse am Fach, Durchhaltevermögen, Zielorientierung, Erfolgszuversicht vs. Misserfolgsängstlichkeit, Wissensdurst/Neugier, (Zielsetzung)	86 / 95
Organisation	Zeitmanagement, Planen, Organisationstalent, Informiertheit, (Prioritäten setzen)	Zeitmanagement , Organisationstalent, Informiertheit, Prioritäten setzen, Planen	81 / 73
Soziale Kompetenz	Kommunikationsfähigkeit, Kontaktfähigkeit, Teamorientierung, Teamfähigkeit, Kollegialität, (Kompromissbereitschaft, selbstbewusstes Auftreten)	Kommunikationsfähigkeit, Kontaktfähigkeit, Teamorientierung, (Teamfähigkeit, Kollegialität, selbstbewusstes Auftreten)	63 / 62
Arbeitsstil, Arbeitsverhalten	Gewissenhaftigkeit, selbständiges Arbeiten, realistisches Anspruchsniveau, (realistisches Einschätzen der eigenen Leistungsfähigkeit)	Gewissenhaftigkeit, Anpassungsfähigkeit, realistisches Einschätzen äußerer Gegebenheiten, realistisches Einschätzen der eigenen Leistungsfähigkeit/Fähigkeiten, selbständiges Arbeiten, (Eigeninitiative)	57 / 49
Intelligenz	Schlussfolgerndes Denken, verbale Intelligenz, Reflexionsvermögen, (numerische Intelligenz, Textverständnis, Flexibilität im Denken, Kreativität)	Schlussfolgerndes Denken, verbale Intelligenz, (Textverständnis, numerische Intelligenz, figurale Intelligenz, Kreativität, kritisches Denken, Umsetzten/Anwenden von Gelerntem)	33 / 33
Belastbarkeit	Frustrationstoleranz, Belastbarkeit allgemein, (Fähigkeit zur Entspannung, Belastbarkeit bei Hindernissen)	Frustrationstoleranz, Belastbarkeit allgemein, (Fähigkeit zur Entspannung)	19 / 28
Lernverhalten	Lernfähigkeit, Konzentration, (Lernstil, Merkfähigkeit)	Lernstil, Lernfähigkeit, (Konzentration)	17 / 24
Selbstbild	(Selbstbewusstsein)	Selbstbewusstsein	2 / 15
Wissen	Fachspezifisches Wissen, (Vorwissen in Programmierung und Softwareentwicklung)	Fachspezifisches Wissen, Vorwissen, (Vorwissen in Mathematik und Programmieren)	4/8

Emotionale Komponenten	Offenheit, (Extraversion)	Distanzierungsfähigkeit, Offenheit, (Selbstständigkeit)	6 / 7
Kontroll-überzeugung	(Internalität und Selbstkonzept eigener Fähigkeiten)	(Attributionsstil, Internalität und Selbstkonzept eigener Fähigkeiten)	1 / 5
Erwartungen	(Realistische Erwartungen)		1 / 0

Tab. 2: Anforderungsprofil der *Test- und Beratungsstelle* für die Studienrichtungen Informatik und Elektrotechnik an der TU Wien (Khorramdel, Weitensfelder & Frebort, unpubl.); die Anforderungsanalysen basieren auf 3 Lehrenden-Interviews und 29 Studierenden-Fragebögen (Informatik) bzw. 5 Lehrenden-Interviews und 26 Studierenden-Fragebögen (Elektrotechnik).

Kategorien	**Anforderungen Maschinenbau**	**Häufig-keit**
Motivation	Engagement, Interesse am Fach, Durchhaltevermögen, Lernbereitschaft, Zielorientierung, Erfolgszuversicht vs. Misserfolgsängstlichkeit	152
Arbeitsstil, Arbeitsverhalten	Gewissenhaftigkeit, Selbständiges Arbeiten, Eigeninitiative, Genaues Arbeiten, Realistische Selbsteinschätzung	94
Intelligenz	Schlussfolgerndes Denken, Flexibilität im Denken, Numerische Fähigkeiten, Technisches Verständnis	84
Organisation	Zeitmanagement, Organisationstalent, Prioritäten setzen, Informiertheit, Planen	69
Lernverhalten	Lernfähigkeit, Lernstil	40
Soziale Kompetenz	Kommunikationsfähigkeit, Kontaktfähigkeit	30
Belastbarkeit	Frustrationstoleranz	20
Kontroll-überzeugung	Internalität	12
Wissen	Vorwissen, Fachspezifisches Wissen	11
Selbstbild	Selbstbewusstsein	6
Rahmen-bedingungen	studienbezogene, personenbezogene	6
Emotionale Komponenten	Offenheit, Ehrlichkeit	5

Tab. 3: Anforderungsprofil für das Maschinenbaustudium (Khorramdel & Maurer, unpubl.), basierend auf 10 Lehrenden-Interviews und 19 Studierenden-Fragebögen

Kategorien	Anforderungen allgemeine Studierfähigkeit (die am häufigsten genannten mit n>30 Nennungen)	Häufig-keit
Motivation	Engagement, Interesse, Durchhaltevermögen, Lernbereitschaft, Zielorientierung, Erfolgszuversicht, Zielsetzung, Wissensdurst/Neugier, Ehrgeiz	1343
Organisation	Organisationstalent, Informiertheit, Zeitmanagement, Prioritäten setzen, Planen	776
Intelligenz	Schlussfolgerndes Denken, Kritisches Denken, Verbale Intelligenz, Flexibilität im Denken, Kreativität, Numerische Intelligenz, Umsetzung/Anwendung von Gelerntem	738
Arbeitsstil, Arbeitsverhalten	Gewissenhaftigkeit, Selbständiges Arbeiten, genaues Arbeiten, Eigeninitiative, realistisches Einschätzen der eigenen Leistungsfähigkeit/Fähigkeiten	591
Soziale Kompetenz	Kommunikationsfähigkeit, Kontaktfähigkeit, Teamfähigkeit, Überzeugungsstärke	567
Belastbarkeit	Frustrationstoleranz	229
Lernverhalten	Lernfähigkeit, Lernstil, Konzentration/zielgerichtete Aufmerksamkeit	224
Emotionale Komponenten	Offenheit, Distanzierungsfähigkeit, Standhaftigkeit, Selbständigkeit	174
Wissen	Fachspezifisches Wissen	78
Selbstbild	Selbstbewusstsein	76
Rahmen-bedingungen	(jeweils n < 30: studienbezogene bzw. umweltbezogene)	38
Kontroll-überzeugung	Internalität	38

Tab. 4: Anforderungsprofil zur allgemeinen Studierfähigkeit (Khorramdel & Maurer, unpubl., in Maurer, 2009), basierend auf Fragebogen-Erhebungen und Interviews von 208 Studierenden und 71 Lehrenden

3. Kognitive vs. nicht-kognitive Prädiktoren für den Studienerfolg

Nach der so genannten „Verhaltensgleichung“ von Westhoff und Kluck (2003) bestimmt sich menschliches Verhalten, somit auch das Verhalten im Studium, aus unterschiedlichen Variablengruppen (Umgebungsvariablen, Organismusvariablen, kognitive, emotionale, motivationale und soziale Variablen) und deren Wechselwirkungen. Um Verhalten bestmöglich vorhersagen zu können, sind all jene zu berücksichtigen. Auch die mittels *Critical Incident Technique* gewonnenen, oben dargestellten (und bis auf die Rahmenbedingugen psycholologischen) Anforderungen lassen sich in kognitive und nicht-kognitive Anforderungen unterscheiden. Während Westhoff und Kluck (2003) auch Arbeitsstil den kognitiven Variablen zuordnen, räumen diese allerdings ein, dass auch andere Variableneinteilungen sinnvoll möglich sind. Entsprechend werden im vorliegenden Fall unter kognitiven Anforderungen primär Merkmale der geistigen Leistungsfähigkeit (wie Intelligenz) verstanden, während Persönlichkeitseigenschaften, Stilmerkmale, emotionale und motivationale Variablen lapidar als „nicht-kognitive“ Anforderungen bezeichnet werden. Dabei drängt sich die Frage auf, ob beide, kognitive wie nicht-kognitive Anforderungen, gleichermaßen zum Studienerfolg beitragen. Untersucht an einer Zusammenstellung von Stichproben aus der Pilotierungsphase der *Wiener Self-Assessments* für *Architektur* (Müller, 2011) bzw. *Maschinenbau* (Mitschek, 2011) zeigt sich, dass es besonders die nicht-kognitiven Prädiktoren sind, die einen besonderen Stellenwert für das untersuchte Erfolgskriterium (nämlich die durchschnittliche Studiengeschwindigkeit) zu haben scheinen (vgl. Kubinger, Weitensfelder, Frebort & Sonnleitner, 2012, siehe auch Anhang B). Tabelle 5 gibt die Korrelationsmatrix für das in der Studie untersuchte Erfolgskriterium wieder: Dabei finden sich geringe Korrelationen zwischen dem Erfolgskriterium und ausgewählten Persönlichkeitseigenschaften, aber keine Korrelationen mit ausgewählten kognitiven (Leistungs-)Testergebnissen. Der Schluss, dass kognitive Prädiktoren für den Studienerfolg vernachlässigbar wären, kann jedoch nicht gezogen werden: Denn einerseits ist die vorliegende Stichprobe bereits insofern vorselektiert, dass nur bereits Studierende (die eine eher leistungsfähige Stichprobe darstellen) einbezogen wurden, andererseits wurden in der Korrelationsmatrix unvollständige Testvorgaben nicht mit berücksichtigt – was aber für den eingesetzten Lerntest (eine Version des Lerntests LAMBDA, s. z.B. Kubinger, Haiden, Karolyi & Maryschka, 2012) bedeutet, dass Personen, die den Test innerhalb der Zeitspanne bis zum automatisch programmierten Abbruch nicht bewältigen konnten, nicht mit einfließen. Eine Nicht-Bewältigung des Tests innerhalb der vorgegebenen Zeit spricht aber für eine schlechte Merkfähigkeit. Berücksichtigt man dies mit, weisen die Daten deskriptiv schon darauf hin, dass eine schlechte Leistung ungünstig für das Erfolgskriterium „Studienerfolg“ zu sein scheint (Tab. 6) – ein gewisses Mindestmaß an Fähigkeiten scheint demnach durchaus vorteilhaft zu sein. Dies geht damit einher mit Ergebissen aus der Literatur, nach denen sich intellektuelle Leistungen als durchaus wichtig für akademischen Erfolg zeigten (Vásquez & de Anglat, 2009), auch wenn die Korrelationen zwischen intellektuellen Fähigkeiten mit Erfolgskriterien oftmals gering ausfallen (z.B. nur $r = 0{,}15$ für das erste Studienjahr und $r = 0{,}11$ für das dritte Studienjahr, insignifikant für das zweite Studienjahr; vgl. Busato, Prins, Elshout & Hamaker, 2000).

		Erfolgs-kriterium	1.	2.	3.	4.	5.
1. Schlussfolgerndes Denken	*r*	.058					
	sig.	.283	-				
	n	99					
2. Benötigte Lernzeit	*r*	-.103	.245*				
	sig.	.208	.039	-			
	n	65	53				
3. Organisations-fähigkeit	*r*	.273*	.211*	-.022			
	sig.	.004	.034	.435	-		
	n	91	76	58			
4. Motivation	*r*	.264*	-.006	-.182	.443*		
	sig.	.006	.478	.084	.000	-	
	n	90	81	59	82		
5. Frustrations-toleranz	*r*	.195*	-.156	-.201	.363*	.457*	
	sig.	.027	.071	.063	.000	.000	-
	n	98	90	59	81	88	
6. Interesse	*r*	.260*	-.057	-.057	.247*	.359*	.223*
	sig.	.010	.317	.345	.019	.001	.026
	n	80	73	52	71	76	76

Tab. 5: Korrelationsmatrix für ausgewählte Prädiktoren und das Erfolgskriterium Studiengeschwindigkeit

Lerntest	**Erfolgskriterium: Durchschnittliche Studiengeschwindigkeit**			**Gesamt**
	Gering	Mittel	Hoch	
Lange Lernzeit	5 (14%)	3 (9%)	5 (17%)	13
Kurze Lernzeit	16 (44%)	17 (49%)	19 (66%)	52
Testabbruch	15 (42%)	15 (43%)	5 (17%)	35
Gesamt	36 (100%)	35 (100%)	29 (100%)	100

Tab. 6: Results in the learning test and academic success

Nichtsdestodtrotz scheinen in den oben erwähnten Ergebnissen insbesondere nicht-kognitive Anforderungen mit dem Erfolgskriterium in Zusammenhang zu stehen – durchaus in Einklang mit der Literatur: In der Studie von Busato et al. (2000) korrelierten nicht-kogntive Prädiktoren wie Leistungsmotivation (r = 0,16 im ersten, r = 0,17 im zweiten und r = 0,12 im dritten Studienjahr) und Gewissenhaftigkeit (r = 0,16 im ersten, r = 0,21 im zweiten und r = 0,18 im dritten Studienjahr) zumindest deskriptiv marginal höher mit akademischem Erfolg als intellektuelle Fähigkeiten. Auch in einer anderen Langzeitstudie zeigten sich Persönlichkeitseigenschaften wie die Big Five (insb. Gewissenhaftigkeit und Introversion) als der intellektuellen Fähigkeiten überlegene Prädiktor für akademische Leistungen (Furnham, Chamorro-Premuzic & McDougall, 2003) – bei der Stichprobe von Furnham et al. (2003) handelt es sich allerdings (noch mehr) um eine bereits hoch selektierte Personengruppe (nämlich Studierende einer britischen Eliteuniversität), sodass hier allgemein von einer hohen Leistungsfähigkeit ausgegangen werden kann.

Davon abgesehen zeigen auch Metaanalysen (Poropat, 2009; Richardson, Abraham & Bond, 2012) die Wichtigkeit von Persönlichkeitseigenschaften für akademischen Erfolg auf und bestätigen diesen, wenngleich der Zusammenhang komplex und von vielen Moderatorvariablen wie z.B. Alter beeinflusst ist (Poropat, 2009).

Von Studienleistung abgesehen scheint es aber auch einen Zusammenhang von nicht-kognitiven Merkmalen mit dem Gefühl, die richtige Studienwahl zu treffen, zu geben: Vergleicht man Studieninteressierte des *Wiener Self-Assessments* für Architektur, die sich in ihrer Studienwahl (eher) sicher sind mit jenen, die es (eher) nicht sind, sind es letztlich nur nicht-kognitive Variablen, die sich in einer Varianzanalyse signifikant zeigen, und von diesen der Großteil der einbezogenen Variablen (Ergebnisse der experimentalpsychologischen Verhaltensdiagnostik wurden aus methodischen Gründen nicht in die Berechnung einbezogen: So ist z.B. im Gegensatz zu anderen Verfahren beim experimentalpsychologisch erfassten Anspruchsniveau sowohl eine zu hohe als auch eine zu niedrige Ausprägung ungünstig). Es zeigten sich signifikante Unterschiede im Arbeitsstil, in der Frustrationstoleranz, in der Motivation, in der Organisation, im Selbstbewusstsein, im Lerninteresse und im Inhaltsinteresse. Die nicht-kognitiven Variablen „Emotionale Komponenten“ und „Soziale Kompetenz“ sind aufgrund der beim multiplen Testen notwendigen Korrektur des Fehlers erster Art (hier durchgeführt mittels Bonferroni-Holm-Korrektur, Holm, 1979; „Emotionale Kompetenz“ müsste als Variable mit dem achtkleinsten *p*-Wert ein *p* von unter 0,05/8 = 0,00625 aufweisen, um signifikant zu werden) nicht mehr signifikant. Tabelle 7 stellt die Deskriptivstatistik sowie die Ergebnisse der einfachen Varianzanalyse dar. Die ersten vier Variablen beziehen sich dabei auf Leistungstests, die folgenden auf Ergebnisse aus Persönlichkeits- bzw. Interessenfragebögen.

Variable	**Gruppe**	***n***	**Mittelwert**	**SD**	**F**	***p***
Lernzeit Lerntest	0	123	293,34	182,088	2,270	0,133
	1	317	266,88	158,395		
Score Matrizentest	0	123	5,20	3,062	0,004	0,947
	1	318	5,22	2,776		
Score Raumvorstellung	0	123	2,48	1,237	3,089	0,080
	1	318	2,25	1,240		
Score Zahlenreihen (Phase 1)	0	123	4,67	1,831	0,109	0,742
	1	317	4,74	1,674		
Arbeitsstil	0	123	5,59	1,746	28,275	0,000
	1	318	6,48	1,513		
Emotionale Komponenten	0	123	5,67	1,491	5,964	0,015
	1	318	6,04	1,407		
Frustrationstoleranz	0	123	4,52	1,363	25,870	0,000
	1	318	5,18	1,160		
Motivation	0	123	11,89	2,970	83,582	0,000
	1	318	14,53	2,629		
Organisation	0	123	9,50	3,074	16,895	0,000
	1	318	10,81	2,944		

Selbstbewusstsein	0	123	3,20	1,664	22,222	0,000
	1	318	4,01	1,581		
Soziale Kompetenz	0	123	7,33	1,474	5,729	0,017
	1	318	7,70	1,499		
Lerninteresse	0	123	12,56	2,829	34,398	0,000
	1	318	14,17	2,475		
Inhaltsinteresse	0	123	11,33	2,982	59,216	0,000
	1	318	13,57	2,650		

Tab. 7: Varianzanalyse für die Gruppe der mit ihrer Studienwahl (eher) Unsicheren (Gruppe 0) bzw. (eher) Sicheren (Gruppe 1)

Zusammengefasst sieht man also, dass Personen, die in ihrer Studienwahl (eher) unsicher sind, gleichzeitig auch niedrigere Interessen, niedrigere Motivation etc. berichten als (eher) sichere Personen. Hinsichtlich Ergebnissen in Leistungstests findet sich jedoch kein Unterschied. Dies könnte bereits einen Risikofaktor für späteren Studienabbruch darstellen, denn auch in einer Längsschnittuntersuchung zum Studienabbruch (Schiefele, Streblow & Brinkmann, 2007) zeigten AbbrecherInnen verglichen mit Weiterstudierenden Unterschiede in nicht-kognitiven Merkmalen wie Motivationswerten; kognitive Variablen wurden allerdings (mit Ausnahme der durchschnittlichen Schulnote, bei der Weiterstudierende ebenfalls etwas bessere Werte erzielten) in der Längsschnittstudie nicht mit berücksichtigt. Für detailliertere Aussagen müsste jedoch auch im Fall des vorliegenden Self-Assessments ebenfalls ein längschnittliches Design verwendet werden.

4. Die Operationalisierung von Anforderungen

Im Rahmen der *Wiener Self-Assessments* für technische Studienrichtungen wurden zahlreiche Anforderungen neu operationalisiert und neue Verfahren entwickelt. Einen Überblick liefern Kubinger, Frebort, Khorramdel und Weitensfelder (2012); am Rande der vorliegenden Dissertation streifen auch die folgenden in Buchartikeln beschriebenen Verfahren die Operationalisierung von Anforderungen zur Studienwahlberatung: Ein Raumvorstellungstest (Weitensfelder, 2012), ein Test zur Erfassung der Spezialfähigkeit zum figuralen Synthetisieren (Weitensfelder & Becker, 2012), eine Leseverständnistest für Technik-Studierende (Weitensfelder & Hofer, 2012) sowie Interessenfragebögen (Weitensfelder, Undeutsch, Khorramdel & Useini, 2012). Beispielhaft seien im Folgenden der neu entwickelte Raumvorstellungstst (als Beispiel für einen Fähigkeitstest) sowie eigens konstruierte Interessenfragebögen (als Beispiel für nicht-kognitive Anforderungen) herausgegriffen. Die Interessenfragebögen (Weitensfelder, subm.a, subm.b) als Beispiel für die Erhebung nicht-kognitiver Anforderungen wurden zudem in eigenen Fachartikeln noch näher beleuchtet.

Da für die Umsetzung in einer psychologisch-diagnostischen Testbatterie allerdings auch das „Gesamtbild" eine Rolle spielt, nämlich dahingehend, dass für die Erstellung einer Testbatterie auch überlegt werden muss, in welcher Reihenfolge die Tests bestmöglich vorzugeben sind, wurde auch diesem Punkt ein eigener Fachartikel gewidmet (Weitensfelder, in press).

4.1 Die Erfassung kognitiver Fähigkeiten am Beispiel eines neu entwickelten Raumvorstellungstests

Mit dem „Test zur Angewandten Raumvorstellung" (TARV; Weitensfelder, 2012, 2013) wurde ein neues Konzept eines Raumvorstellungstests umgesetzt. Dabei stellte sich zunächst die Frage, wie Raumvorstellung denn verstanden werden sollte: Denn obgleich das Konstrukt der Raumvorstellung bereits lange untersucht wird (z.B. El Koussy, 1935, Thurstone, 1938), gibt es bis zum heutigen Tag keine einheitliche Definition der Fähigkeit – es ist sogar umstritten, ob Raumvorstellung eine einzige Fähigkeit ist oder aus mehreren (und wenn, wie vielen) Dimensionen oder Faktoren besteht. Erschwerend für eine Einteilung möglicher Faktoren oder Dimensionen kommt hinzu, dass Raumvorstellung manchmal breiter, manchmal weniger breit verstanden wird, und mit unterschiedlichsten Herangehensweisen untersucht wurde. Ein Überblick über die geläufigsten bisherigen Einteilungen findet sich z.B. bei Hegarty und Waller (2006).

Historisch bedeutsam ist die Einteilung Thurstones (1938), der Raumvorstellung zunächst noch als einheitlichen Faktor ansah, später (1950) jedoch drei Subfaktoren definierte: S_1 als die Fähigkeit, sich ein Objekt von verschiedenen Seiten vorzustellen oder es wiederzuerkennen, S_2 als die Fähigkeit, sich Konfigurationen vorzustellen, bei denen eine Bewegung oder Verschiebung zwischen den Teilen vorherrscht, und S_3 als Fähigkeit, die eigene Person richtig im Raum einordnen zu können. In der Literatur finden sich die Subfaktoren oft auch unter den Bezeichnungen *Spatial Relations*, *Visualization* und *Spatial Orientation* (vgl. z.B. Rost, 1976). Eine Einteilung unter dieser Namensgebung wurde später auch von Lohman (1979) übernommen, wobei Lohmans drei Faktoren inhaltlich nicht

deckungsgleich mit jenen Thurstones sind: Unter *Spatial Relations* (später als *Speeded Rotation* bezeichnet, vgl. Hegarty und Waller, 2006) versteht Lohman, einfache mentale Rotationsaufgaben rasch zu lösen; *Visualization* wird von komplexen Aufgaben ohne Zeitdruck gefordert, während sich *Spatial Orientation* auf die Fähigkeit bezieht, sich ein Objekt von einer anderen Seite vorstellen zu können. Eine Abgrenzung von Rotation zur Orientierung erscheint hier freilich schwierig, wie sogar von Lohman (1979, S. 188) selbst beschrieben: „However, the factor is difficult to measure since tests designed to tap it are often solved by mentally rotating the stimulus rather by reorienting an imagined self". Nicht zuletzt solchen Abgrenzungsproblemen mag es zuzuschreiben sein, dass sich teilweise hohe Korrelationen zwischen diesen finden (z.B. Hegarty & Waller, 2004, berichten von Korrelationen um 0,8-0,81 zwischen dem Einnehmen anderer Perspektiven und mentaler Rotation). Solch hohe Korrelationen legen den Verdacht nahe, dass es möglicherweise gar nicht so viele unterschiedliche räumliche Fähigkeiten gibt, wenngleich manche Autoren (z.B. Hegarty & Waller, 2006) betonen, dass Raumvorstellung aus mehreren, getrennten Fähigkeiten besteht. Gegenteilig gibt es aber auch einige wenige Ergebnisse, die einen einzigen zugrundeliegenden räumlichen Faktor im Sinne einer „General Spatial Ability" nahelegen (Colom, Contreras, Botella & Santacreu, 2002; Colom, Contreras, Shih & Santacreu, 2003).

4.1.1 Konstruktion des TARV

Der Test zur Angewandten Raumvorstellung (TARV, Weitensfelder, 2012, 2013) basiert mit der Erfassung unterschiedlicher Teilbereiche auf einem sehr pragmatischen Ansatz: Zum einen erscheinen die Anforderungen an räumliche Fähigkeiten in der Praxis mannigfaltig und komplex, zum anderen ist aber aus bisherigen Forschungsergebnissen weiterhin unklar, ob es überhaupt unterschiedliche Faktoren oder Dimensionen räumlicher Fähigkeiten gibt (und wenn, wie viele). Daher wurde mit dem TARV versucht, unterschiedliche Anforderungen an räumliche Leistungen in einem einzigen Verfahren zu erfassen, inspiriert von praktischen Überlegungen: So erscheint es für ArchitektInnen beispielsweise notwendig, sich anhand zweidimensionaler Pläne dreidimensionale Gebilde vorstellen zu können (s. z.B. die Umsetzung in einem Item in Abb. 1). Zudem ist es auch notwendig, dass ArchitektInnen sich auf den Plänen zurechtfinden, oder sich vorstellen können, wie Gebäude von einer anderen Seite aussehen etc. (Abb. 2). Dieser Komplexität sollte durch Einbeziehung unterschiedlicher räumlicher Teilbereiche Genüge getan werden. Diese Einbeziehung erfolgte nicht zuletzt unter dem Aspekt der Genderfairness, denn der vielfach zitierete Geschlechtsunterschied in räumlichen Fähigkeiten ist gar nicht in allen Teilaspekten gleich relevant (Linn & Petersen, 1985; Voyer, Voyer & Bryden, 1995), sondern besteht am stärksten im Aspekt der mentalen Rotation, inbesondere im *Mental Rotations Test* (Vandenberg & Kuse, 1978), wo er sich mit einer nach Hedges berechneten Effektgröße von 0,94 doch beachtlich zeigte (Linn & Petersen, 1985). Der Unterschied im *Mental Rotations Test* wurde auch in späteren Studien bestätigt (z.B. Voyer & Hou, 2006), ein Geschlechtsunterschied im mentalen Rotieren ist aber ohne Zeitbeschränkung (Glück & Fabrizii, 2010) oder bei dreidimensionaler Vorgabe (McWilliams, Hamilton & Muncer, 1997; Larson et al., 1999; Parsons et al., 2004; Neubauer, Bergner & Schatz, 2010) nicht mehr nachweisbar. Vor diesem Hintergrund erscheint aus Gründen der Genderfairness notwendig, räumliche Fähigkeiten so zu erfassen, dass diese –

zumindest, sofern in der Praxis auch andere Teilaspekte von Raumvorstellung benötigt werden – eben auch anhand anderer Aspekte als „nur“ mittels Rotationsaufgaben gemessen werden; ebenso erscheint es notwendig, auf eine Zeitkomponente zu verzichten, sofern diese in der Praxis nicht gefordert ist (zur Problematik von Speed- und Powertests siehe zudem Kubinger, 2009).

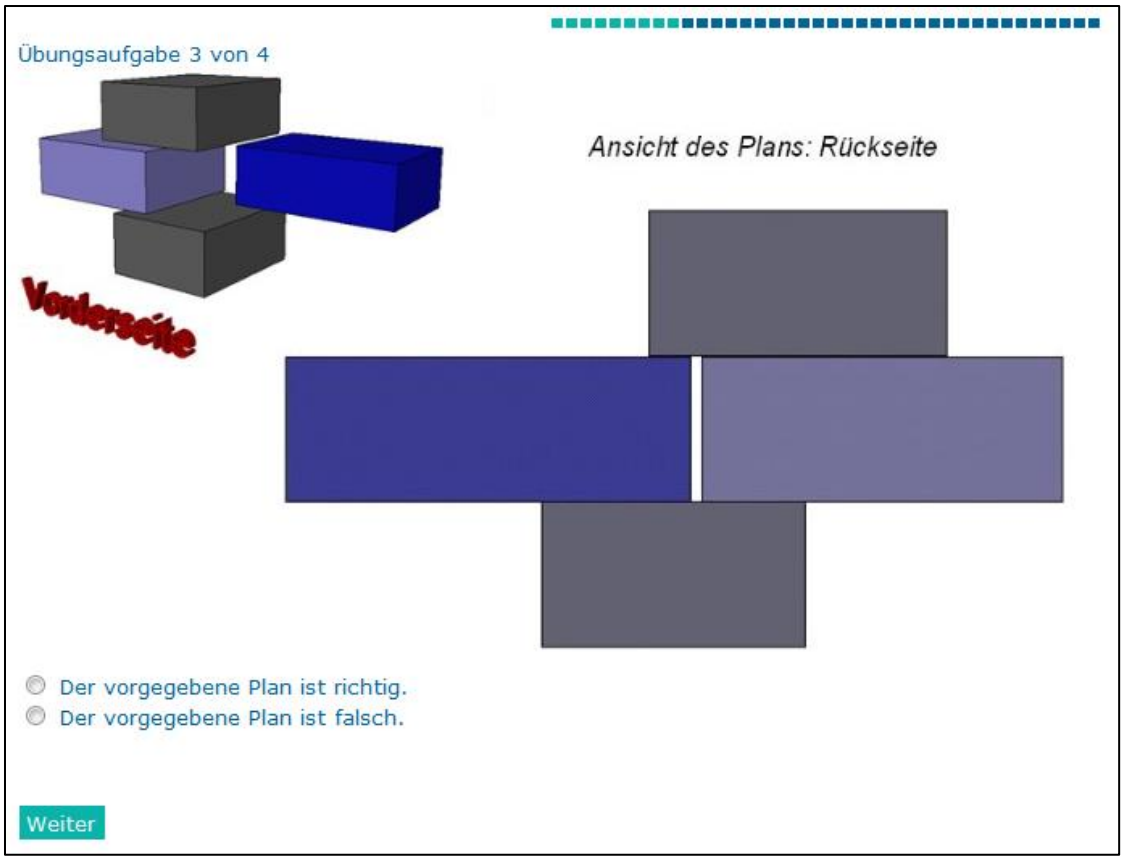


Abb. 1: Ein Instruktionsitem des TARV; der vorgegebene Plan ist falsch, da der Abstand zwischen den mittleren Teilen zu gering dargestellt ist.

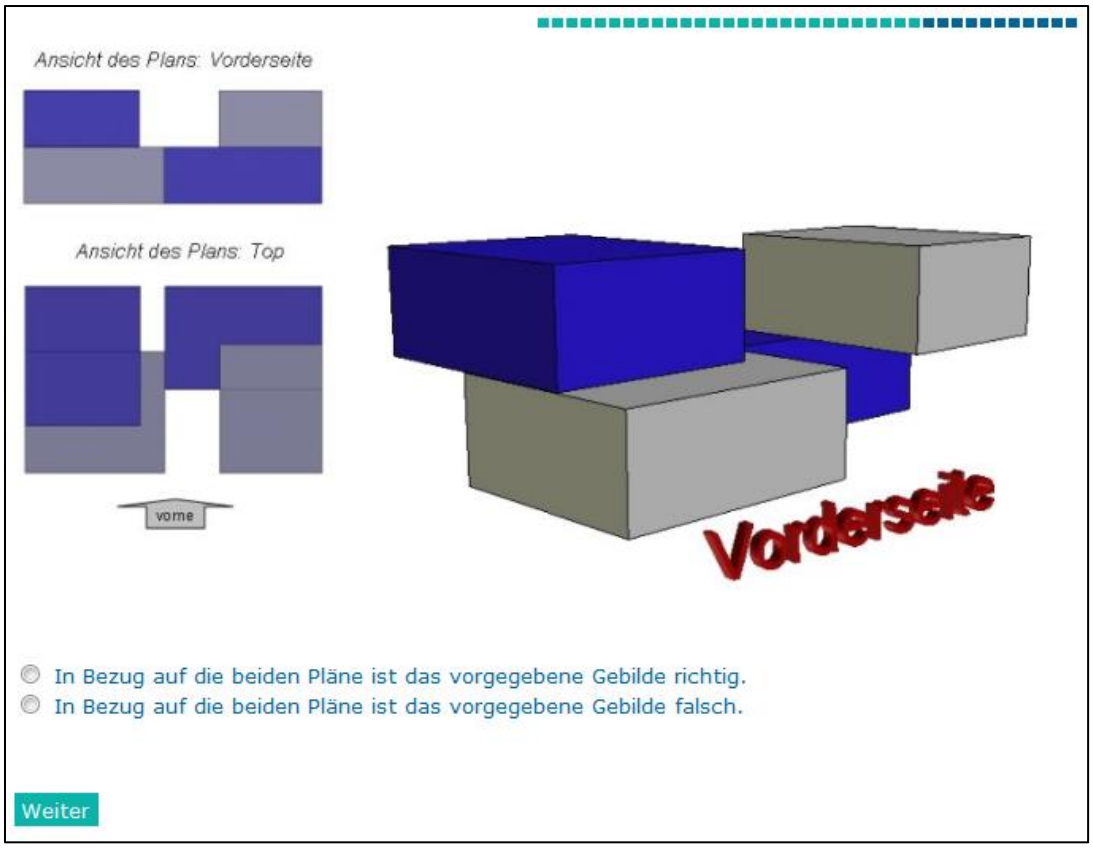


Abb. 2: Ein Instruktionsitem des TARV; das vorgegebene Gebilde ist korrekt.

Abgesehen davon, dass mit dem Test mehrere Aspekte gleichzeitig erfasst werden sollten, und dies möglichst genderfair passieren sollte, besteht eine weitere – und in der Entstehungsgeschichte des Tests nochmals grundlegendere – Idee des Tests (in einem ökologisch validen Sinn) darin, zweidimensionale Planansichten und dreidimensionale Gebilde zu vergleichen und „Fehler" in dem Vergleich zu identifizieren. Neben diesem grundsätzlichen Schließen von zweidimensionalen auf dreimdimensionale Darstellung (Abb. 2) und umgekehrt (Abb. 1) sollen eben unterschiedliche Teilbereiche (im Sinn mentaler Operationen) angesprochen werden. Diese wurden im TARV als „Facetten" bezeichnet, da zum Zeitpunkt der Testkonstruktion (und somit vor ersten Ergebnissen) nicht klar sein konnte, ob es sich dabei um eigenständige Faktoren oder gar Dimensionen handeln würde. Die Facetten wurden in der Bezeichnung von der Einteilung Lohmans (1979) inspiriert, wobei die inhaltliche Umsetzung allerdings nicht genau Lohmans Einteilung entspricht. Stattdessen wurde versucht, die Facetten inhaltlich möglichst klar voneinander abzugrenzen. Die erfassten Facetten im TARV sind:

Facette 1: Relationen – Erkennen von Größen- und Abstandsverhältnissen
Diese Facette zielt (im Gegensatz zu Lohmans „Spatial Relations") auf das ab, was im deutschen Sprachgebrauch unter „Relationen" verstanden wird: Nämlich auf Größen- oder Abstands-Verhältnisse von Dingen zueinander. Entwickelte Testitems zu dieser Facette beziehen sich darauf, dass Testpersonen im Vergleich zu einer Vorgabe Fehler bei Abstands- oder Größenverhältnissen erkennen sollen (s. auch in Abb. 1). Im Alltag könnte man diese Facette mit der Anforderung vergleichen, z.B. beim Auto-Einparken zu entscheiden, ob das Auto in eine Parklücke Platz finden würde oder eben nicht. Während Lohman diesen Aspekt

nicht einbezieht, findet sich in der Literatur zum Abstand-Schätzen Hinweise, dass diese Fähigkeit einen eigenen Faktor darstellen könnte (Carroll, 1993).

Facette 2: Rotation
Mit der Facette „Rotation“ soll die Fähigkeit erfasst werden, mentale Rotationen (ganzer Objekte oder von Objektteilen) vorzunehmen bzw. Rotationsfehler im Vergleich von Plänen und dreidimensional dargestellten Gebilden zu erkennen. Im Alltag wird mentale Rotation benötigt, wann immer man sich Objekte von einer anderen Seite vorstellen muss. Items, die auf diese Facette abzielen, werden im Test beispielsweis öfters von Ansichten dargestellt, die aus der Vorgabe nicht gleich ersichtlich sind, sondern bei denen für den Plan-Objekt-Vergleich erst eine Ansicht in der Vorstellung stark gedreht werden muss (z.B. bei einer Planansicht von der Unterseite). Zudem beinhalten Items auch häufig fehlerhafte Neigungen bzw. Rotationen (fehlerhafte Winkelverhältnisse) von Objektteilen im Vergleich zur Vorgabeansicht, die erkannt werden sollen.

Facette 3: Orientierung
Mit der Facette „Orientierung“ soll schließlich die Fähigkeit erfasst werden, die eigene Person oder Objekte im Raum lokalisieren zu können, besonders in Bezug auf die Richtung. Im Alltag könnte man diese Facette mit der Anforderung vergleichen, mittels Landkarte die richtige Richtung einzuschlagen. Im Test beziehen sich Items zu der Facette oft auf darauf, dass Testpersonen Links-Rechts-Verwechslungen in Bezug auf das Vergleichsobjekt bzw. den Vergleichsplan erkennen können.

4.1.2 Weitere Gestaltungsprinzipien des Tests
Der Test wurde als reiner Powertest für Anforderungen entwickelt, bei denen für räumliche Herausforderungen kein (extremer) Zeitdruck vorherrscht, wie dies beim Lesen oder Erstellen von Plänen in technischen Studienrichtungen der Fall ist. Im Gegensatz dazu wäre z.B. beim Steuern von Flugzeugen neben der Korrektheit auch die Schnelligkeit räumlicher Leistungen bedeutsam. Das Itemmaterial wurde inhaltlich zudem bewusst abwechslungsreich gestaltet, indem sich eckige Formen mit eher runden, anatomisch anmutenden Items abwechseln sowie sowohl abstrakte, technisch anmutende Formen mit konkreten (beispielsweise an Gebäude erinnernde) Objekten als Items beinhaltet sind.
Zudem wurde versucht, die o.g. Facetten so zu erfassen, dass möglichst wenige Überlappungen mit anderen Facetten bestehen (z.B., dass bei einer Orientierungsaufgabe keine Ansichten vorgegeben werden, bei die für einen Vergleich die Pläne stark mental rotiert werden müssen) – eine vollkommene Trennung ist allerdings allein schon aufgrund der perspektivischen Darstellung der Items nicht immer möglich.
In seiner ursprünglichen Variante wird der TARV in einem sequentiellen Antwortformat vorgegeben, bei dem zwei (in einer Pilotversion sogar noch drei) Ansichten korrekt beurteilt werden müssen, um ein Item zu lösen; mittlerweile existiert jedoch auch eine Multiple-Choice-Version des Tests (s. bei Haberstroh, 2012; Weitensfelder, 2012). Es wurden sowohl Items konstruiert, die ein dreidimensional dargestelltes Objekt als Ankerreiz haben (Schließen von 3D auf 2D, vgl. Abb. 1) als auch mit zweidimensionalen Plänen als Ankerreiz (Schließen von 2D auf 3D, vgl. Abb. 2). Ursprünglich waren in beiden Aufgabenarten Items zu jeder

Facette enthalten, in den gekürzten Versionen aus ökonomischen Gründen ist dies nunmehr nicht mehr der Fall. Allerdings zeigten Analysen des Tests, dass der Test über alle Facetten und Aufgabenarten hinweg dem dichotom logistischen Modell nach Rasch (Rasch, 1960) entspricht, somit eindimensional misst (vgl. Haberstroh, 2012; Weitensfelder, 2011, 2012; s. Anhang B für Weitensfelder, 2011): Für die Lösung aller Aufgaben scheint demnach eine einzige Fähigkeit grundlegend zu sein. Damit schließt sich der Test denjenigen Forschungsergebnissen an (z.B. Colom et al., 2002, 2003), die von einer einzigen räumlichen Fähigkeit ausgehen.

4.2 Die Erfassung nicht-kognitiver Anforderungen am Beispiel studienfachspezifischen Interesses

Interessenfragebögen sind in Laufbahnberatungen nicht mehr wegzudenken. Im Ausbildungskontext spielen Interessen über die gesamte Laufbahn eine Rolle: Angefangen bei der Ausbildungswahl (für Studienrichtungen s. z.B. Päßler & Hell, 2012), sind sie beim Lernen von Relevanz (Hidi, 1990; Krapp, 1999) und können letztlich auch bei einem Ausbildungsabbruch eine Rolle spielen (s.o.: Bei einer Befragung für den Grund des Studienabbruchs gaben immerhin 19% der AbbrecherInnen motivationale Gründe an; Heublein et al., 2010).

Überraschenderweise gibt es wenig Belege, dass „passende“ Interessen auch zur Zufriedenheit beitragen. So sind empirische Befunde über Zusammenhänge zwischen Passung (Kongruenz) der Interessen und (Arbeits-)Zufriedenheit oft mehr als bescheiden, wenn überhaupt vorhanden (Dik, Strive & Hansen, 2010; Lent & Lopez, 1996; Tranberg, Slane & Ekeberg, 1993; Wille, Tracey, Feys & De Fruyt, 2014), ein Grund dafür dürfte jedoch methodisch begründet sein (Rolfs & Schuler, 2002): Bei einer durchdachten methodischen Herangehensweise und entsprechend gewählten Kongruenzmaßen konnten Rolfs und Schuler (2002) doch schwache bis mittlere Zusammenhänge zwischen Interessenkongruenz und Wohlbefinden bzw. Bewertung des Studiums finden.

Da für Interesse keine einheitliche Definition existiert, verwundert es nicht, dass sich hinter dem Begriff auch zwei scheinbar unterschiedliche Forschungsströmungen verbergen: Nämlich jene, die Interesse aus einer individuellen (oft auch: „dispositionalen“) Perspektive betrachten, und jene, die Interessen aus einer situativen Perspektive betrachten (Boekarts & Boscolo, 2002). Krapp (2002) argumentierte allerdings, dass situatives Interesse durchaus in individuelles Interesse übergehen kann. Die somit durchaus vereinbaren Ströme sind beide im Ausbildungskontext relevant: Während beim Lernen der Fokus auf siutativem Interesse liegt (z.B. über die Aufmerksamkeitszuwendung, Hidi, 1990), geht es bei der Wahl der richtigen Ausbildungsrichtung vielfach um individuelles bzw. dispositionales Interesse.

Häufig werden dabei Fragebögen eingesetzt, die auf Holland’s (1959, 1997) RIASEC-Modell basieren, in dem sechs unterschiedliche Interessengebiete erfasst werden (R – Realistic, I – Investigative, A – Artistic, S – Social, E – Enterprising, C – Conventional) und die dem Modell seinen Namen geben. Wenngleich das daraus resultierende Interessenprofil für Beratungszwecke als große Errungenschaft gesehen werden kann (vgl. z.B. Sodano, 2015), ist die Bestimmung des Kongruenzmaßes mit dem letztlich angestrebten Berufs- oder Ausbildungsfeld oftmals problematisch (Lent & Lopez, 1996; Rolfs & Schuler, 2002). Im Rahmen der *Wiener Self-Assessments* wurde daher ein gänzlich anderer Ansatz gewählt:

Erfasst werden studienrichtungsspezifische Interessen, maßgeschneidert für die Studienrichtung Maschinenbau bzw. Architektur an der TU Wien, wobei die Items sich an tatsächliche oder zumindest mögliche Inhalte des Studienplans anhalten (s. Weitensfelder, subm.a, subm.b). Die Interessenfragebögen bieten somit auch den Vorteil, dass sie gleich einen konkreten Einblick in die Studienrichtung gewähren (Weitensfelder et al., 2012) und somit eine ökologische, wenn nicht gar „triviale" Validität aufweisen. Im Interessenfragebogen Architektur, nicht jedoch Maschinenbau, wird zudem auch noch das Interesse an Wissensaneignung (für die studiumsspezifischen Inhalte) berechnet, inspiriert vom Modell von Brickenkamp (1990), der neben den Inhaltsklassen auch unterschiedliche Verhaltensklassen von Interesse berücksichtigte: Nämlich rezeptive (die Aufnahme neuer Informationen betreffend), reproduktive (als aktive Auseinandersetzung im Sinne des Nachmachens) und kreative Interessen (bei denen z.B. auch eigene Ideen eingebracht werden). Brickenkamp (1990) geht von einer hierarchischen Entwicklung dieser Verhaltensklassen aus; nach Lehrendenauskunft der Studienrichtung Architektur sei gerade bei Architekturstudierenden allerdings öfter das kreative Interesse vorhanden, während das Interesse an Wissensaneignung eher gering sei. Deshalb wurden im Interessenfragebogen für das Architekturstudium zwei unterschiedliche Interessenarten (nämlich „Inhaltsinteresse", erfragt über reproduktive und kreative Aufgabeninhalte, und Interesse an Wissenseignung/„Lerninteresse", erfragt über Items, in denen der Inhalt z.B. zu lernen sei) hierarchisch gleichwertig nebeneinandergestellt. In Analysen des Interessen-Fragebogens für das Architekturstudium erwies sich das Interesse an Wissensaneignung allerdings hinsichtlich der Geltung des Rasch-Modells als äußerst problematisch (Weitensfelder, subm.b), sodass sogar empfohlen werden kann, diese Skala überhaupt aus dem Test zu kürzen.

Während für detailliertere Beschreibungen und Analysen zu den Interessenfragebögen für Maschinenbau bzw. Architektur auf die Fachartikel verwiesen sei (Weitensfelder, subm.a, subm.b), sei zusammenfassend erwähnt, wie wichtig sich gerade in diesem Kontext die Geltung des dichotom logistischen Modells nach Rasch (1960), kurz „Rasch-Modell", erweist: So kann nämlich beispielhaft gezeigt werden, wie die Fragebögen durch Kürzungen einzelner Items bzw. auch durch Teilung einer Gesamtskala in mehrere Subskalen so adaptiert werden können, dass der rückgemeldete Wert eindimensional ist und somit ein faires Maß darstellt. Adaptionen sind allerdings – wohl wegen der vielseitigen Studieninhalte – von den Erstversionen notwendig, denn gänzlich ohne Anpassungen entsprechen die Fragebögen nicht dem Rasch-Modell; Kürzungen (in Architektur) bzw. Teilungen in Subskalen (in Maschinenbau) erscheinen diesbezüglich jedoch vielversprechend. Selbst für das Teilungskritierum „Geschlecht" finden sich (mit Ausnahme der allgemein kritischen Skala „Lerninteresse") keine oder kaum Items, die dem Modell nicht entsprechen. Dass sich kein Hinweis auf geschlechtsspezifisches „Differential Item Functioning" (DIF) findet, ist allerdings keine Selbstverständlichkeit bei Interessenfragebögen: Studien, die sich mit geschlechtsspezifischem DIF bei allerdings breiter gefassten Interessenfragebögen beschäftigen (z.B. Einarsdóttir & Rounds, 2009; Pässler, Beinicke & Hell, 2014; Wetzel & Hell, 2013) berichten durchaus, dass Frauen und Männer Items trotz der selben zugrundeliegenden Dimension vielfach anders interpretieren; so finden sich beispielsweise in jedem Bereich des *Allgemeinen Interessen-Struktur-Tests AIST-R* (Bergmann & Eder, 2005)

Items mit DIF, am ausgeprägtesten im Bereich der praktisch-technischen Interessen (Wetzel & Hell, 2013).

4.3 Allgemeine Gestaltungsprinzipien: Testreihenfolgen

Während Effekte unterschiedlicher Itemreihenfolgen in Tests durchaus bestätigt sind (z.B. Franke, 1997; Knowles, 1988; Ortner, 2008), gibt es relativ wenige Studien, die sich mit Testreihenfolgen bzw. der optimalen Zusammenstellung von Tests in einer Testbatterie beschäftigen. Dabei gibt es auch hier Hinweise, dass es hinsichtlich ganzer Tests in einer Testbatterie zu Reihenfolgeeffekten kommen könnte (Khorramdel & Frebort, 2011): Bei einer Variation der Testreihenfolge in Testbatterien einer personalauswahl-ähnlichen Situation für Führungspersonen zeigten sich zwar keine Ermüdungseffekte hinsichtlich kognitiver Tests, allerdings Unterschiede in Ergebnissen der experimentalpsychologischen Verhaltensdiagnostik, wo Testpersonen mit längerer Testzeit höhere Entschlussfreudigkeit und geringere Frustrationstoleranz zeigten. Die von Khorramdel & Frebort (2011) gefundenen Testreihenfolgeeffekte konnten allerdings in einer späteren Studie nicht repliziert werden (Schünemann, 2013). Zumindest der Teil, dass sich keine Ermüdungseffekte fanden, mag zunächst überraschen, aber auch eingie andere Studien kommen zu dem Schluss, dass sich eine längere Testdauer oder spätere Vorgabe in einer Testbatterie nur marginal (z.B. Ryan, Glass, Hinds & Brown, 2010; Zhu & Tulsky, 2000) oder gar nicht auf die erbrachte Leistung auswirkt (z.B. Ackerman & Kanfer, 2009; Liu, Allspach, Feigenbaum, Oh & Burton, 2004; Tulsky & Zhu, 2000), selbst Berichte über eine Leistungssteigerung bei längerer Vorgabedauer existieren (Jensen, Berry & Kummer, 2013). Bei längerer Vorgabedauer wird allerdings eine stärkere subjektiv empfundene Ermüdung berichtet (Ackerman & Kanfer, 2009). Dies legt den Schluss nahe, dass es möglicherweise deshalb keine Leistungseinbußen gibt, weil Personen ihre Ermüdung bemerken und mit mehr Anstrengung zu kompensieren versuchen (ebd.) – insbesondere in Testsituationen, in denen die Ergebnisse für die Testpersonen Konsequenzen mit sich bringen, macht diese Erklärung durchaus Sinn. Fraglich ist allerdings, wie dies in Testsituationen aussieht, in denen die Ergebnisse mit kaum oder keinen Konsequenzen für die Testperson behaftet sind – wie beispielsweise in den *Wiener Self-Assessments*.

Daher eröffnet dies die Frage, wie denn Self-Assessments in ihrer Zusammenstellung gestaltet werden sollten, um z.B. Testpersonen nicht zu sehr zu ermüden, oder auch durch ungünstige Testreihenfolgen hohe Abbrecherraten zu provozieren. In Hinblick auf Online Testbatterien (somit auch auf online zugängliche Self-Assessments) existieren Empfehlungen dahingehend, dass (um AbbrecherInnen möglichst zu Beginn abzufangen) mit einer hohen Hürde begonnen werden sollte, und dass es eine „Aufwärmphase“ geben sollte (Reips, 2002). Insbesondere zur Reduktion der ohnehin hohen Dropout-Raten in Self-Assessments wäre daher interessant zu untersuchungen, ob sich eine Veränderung der Testreihenfolge positiv auf diese auswirken könnte. Eine Vorstudie (Weitensfelder, Frebort, Müller & Mitschek, 2011; siehe Anhang B) anhand einer ex-post-facto Situation legte den Schluss nahe, dass Testreihenfolgen möglicherweise auch für Dropouts eine Rolle spielen könnten – experimentell wurde dies später bestätigt (Weitensfelder, in press): Es zeigte sich bei Vorschaltung eines kurzen Erwartungsfragebogens vor der größten Hürde doch eine um 8% geringere Dropout-Rate im Verlauf. Darüber hinaus zeigte sich bei Reihenfolgeänderungen allerdings auch

Leistungseinbußen im Testverlauf des *Wiener Self-Assessments* für Architektur (Weitensfelder, in press): Im Matrizentest wurde bei späterer Vorgabe im Schnitt über ein Item weniger (ca. 0,44 Standardabweichungen, Hedges *g*) gelöst, und auch im Raumvorstellungstest zeigten sich ähnliche Ermüdungseffekte (0,42 Standardabweichungen, Hedges *g*) – womit dieses Ergebnis gewissermaßen einen Kontrapunkt zu bisherigen Forschungsergebnissen darstellt. Eine Erklärung dafür könnte in der *Low-Stakes*-Testsituation begründet sein, das heißt darin, dass mit den Ergebnissen keinerlei (gravierende) Konsequenzen für die Testpersonen verbunden sind, und die von Ackerman und Kanfer (2009) vermutete Mehranstrengung zur Kompensation der Ermüdung hier nicht geleistet wurde. Jedenfalls machen die Ergebnisse deutlich, dass auch die Testreihenfolge in Self-Assessments berücksichtigt werden sollte – und zwar nicht nur, um die Dropout-Raten möglichst gering zu halten, sondern auch aufgrund möglicher Ermüdungseffekte.

5. Diskussion und Ausblick

Die beiden Fachzeitschriftenartikel zum Interesse (Weitensfelder, subm.a, subm.b) liefern Anregungsmöglichkeiten, wie die nicht-kognitive Anforderung Interesse alternativ zu herkömmlichen Interessenfragebögen erfasst werden kann. Die Anlehnung der Items an konkrete Inhalte des Studienplans erscheint besonders vielversprechend, da gleich ein Einblick in den Studienplan und somit eine Informationsfunktion an angehende Studierende gewährleistet ist (vgl. auch Weitensfelder et al., 2012), während die psychometrischen Analysen (Weitensfelder, subm.a, subm.b) zeigen, wie mit derart „praktisch" gewonnenen und aufgrund des Studienplans vielseitigen Iteminhalten verfahren werden kann, damit die Fragebögen dennoch psychometrischen Anforderungen genügen und der rückgemeldete Testwert ein faires Maß darstellt. Allein aufgrund der Itemkonstruktion (mit der Anlehnung an Studieninhalte) kann den Fragebögen „triviale" Validität zugesprochen werden, Studien in Hinblick auf die prognostische Validität sind allerdings noch ausständig; in diesem Zusammenhang wäre es insbesondere interessant, zu untersuchen, welchen Mehrwert derart spezifische Fragebögen im Vergleich zu herkömmlichen Interessenfragebögen tatsächlich aufweisen. Insbesondere im Falle eines Mehrwerts erscheint der Einsatz möglichst spezifischer Fragebögen auch für andere Einsatzgebiete/Studienrichtungen empfehlenswert, besonders, wenn Studieninteressierte ohnehin bereits ihr zukünftiges Studienfeld stark eingrenzen können, sodass sie keinen allgemeinen Überblick über ihre Interessen mehr benötigen. Nichtsdestotrotz erscheint überlegenswert, ob zu Beratungszwecken nicht auch zusätzlich eine Skala mit Items eingeführt werden soll, die sich auf Inhalte bezieht, die im Studium eben *nicht* abgedeckt sind. In Hinblick darauf, dass auch für Interessen ein „Generalfaktor" besteht (Tracey, 2012), der sich nachteilig auf sehr eng gefasste Interessenfragebögen auswirkt, könnte Studieninteressierten damit nämlich besser geraten werden. Auf der anderen Seite könnten keinesfalls auf den Studienplan zutreffende Iteminhalte aber auch falsche Erwartungen bei Studieninteressierten wecken, sodass damit wieder die gleichzeitige Informationsfunktion über das Studium gefährdet wäre, es bedarf hier also einer sorgfältigen Abwägung. In jedem Fall empfiehlt es sich, die Spezifität der Fragenformulierung beizubehalten - denn abgesehen von der Informationsfunktion scheint die spezifische Formulierung noch einen weiteren Vorteil aufzuweisen: Bis auf bei einer allgemein problematischen Skala zeigen sich nämlich trotz des Bezugs zu praktisch-technischen Inhalten keine Hinweise auf geschlechtsspezifisches „Differential Item Functioning" (DIF) – keineswegs eine Selbstverständlichkeit. Dies eröffnet in einem viel breiter gefassten Rahmen die Frage, ob nicht manche geschlechtsspezifischen Unterschiede im Abschneiden von Interessenfragebögen teilweise durch allgemeine Formulierungen zu erklären sind, bei denen Frauen sich ein anderes Bild machen als Männer, und dem durch eine konkrete Ausformulierung entgegengewirkt werden könnte (zu globalen Interessenunterschieden, bei denen insb. auf der Dimension mit Sach- bzw. Personenbezug Geschlechtsunterschiede zeigen, siehe z.B. Su, Rounds & Armstrong, 2009). Hier eröffnet sich in jedem Fall weiterer, auch über Studienwahl und Studieneignung hinausgehender, Forschungsbedarf.

Abgesehen von den Beiträgen zur Erfassung konkreter Anforderungen bezieht sich die vorliegende Dissertation aber mit ihrem dritten vorliegenden Fachzeitschriftenartikel

(Weitensfelder, in press) auch auf eine breiter gefasste Frage allgemeiner Gestaltung. Nachdem bereits eine Vorstudie (Weitensfelder et al., 2011, s. Anhang B) Hinweise lieferte, dass sich die Testreihenfolge auf die Abbruchrate in Self-Assessments auswirken könnte, wurde dies auch in der nachfolgenden Untersuchung bestätigt. So konnte mit Vorschaltung eines Erwartungsfragebogens vor der höchsten Hürde zwar eine bescheidene (8%), aber dennoch nicht unwesentliche Verringerungen in der Abbruchquote erzielt werden. In Übereinstimmung mit den Empfehlungen von Reips (2002) kann daher bekräftigt werden, dass Self-Assessments mit einer kurzen „Aufwärmphase" (hier: einem Erwartungsfragebogen) und sodann gleich mit einer möglichst hohen Hürde begonnen werden sollten. Besonders beachtlich zeigte sich jedoch, dass die Testreihenfolge sich – entgegen Erkenntnissen aus der Literatur – auf die Leistung in nachfolgenden Verfahren auswirkte (Ermüdungseffekte). Eine möglich Erklärung liegt darin, dass eine kompensatorische Anstrengung, die Leistungseinbußen ausgleichen soll, und daher zu starkem subjektivem Ermüdungsgefühl führt (Ackerman & Kanfer, 2009), in Self-Assessments zu Beratungszwecken nicht zu tragen kommt: Letztlich sind ja für die Studieninteressierten damit keine gewichtigen Konsequenzen verbunden. Diese Problematik sollte allerdings weiter untersucht werden, denn letztlich könnten auch in anderen Testsituationen abseits des klassischen *High-Stakes-Testing* Ermüdungseffekte auftreten. Dass es bei *Low-Stakes-Testing*, also Testsituationen, bei denen die Testergebnisse keine persönlichen Konsequenzen für die Testpersonen darstellen, motivationale Probleme geben kann, wurde in der Literatur bereits thematisiert (z.B. Wise & DeMars, 2005). Zumindest in abgeschwächter Form könnte dies auch für Self-Assessments gelten, wenngleich hier Nützlichkeit und Wichtigkeit für die Testpersonen offensichtlich sein sollten – Faktoren, die sich in herkömmlichen *Low-Stakes*-Testsituationen widerum positiv auf die Leistung auswirken (Cole, Bergin & Whittaker, 2008). Aufgrund der Ergebnisse sollte jedenfalls auch der Gestaltung von Testreihenfolgen in Situationen abseits des *High-Stakes-Testing* Beachtung geschenkt werden, denn Reihenfolgeeffekte können nach den vorliegenden Ergebnissen nicht automatisch ausgeschlossen werden.

Als kleine Forschungsanregung sei auch auf ein Ergebnis hingewiesen, das sich am Rande dieser Arbeit herausgestellt hat, nämlich aus der Entwicklungsarbeit für den Raumvorstellungstest TARV: Da sich der Test über alle Facetten und Aufgabenarten hinweg als eindimensional erwiesen hat, wäre es interessant zu untersuchen, ob sich nicht manche in der Literatur gefundenen Raumvorstellungsfaktoren artifiziell durch die unterschiedliche Test- und Itemgestaltung ergeben.

Grundlegender Forschungsbedarf besteht jedoch in einer Längsschnittstudie, in der die unterschiedlichen Anforderungen mit diversen Studienerfolgskriterien sowie mit Studienzufriedenheit in Verbindung gebracht werden. Gerade an Erkenntnissen in Hinblick auf die Kombination von kognitiven und nicht-kognitiven Merkmalen und deren Wechselwirkungen besteht hier noch Bedarf, um Studieninteressierte wie Hochschulen bestmöglich unterstützen zu können.

Zusammenfassend kann die vorliegende Forschungsfrage, welche Eigenschaften für technische Studienrichtungen benötigt werden, und wie diese bestmöglich zu erfassen seien, jedenfalls folgend beantwortet werden: Abgesehen von einigen Spezialfähigkeiten für einzelne Studienrichtungen ergeben sich für technische Studienrichtungen die selben

Anforderungen wie für allgemeine Studierfähigkeit. Zur Erhebung ausgewählter spezifischer Anforderungen wurden neue Testkonzepte (zur Raumvorstellung bzw. studienrichtungsspezifischen Interessen) präsentiert, wobei sich zeigte, dass für eine Operationalisierung in einer Testbatterie auch Testreihenfolgeeffekte beachtet werden sollten.
Die vorliegenden Ergebnisse einschränkend ist zu beachten, dass die Daten der Fachartikel von anonymen Teilnehmenden aus online zugänglichen Self-Assessments gewonnen wurden. Zwar stellen genau diese die Zielpopulation von Self-Assessments dar, es kann jedoch nicht ausgeschlossen werden, dass diese sich in manchen Eigenschaften von denjenigen Studieninteressierten unterscheiden, die keine solche Testbatterie durchlaufen wollen würden. Denn die technologische Präferenz kann durchaus Zusammenhänge mit Persönlichkeitseigenschaften aufweisen: So zeigten beispielsweise Buchanan und Reips (2001), dass Mac-User höher auf der Skala „Offenheit“ scoren als PC-User; und auch eine andere Studie (Ehrenberg, Juckes, White & Walsh, 2008) kommt zu dem Schluss, dass unterschiedliche Arten der Handynutzung mit Persönlichkeitseigenschaften in Verbindung stehen. Eine neuere Studie (Germine et al., 2012) kommt zwar zu dem Schluss, dass auch Online-Stichproben an Freiwilligen eine gute Datenqualität aufweisen können und es keine systematischen Unterschiede zwischen Web- und Laborstichproben gibt; in der Einzelbetrachtung finden sich in der Studie aber doch einige (unsystematische) Unterschiede, sodass aufgrund derartiger Aspekte vor einer vorschnellen Generalisierung auf Gruppen abseits der Zielpopulation abgeraten werden sollte. Eine Untersuchung, inwieweit die Personengruppen, die an Self-Assessments für eine bestimmte Studienrichtung teilnehmen, repräsentativ für Studieninteressierte im Allgemeinen sein können, stellt daher noch eine letzte Forschungsanregung aus der vorliegenden Arbeit dar.

6. Literatur

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology, 15,* 163–181.

Brickenkamp, R. (1990). Die Generelle Interessen-Skala (GIS). Göttingen: Hogrefe.

Bergmann, C., & Eder, F. (2005). *AIST-R. Allgemeiner Interessen-Struktur-Test mit Umwelt-Struktur-Test (UST-R) – Revision*. Göttingen: Beltz Test GmbH.

Boekarts, M. & Boscolo, P. (2002). Interest in learning, learning to be interested. *Learning and Instruction, 12,* 375-382.

Buchanan, T., & Reips, U.-D. (2001). Platform-dependent biases in Online Research: Do Mac users really think different? In K. J. Jonas, P. Breuer, B. Schauenburg & M. Boos (Eds.), *Perspectives on Internet Research: Concepts and Methods.* Verfügbar unter http://www.uni-konstanz.de/iscience/reips/pubs/papers/Buchanan_Reips2001.pdf [09.04.2017]

Busato, V. V., Prins, F. J., Elshout, J. J., & Hamaker, C. (2000). Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Personality and Individual Differences, 29,* 1057–1068.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge: Cambridge University Press.

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33,* 609-624.

Colom, R., Contreras, M.J., Botella, J., & Santacreu, J. (2002). Vehicles of spatial ability. *Personality and Individual Differences, 32,* 903-912.

Colom, R., Contreras, M. J., Shih, P. C., & Santacreau, J. (2003). The Assessment of Spatial Ability with a Single Computerized Test. *European Journal of Psychological Assessment, 19,* 92-100.

Dik, B. J., Strife, S. R., & Hansen, J.-I. C. (2010). The Flip Side of Holland Type Congruence: Incongruence and Job Satisfaction. *The Career Development Quarterly, 58,* 352-358.

DIN Deutsches Institut für Normung e.V. (2002). *Anfoderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen: DIN 33430.* Berlin: Beuth.

Ehrenberg, A., Juckes, S., White, K. M., & Walsh, S. P. (2008). Personality and self-esteem as predictors of young people's technology use. *CyberPsychology & Behavior, 11,* 739-741.

El Koussy, A. A. H. (1935). The Visual Perception of Space. *British Journal of Psychology* (Monographs Supplement), *20.*

Einarsdóttir, S., & Rounds, J. (2009). Gender bias and construct validity in vocational interest measurement: Differential item functioning in the Strong Interest Inventory. *Journal of Vocational Behavior, 74,* 295-307.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327-358.

Franke, G. H. (1997). "The Whole is More than the Sum of its Parts": The Effects of Grouping and Randomizing Items on the Reliability and Validity of Questionnaires. *European Journal of Psychological Assessment, 13,* 67-74.

Furnham, A., Chamorro-Premuzic, T., & McDougall F. (2003). Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance. *Learning and Individual Differences, 14,* 49-66.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review, 19,* 847-857.

Glück, J., & Fabrizii, C. (2010). Gender Differences in the Mental Rotations Test Are Partly Explained by Response Format. *Journal of Individual Differences, 31,* 106-109.

Haberstroh, S. (2012). *Kalibrierung eines Tests zur Angewandten Raumvorstellung unter Berücksichtigung unterschiedlicher Facetten der Raumvorstellung.* Unveröffentlichte Diplomarbeit, Universität Wien.

Hasenberg, S., & Schmidt-Atzert, L. (2014). Internetbasierte Selbsttests zur Studienorientierung. *Beiträge zur Hochschulforschung, 36,* 8-28.

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence, 32,* 175-191.

Hegarty, M., & Waller, D. A. (2006), Individual Differences in Spatial Abilities. In P. Shah & A. Miyake (Hrsg.), *The Cambridge Handbook of Visuospatial Thinking* (S. 121-169). Cambridge: Cambridge University Press.

Hell, B., Trapman, S., & Schuler, H. (2008). Synopse der Hohenheimer Metaanalysen zur Prognostizierbarkeit des Studienerfolgs und Implikationen fuer die Auswahl- und Beratungspraxis. In Schuler, H., Hell, B. (Hrsg.) *Studierendenauswahl und Studienentscheidung* (S. 43-54). Goettingen: Hogrefe.

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 2007/2008.* HIS: Forum Hochschule 2/2010. Hannover: HIS. Verfügbar unter http://www.dzhw.eu/pdf/pub_fh/fh-201002.pdf [27.03.2017].

Heublein, U., Richter, J., Schmelzer, R., & Sommer, D. (2014). *Die Entwicklung der Studienabbruchquoten an den deutschen Hochschulen. Statistische Berechnungen auf der Basis des Absolventenjahrgangs 2012.* Forum Hochschule 4/2014. Hannover: HIS. Verfügbar unter http://www.dzhw.eu/pdf/pub_fh/fh-201404.pdf [02.04.2017].

Heublein, U., & Wolter, A. (2011). Studienabbruch in Deutschland. Definition, Häufigkeit, Ursachen, Maßnahmen. *Zeitschrift für Pädagogik, 57,* 214-236.

Hidi, S. (1990). Interest and Its Contribution as a Mental Resource for Learning. *Review of Educational Research, 60,* 549-571.

Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology, 6,* 35-45.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Holm, S. (1979): A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics, 6,* 65-70.

Hornke, L. F., Wosnitza, M., & Bürger, K. (2013). SelfAssessment: Ideen, Hintergründe, Praxis und Evaluation [SelfAssessment: Ideas, Backgrounds, Practical Experience and Evaluation]. *Wirtschaftspsychologie, 15,* 5-16.

Jensen, J. L., Berry, D. A., & Kummer, T. A. (2013). Investigating the Effects of Exam Length on Performance and Cognitive Fatigue. *PLoS ONE, 8:* e70270. Verfügbar unter http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0070270 [04.03.2016].

Karner, T. (2002). The volunteer effect of answering personality questionnaires. *Psychologische Beiträge, 44,* 42-49.

Khorramdel, L., & Frebort, M. (2011). Context Effects on Test Performance. What About Test Order? *European Journal of Psychological Assessment, 27,* 103-110.

Khorramdel, L., Maurer, M., Frebort, M., & Kubinger, K. D. (2012). Ein Anforderungsprofil als Voraussetzung eines Self-Assessments zur Studienwahlberatung – am Beispiel „Architektur". In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder (Hrsg.) („Wiener Autorenkollektiv Studienberatungstests"), *Self-Assessment: Theorie und Konzepte* (S. 49-62). Lengerich: Pabst.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*, 312-320.

Krapp, A. (1999). Interest, motivation and learning: An educational-psychological perspective. *European Journal of Psychology of Education, 14,* 23-40.

Krapp, A. (2002). Structural and dynamic aspects of interest development: theoretical considerations from an ontogenetic perspective. *Learning and Instruction, 12,* 383-409.

Kubinger, K. D. (2009). *Psychologische Diagnostik – Theorie und Praxis psychologischen Diagnostizierens* (2., überarb. u. erweiterte Aufl.). Göttingen: Hogrefe.

Kubinger, K. D. (2015). Kritische Reflexionen zu Self-Assessments im Rahmen der Studienberatung. *Das Hochschulwesen, 63,* 76-80.

Kubinger, K. D., Frebort, M., & Müller, C. (2012). Self-Assessments im Rahmen der Studienberatung: Möglichkeiten und Grenzen. In Weitensfelder, L. (2012). Test zur Angewandten Raumvorstellung. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder (Hrsg.) („Wiener Autorenkollektiv Studienberatungstests"), *Self-Assessment: Theorie und Konzepte* (S. 9-24). Lengerich: Pabst.

Kubinger, K. D., Frebort, M., Khorramdel, L., & Weitensfelder, L. (Eds.) „Wiener Autorenkollektiv Studienberatungstests" (2012). *Self-Assessment: Theorie und Konzepte.* Lengerich: Pabst.

Kubinger, K. D., Frebort, M., Khorramdel, L., & Weitensfelder, L. (2013). Prinzipien und Verfahren der Self-Assessments vom „Wiener Autorenkollektiv Studienberatungstests". *Wirtschaftspsychologie, 15,* 17-24.

Kubinger, K. D., Haiden, A., Karolyi, M., & Maryschka, C. (2012). Diagnostik des Lernstils: Der Lerntest LAMBDA. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests") (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 71-82). Lengerich: Pabst.

Kubinger, K. D., Weitensfelder, L., Frebort, M., & Sonnleitner, P. (2012). *Non-cognitive vs. cognitive predictors for academic success – opportunities and limits of self-assessments illustrated by the Viennese products.* Paper presented at the AERA [American Educational and Research Association] Annual Meeting, Vancouver, April 13th - April 17th 2012.

Larson, P., Rizzo, A. A., Buckwalter, J. G., Van Rooyen, A., Kratz, K., Neumann, U., Kesselman, C., Thiebaux, M., & Van der Zaag, C. (1999). Gender issues in the use of virtual environments. *Cyberpsychology & Behavior, 2,* 113−123.

Lent, E. B., & Lopez, F. G. (1996). Congruence from Many Angles: Relations of Multiple Congruence Indices to Job Satisfaction among Adult Workers. *Journal of Vocational Behavior, 49,* 24-37.

Linn, M. C., & Petersen, A. C. (1985). Emergence and Characterization of Sex Differences in Spatial Abilities: A Meta-Analysis. *Child Development, 56,* 1479-1498.

Liu, J., Allspach, J. R., Feigenbaum, M., Oh, H.-J., & Burton, N. (2004). *A Study of Fatigue Effects from the New SAT®.* College Board Research Report No. 2004-5. New York: The College Board. Verfügbar unter http://www.ets.org/Media/Research/pdf/RR-04-46.pdf [20.02.2016].

Lohman, D. F. (1979). *Spatial Ability: A Review and Reanalysis of the Correlational Literature.* Technical Report No. 8, Aptitude Research Project, School of Education, Stanford University.

Mayring, P. (2008). *Qualitative Inhaltsanalyse: Grundlagen und Techniken.* Weinheim: Beltz.

Maurer, M. (2009). *Wiener Studieneignungs-Persönlichkeitsinventar (WSP: Formulierung eines Regelkataloges zur Itemgenerierung und dessen Anwendung zur Erstellung eines Itempools zur Erfassung studienrelevanter Eigenschaften.* Unveröffentlichte Diplomarbeit, Universität Wien.

McWilliams, W., Hamilton, C. J., & Muncer, S. J. (1997). On mental rotation in three dimensions. *Perceptual and Motor Skills, 85,* 297−298.

Milbradt, A., Zettler, I., Putz, D., Heukamp, V., & Hornke, L. (2008). *Ziele von Self-Assessments und ihre Bedeutung für Entwicklung und Evaluation*. *Report Psychologie*, *33*, 352-362.

Mitschek, B. (2011). *Ermittlung prototypischer Testkennwerte im Wiener Self-Assessment Maschinenbau anhand erfolgreicher Studierender*. Unveröffentlichte Diplomarbeit: Universität Wien.

Müller, C. E. (2011). *Ermittlung prototypischer Testkennwerte für das Wiener Self-Assessment für Architektur anhand erfolgreich und wenig(er) erfolgreich Studierender.* Unveröffentlichte Diplomarbeit: Universität Wien.

Neubauer, A. C., Bergner, S., & Schatz, M. (2010). Two- vs. three-dimensional presentation of mental rotation tasks: Sex differences and effects of training on performance and brain activation. *Intelligence, 38,* 529-539.

Ortner, T. M. (2008) Effects of changed item order: A cautionary note to practitioners on jumping to computerized adaptive testing for personality assessment. *International Journal of Selection and Assessment, 16,* 249-257.

Parsons, T. D., Larson, P., Kratz, K., Thiebaux, M., Bluestein, B., Buckwalter, J. G., & Rizzo, A. A. (2004). Sex differences in mental rotation and spatial rotation in a virtual environment. *Neuropsychologia, 42*, 555−562.

Pässler, K., Beinicke, A., & Hell, B. (2014). Gender-Related Differential Validity and Differential Prediction in Interest Inventories. *Journal of Career Assessment, 22,* 138-152.

Päßler, K., & Hell, B. (2012). Do Interests and Cognitive Abilities Help Explain College Major Choice Equally Well for Women and Men? *Journal of Career Assessment, 20,* 479-496.

Pietrangeli, S., Sindern, E., & Rudinger, G. (2013). Entscheidung für das Studienfach mit Online-Self-Assessment (OSA). *Wirtschaftspsychologie, 15,* 34-45.

Poropat, A. E. (2009). A Meta-Analysis of the Five-Factor Model of Personality and Academic Performance. *Psychological Bulletin, 135,* 322-338.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute of Education Research. (Expanded Edition, 1980. Chicago: University of Chicago Press.)

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological Correlates of University Students' Academic Performance: A Systematic Review and Meta-Analysis. *Psychological Bulletin, 138,* 353-387.

Reips, U.-D. (2002). Standards for Internet-Based Experimenting. *Experimental Psychology, 49,* 243-256.

Rolfs, H., & Schuler, H. (2002). Berufliche Interessenkongruenz und das Erleben im Studium. *Zeitschrift für Arbeits- und Organisationspsychologie, 46,* 137-149.

Rost, D. (1976). *Der Begabungsfaktor "Raumvorstellung". Theorie und Training.* Unveröffentlichte Dissertation, Universität Hamburg.

Ryan, J. J., Glass, L. A., Hinds, R. M., & Brown, C. N. (2010) Administration Order Effects on the Test of Memory Malingering. *Applied Neuropsychology, 17,* 246-250.

Scheck, T., Becker, F., Weitensfelder, L., & Frebort, M. (2012). Rahmenbedingungen von Self-Assessments zur Studienwahlberatung. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests") (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 25-48). Lengerich: Pabst.

Schiefele, U., Streblow, L., & Brinkmann, J. (2007). Aussteigen oder Durchhalten. Was unterschiedet Studienabbrecher von anderen Studierenden? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 39,* 127-140.

Schünemann, A. L. (2013). *Testreihenfolgeneffekte bei Persönlichkeits- und Leistungsdiagnostik in realen Auswahlsituationen: Ein Experiment in der Personalauswahl*. Unveröffentlichte Diplomarbeit, Universität Wien.

Sodano, S. M. (2015). Meaning, measurement, and assessment of vocational interests for career intervention. In Paul J. Hartung, Mark L. Savickas, Bruce W. Walsh (Eds.), *APA handbook of career intervention, Volume 1: Foundations. APA handbooks in psychology* (pp. 281-301). Washington, DC: American Psychological Association.

Sonnleitner, P., Kubinger, K. D., & Frebort, M. (2009). Das Wiener Self-Assessment Psychologie mit seinen Verfahren der experimentalpsychologischen Verhaltensdiagnostik zur Messung studienfachübergreifender Soft Skills. In G. Rudinger & K. Hörsch (Hrsg.), *Self-Assessment an Hochschulen: Von der Studienfachwahl zur Profilbildung* (S. 63-72). Göttingen: V&R unipress.

Spiel, D., Litzenberger, M., & Haiden, D. (2007). Bildungswissenschaftliche und psychologische Aspekte von Auswahlverfahren. In C. Badelt, W. Wegscheider & H. Wulz (Hrsg.), *Hochsuchulzugang in Österreich* (S. 479-552). Graz: Universitätsverlag.

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and Things, Women and People: A Meta-Analysis of Sex Differences in Interests. *Psychological Bulletin, 135*, 859-884.

Thurstone, L. L. (1938). *Primary mental abilities.* Psychometric Monographs, 1. Chicago: University of Chicago Press.

Thurstone, L. L. (1950). Some primary abilities in visual thinking. *Proceedings of the American Philosophical Society, 94,* 517-521.

Tracey, T. J. G. (2012). Problems with single interest scales: Implications of the general factor. *Journal of Vocational Behavior, 81, 378-384.*

Tranberg, M., Slane, S., & Ekeberg, S. E. (1993). The Relationship between Interest Congruence and Satisfaction: A Metaanalysis. *Journal of Vocational Behavior, 42,* 253-264.

Trapmann, S., Hell, B., Weigand, S., & Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs - eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie, 21*,11-27.

Tulsky, D. S., & Zhu, J. (2000). Could Test Length or Order Affect Scores on Letter Number Sequencing of the WAIS-III and WMS-III? Ruling Out Effects of Fatigue. *The Clinical Neuropsychologist, 14,* 474-478.

Vandenberg, S.G., & Kuse, A. R. (1978). Mental rotations: A group test of three-dimensional spatial visualization. *Perceptual and Motor Skills, 47,* 599–604.

Vásquez, S. M., & de Anglat, H. D. (2009). Academic achievement and formal thought in engineering students. *Electronic Journal of Research in Educational Psychology, 7,* 653–672.

Voyer, D., & Hou, J. (2006). Type of items and the magnitude of gender differences on the Mental Rotations Test. *Canadian Journal of Experimental Psychology, 60,* 91–100.

Voyer, D., Voyer, S., & Bryden, M.P. (1995). Magnitude of sex differences in spatial abilities:Ameta-analysis and consideration of critical variables. *Psychological Bulletin, 117,* 250–270.

Weitensfelder, L. (2011). *Conception of a facet-oriented spatial ability test (TARV – Test of Applied Relations and Visuo-spatial abilities).* Vortrag auf der 11th European Conference on Psychological Assessment, Riga, August 31st - September 3rd 2011.

Weitensfelder, L. (2012). Test zur Angewandten Raumvorstellung. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 181-195). Lengerich: Pabst.

Weitensfelder, L. (2013). *TARV Test zur Angewandten Raumvorstellung.* Mattersburg: PT Verlag.

Weitensfelder, L. (in press). Test Order Effects in an Online Self-Assessment: An Experimental Study. *Psychological Test and Assessment Modeling.*

Weitensfelder, L. (subm.a). *Implementation of a Field-Specific Interest Questionnaire for Mechanical Engineering: Lessons Learned from Psychometric Analyses.* Submitted to: Testing, Psychometrics, Methodology in Applied Psychology.

Weitensfelder, L. (subm.b). *Implementation, Psychometric Analyses and Gender-Fairness of A Field-Specific Interest Questionnaire for Architecture.* Submitted to: Journal of Career Assessment.

Weitensfelder, L., & Becker, F. (2012). Optische Differenzierungs- und Gliederungsfähigkeit: Der Test Figurales Synthetisieren. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 207-217). Lengerich: Pabst.

Weitensfelder, L., Frebort, M., Müller, C. E., & Mitschek, B. (2011). *Processing Characteristics and Reasonableness regarding self-assessments.* Vortrag auf der 11th European Conference on Psychological Assessment, Riga, August 31st - September 3rd 2011.

Weitensfelder, L., & Hofer, S. (2012). Verstehen von Texten: Leseverständnistest für Technik-Studierende. In K.D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 197-205). Lengerich: Pabst.

Weitensfelder, L., Undeutsch, N., Khorramdel, L., & Useini, C. (2012). Intrinsische Studienmotivation: Interesse, Erwartungen und Selbstkonzept eigener Fähigkeiten. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 119-142). Lengerich: Pabst.

Westhoff, K., & Kluck, M.-L. (2003). *Psychologische Gutachten schreiben und beurteilen.* (4., vollständig überarbeitete und erweiterte Auflage). Berlin: Springer.

Wetzel, E., & Hell, B. (2013). Gender-Related Differential Item Functioning in Vocational Interest Measurement. An Analysis of the AIST-R. *Journal of Individual Differences, 34,* 170-183.

Wille, B., Tracey, T. J. G., Feys, F., & De Fruyt, F. (2014). A longitudinal and multi-method examination of interest-occupation congruence within and across time. *Journal of Vocational Behavior, 84,* 59-71.

Wise, S. L., & DeMars, C. E. (2005). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment, 10,* 1-17.

Zhu, J., & Tulsky, D. S. (2000). Co-norming the WAIS-III and WMS-III: Is There a Test-Order Effect on IQ and Memory Scores? *The Clinical Neuropsychologist, 14,* 461-467.

Anhang A:
Fachzeitschriftenbeiträge (Originalia)

1) Weitensfelder, L. (in press). Test Order Effects in an Online Self-Assessment: An Experimental Study. *Psychological Test and Assessment Modeling.* [Version vor noch ausständiger Druckfahnenkorrektur]
2) Weitensfelder, L. (subm.a). *Implementation of a Field-Specific Interest Questionnaire for Mechanical Engineering: Lessons Learned from Psychometric Analyses.* Submitted to: Testing, Psychometrics, Methodology in Applied Psychology.
3) Weitensfelder, L. (subm.b). *Implementation, Psychometric Analyses and Gender-Fairness of A Field-Specific Interest Questionnaire for Architecture.* Submitted to: Journal of Career Assessment.

Test Order Effects in an Online Self-Assessment:

An Experimental Study

Author:

MMag. Lisbeth Weitensfelder

Doctoral Student at the University of Vienna, Faculty of Psychology (Division of Psychological Assessment and Applied Psychometrics), Liebiggasse 5, 1010 Vienna, Austria

lisbeth.weitensfelder@univie.ac.at

Abstract

While computer based online self-assessments become increasingly popular, not much research can be found yet how such test batteries should be designed regarding test order, though effects on motivation and performance can't be ruled out. The given study examines in an experimental setting how test order influences dropout rates and fatigue in a computer based online self-assessment. Test order could at the most reduce overall dropouts to a limited degree – especially a combination of a warming up phase (e.g. a short questionnaire) with a subsequent high hurdle (e.g. a strenuous test) can be recommended. Contrary to studies that focused on high stakes situations, fatigue effects in strenuous tests could be found.

Key words: Context effects; self-assessment; test order; fatigue; dropout rates

Computer based online self-assessments, where testees can assess their abilities, proficiencies, interests or attitudes without supervision of trained test administrators, are becoming increasingly popular in educational contexts: When it comes to provide support for students to select the field of study that best matches their interests and abilities, online self-assessments are already an important assessment tool in German-speaking countries (for examples see Hornke, Wosnitza & Bürger, 2013, or Kubinger, Frebort, Khorramdel & Weitensfelder, 2012, 2013). For the design and development of such online test batteries some recommendations exist with regard to the use of certain tests and instruments: Kubinger (2015) suggests the use of objective personality tests based on experimental-based behavior tasks in addition to personality questionnaires based on self-ratings, he also enunciates against too brief test batteries, which might not meet psychometric quality standards. However, not much research can be found yet on how such online-assessments should be put together regarding the order of the contained tests – this paper aims to close this gap.

Context Effects Regarding Test Order and Test Length

While also the presentation mode of online vs. traditional testing might have an effect, making it possible that online versions measure partly different constructs (e.g. Buchanan, 2001, 2002), the following study focuses solely on online self-assessments. Previous research regarding test order effects or effects of test length usually focuses on “traditional” testing situations. Hereby, the results are contradictory. While for the change of item order, resulting context effects can be seen as proven (e.g. Franke, 1997; Knowles, 1988; Ortner, 2008), results about effects caused by test order are inconsistent. Khorramdel and Frebort (2011) showed at a sample of managers that test order might have effects on the performance of experimental-based behavior tasks, but they found no effects on cognitive ability tests. The authors concluded that test order rather affects simple than complex tasks. However, these results could not be reproduced in a later study, where no test order effects could be found (Schünemann, 2013).

Test order effects might be closely related to test length effects, since with ongoing test time users might experience fatigue. But several studies find no (e.g. Tulsky & Zhu, 2000) or only small (e.g. Zhu & Tulsky, 2000; Ryan, Glass, Hinds & Brown, 2010) fatigue effects on performance scores. It might sound surprising that longer test length does not necessarily lead to fatigue effects, but that goes along with findings of Davis (1946) who found three patterns of performance change with ongoing test time of a 70-minute task: A stable performance (shown by about three-quarters of the sample), an increase of performance (approx. 17%) and a decrease (less than 8%). A study of Ackerman and Kanfer (2009) affirms that performance does not necessarily have to go down with long test length conditions: Testing three versions of the in the U.S. widely used test for scholastic aptitude SAT (with test lengths between 3.5 and 5.5 hours), they found that performance even increased in the longer versions, being dissociated with subjective fatigue which rather increased with longer test length. Similar results could be found in a small sample earlier (Liu, Allspach, Feigenbaum, Oh & Burton, 2004), where prolonging the SAT with an essay did not affect performance negatively. Jensen, Berry and Kummer (2013) even showed that a lengthier version of a biology exam resulted in better performance than a shorter version. However, that does not lead to conclusions for the context of self-assessments. In high stakes tests as the SAT, the results have a powerful implication on the testees' educational pathway, which does not necessarily apply to self-assessments.

Effort in Low Stakes Testing

One possible explanation for the dissociation between subjective fatigue and performance as found by Ackerman and Kanfer (2009) was seen in compensation mechanisms: The authors suggested that people realized their fatigue and therefore put in more effort – an explanation that makes sense in high stakes testing situations. In low stakes testing situations, where test results have no or only a small implication, the situation could be different. Self-assessments for counseling purposes surely can't be classified as high stakes

situations (like entrance exams or the SAT test), but they also differ from "classical" low-stakes testing situations (like PISA) since users undergo the assessment for counseling purposes. Lack of self-relevance has been identified as one of the problems of low-stakes testing (e.g. Finn, 2015, for an overview). Therefore, the wish for appropriate feedback could be a motivator to show some effort in the test battery.

As Wise and DeMars (2005) pointed out, motivation has a big impact on performance in low stakes testing, with more motivated students scoring an average *g* effect size of .59 better than less motivated students. But this might not apply to everyone and every sort of test equally: Barry, Horst, Finney, Brown and Kopp (2010) identified three types of effort in low-stakes testing. Two groups differed in their overall level of effort, but both showed less effort in a cognitive test than in non-cognitive ones (test-taking effort was assessed via a set of items that participants responded to after each contained test). A third group showed moderately high effort over both cognitive and non-cognitive tests. Hints in literature show that such differences could at least partly be caused by relatively stable traits: E.g. Brown and Finney (2011) found that non-responders (representing a group with low motivation) in a low stakes assessment show a different level of reactance than responders. When domain-specific motivational traits (like self-concept) are controlled, the correlation between situation-specific motivation and performance turns out to be smaller, though it still exists (Penk, 2015).

Dropout Rates in Online Surveys and Self-Assessments

While online research offers several benefits like easy access for testees or greater external validity (e.g. Reips, 2002), one of its disadvantages can be seen within high dropout rates (Frick, Bächtiger & Reips, 2001; Reips, 2002). There has been some research on how to reduce those dropout rates in online research (Frick et al., 2001; Göritz & Stieger, 2008): Personal information asked in the beginning can reduce dropout rates, as well as incentives like the possibility of winning a lottery (Frick et al., 2001). However, in the case of self-assessments, incentives are no realistic possibility, and asking personal information in the

beginning could bear the risk of scaring away users who are concerned about their privacy. Therefore, another method seems promising: Putting high hurdles, which means motivational adverse factors, close to the beginning, so that most dropouts happen in the beginning (Reips, 2002). Hurdles can come in diverse forms, such as the need for patience (e.g. via loading times or long texts in the beginning), or technical pretests (Reips, 2002). But as Göritz and Stieger (2008) showed, not all high hurdles might indeed be beneficial: Increasing loading times of webpages to filter out less motivated participants and reach a higher data quality did not only fail to reach the intended effect, but even led to counterproductive results (a lower likelihood of responding to the study).

Reips (2002) suggests a high hurdle in combination with a warm-up phase for internet-based assessment. Seen in the light the theory of sunk costs (Arkes & Blumer, 1985), such a suggested warm-up phase could also have the benefit that testees have already invested time and effort in the test battery, making it harder for them to quit.

Results of a pre-study. In the *Viennese Self-Assessments* (Kubinger et al., 2012, 2013), dropout rates are especially high for two technical fields of study, namely *Architecture* and *Mechanical Engineering*, where they outnumber finishing rates. As shown in a pre-study (Weitensfelder, Frebort, Müller & Mitschek, 2011), most of the dropouts happen at the beginning of the test batteries, but with different gradients: While a majority of dropouts in *Mechanical Engineering* happens during the first test, the dropout process in *Architecture* is more lingering (figure 1). The test batteries are similar, yet include some differences regarding contained tests/items and a different test order, where the position of two strenuous tests is switched: The test battery in *Mechanical Engineering* starts with a (high-hurdle-like) demanding learning test (Kubinger, Haiden, Karolyi & Maryschka, 2012), which is combined with a number series test (Poinstingl, in prep.; see Berndl, Steinfeld & Poinstingl, 2012), while the test battery in Architecture starts with a matrices test (Undeutsch, 2012); the learning test appears much later in the test battery.

[Insert figure 1 about here]

The comparison of the dropout processes leads to the question whether the differences might be caused by test order and in how far a changed test order can alternate dropout rates. Regarding reasonability, as being the amount that a testee's resources are preserved (Testkuratorium, 1986), the dropout process in *Architecture* seems especially disadvantageous for users, where they invest much time working on tests before dropping out to the same amount as testees in *Mechanical Engineering*. The ex-post-facto-situation does not allow a causal interpretation (e.g. there might be a chance of trait differences between the groups), so the previous study examines the effect of test order in an experimental setting. (The formulation of hypotheses refers to the study design and therefore can be found in the method section.)

Method

Design

For an experimental design, test order for the test battery in *Architecture* was varied weekly for a survey period of 18 weeks in 2012. Three different test orders were tested:

Version A: Starting the test battery with a matrices test, placing the demanding learning test (combined with the number series test) in the middle of the test battery, as it happened in the pre-study. This test order offers the benefit of a rather high hurdle (the matrices test) in the beginning, yet leading testees to invest already some time before the highest hurdle, so that they might feel determined to finish the test battery.

Version B: Starting with the test that seemed to be the most strenuous one regarding the dropout rates in the pre-study.

Version C: Starting with a short questionnaire about expectations for the field of study, afterwards following the test order as in version B. This test order should have the benefit of a very high hurdle rather in the beginning, but also having a warm-up-phase up front.

Table 1 shows an overview of the test order in the three different versions including short test descriptions. Please note that the learning test is conjoined with a filling test (the number series test, which also serves as an intermediate task between different recalls) which is listed as a separate test, therefore test order in version B seems to be shifted by one position.

[Insert table 1 about here.]

Hypotheses

Due to the results of the pre-study, it is not assumed that a simple position exchange of the two strenuous tests (versions A vs. B) might be able to reduce overall dropout rates, yet that shall be tested. However, it seems plausible that a warm-up phase (as suggested by Reips, 2002) lowers the dropout rate. Therefore the first hypotheses are:

Hypothesis 1: Test order of version A leads to a significantly lower dropout rate than version B.

Hypothesis 2: A short warm-up phase as in version C can significantly lower the dropout rate compared to version B.

Apart from the consequences on dropout rates, it shall be researched whether test order influences test results. Feedback, especially when being criterion oriented, does not take test order into account, implying that there are no mentionable score differences depending on test order. Therefore test results of the first six tests as in version A should closely be looked at regarding order effects:

Hypothesis 3: There are significant mean differences in test scores between testees from version A and testees from version B.

Hypotheses 1 and 2 are to be tested with simple Pearson chi-square-tests (one-sided). Regarding Hypothesis 3, a multivariate analysis of variances (MANOVA) is conducted for the results of the first test. Additionally, proficiency level is added as a variable in the coding test, even if it is not reported in the feedback for the users: It represents a very simple task and

therefore could show different fatigue levels than complex tasks (Khorramdel & Frebort, 2011). All variables are used in raw scores, alpha is set to 5%.

Sample

In the 18 weeks lasting survey period, 676 users logged in for the test battery (users with prior or multiple logins as well as test users and users, where research purposes were known or assumed, were deleted beforehand). A closer look at the data showed that a considerable amount (namely 100 users) seemed to have paused the test battery for a short or longer period of time (of those who finished the test battery, only the ones with incomplete test parts and the ones who took 4 hours or longer for the test battery were manually controlled). Due to their considerable amount, interrupters were not automatically erased from the sample. They were only erased if their break caused problems like a meanwhile changed test order or technical problems as incomplete tests/test batteries (which appeared to $n = 62$ interrupters and happened especially when the learning test was interrupted). The remaining interrupters were considered as finishers, if they finished within the survey period ($n = 5$) and were considered as dropouts, if they did not finish while the survey period lasted ($n = 33$), no matter whether they still logged in or finished after.

After these adjustments 614 users remain in the sample, including 419 dropouts.

Results

Results Regarding Dropouts

Crosstabs and descriptives for all three versions are shown in table 2. As expected, chi-square-tests do not show significant differences between versions A and B (one-sided Pearson chi-square: $p = .165$). The process (figure 2) seems to show differences though and is comparable to the one found in the pre-study, with a more lingering dropout process in version A. Investigating the effect of a short questionnaire up front, the one-sided hypothesis shows a significant difference between versions B and C (one-sided Pearson chi-square: $p = .041$), meaning that a preceding presentation of a short expectancy questionnaire lowers

overall dropouts. The overall effect with approx. 8% less dropouts after presenting a short preceding questionnaire before the highest hurdle can be seen as modest. It should be noted though, that the effect would remain insignificant, if all disrupters had been expelled from the sample (remaining samples then would be: n = 164 in version B with 116 dropouts and n = 201 in version C with 129 dropouts; one-sided Pearson chi-square: p = .093).

[Insert table 2 about here.]

[Insert figure 2 about here.]

Descriptively, version B has the highest dropout rate and version C the lowest, but only the difference between B and C remains significant (a comparison of A and C is not part of the study design, but shows only insignificant differences; Pearson chi-square = .721, two-sided p = .396). Regarding the dropout process, B and C look comparable (figure 2) with only a small degree of dropouts happening after the highest hurdle.

Results Regarding Performance Effects

To investigate hypothesis 3, a multivariate analysis of variances (MANOVA) was conducted to compare means in test scores between version A and version B. Results of the first six tests of version A served as variables. To ensure that the groups were comparable and that breaks wouldn't level out test order effects, only data from users who completed the test battery continuously was used. The Box's M Test proved to be insignificant (p = .076), so that the resulting F-values of the MANOVA can be interpreted fairly.

Table 3 shows the results of the MANOVA and descriptive statistics. Two variables show significant test order effects: the matrices test score (p = .026) and the test for spatial abilities (p = .034). Testees did score a bit lower in both tests, when they appeared later in the test battery (part. η^2 = .045 and Hedges g = .44 for the matrices test; part. η^2 = .041 and Hedges g = .42 for the test for spatial abilities). The other tests did not show effects on performance.

[Insert table 3 about here.]

Discussion

The results of an experimental variation of test order in an online self-assessment show that test order has none (hypothesis 1) or only a small impact (hypothesis 2) on overall dropout rates. When comparing two different versions that both start with a high hurdle, there is no difference in overall dropout-rates (therefore, hypothesis 1 has to be rejected). However, differences in the dropout process seem visible: When starting with a high, not yet the highest hurdle, more dropouts happen along the ongoing test battery (figure 2). Therefore, regarding reasonability, it seems advisable to put the highest suspected hurdle rather in the beginning of a test battery, so that users who might drop out at a later stage already drop out sooner. A high hurdle rather soon could also be a method to filter out less motivated students; especially for low-stakes assessment, motivation filtering seems advisable anyways (Wise and DeMars, 2005). Of course a very high hurdle as a start might also discourage some users: Descriptively (but not to a significant amount), the version with the highest hurdle in the very beginning has more dropouts than the one that starts with a high, not yet the highest hurdle.

While the position exchange of two strenuous tests did not affect dropout-rates, a short expectancy questionnaire up front did (affirming hypothesis 2). The effect of about 8% less dropouts in version C compared to version B can be seen as small, but satisfying given the background that motivational differences in low-stakes assessment might also be caused by relatively stable traits (e.g. Brown & Finney, 2011). However, it remains unclear whether the different dropout rate between versions B and C is caused by a dropout increase after the highest hurdle in the beginning, a dropout decrease after the short preceding questionnaire or a combination of both. It is possible that putting the highest hurdle in the very beginning (as it happened in version B compared to version A) leads to a dropout increase that is too small to become significant. But even in this case, a possible negative impact of the highest hurdle up front would be outbalanced with a positive impact of the preceding questionnaire. The effect of the questionnaire could be caused either by a general positive effect of a “warming up”

phase, or because of the questionnaire content which (by asking for practical expectations for the aspired field of study) might have appeared especially relevant to the testees and increased their commitment. A third explanation, namely that the lower dropout rate is caused by a sunk cost effect, does not seem plausible given the background that another strenuous and time-consuming test in the beginning (as it happened in version A) did not lower dropout rates for the later coming highest hurdle – even though that represents much higher sunk costs in case of a dropout.

Overall, dropout rates are still very high and test order as investigated in the study can reduce dropouts only to a small amount. It is possible that other methods (e.g. incentives, continuous feedback) might be more effective in reducing dropouts, but not all methods can realistically be implemented in self-assessment test batteries as the one investigated. Test orders however can be changed easily, so if other methods are not available, at least a limited decrease of dropouts can be accomplished.

Apart from an effect on dropout rates, test order led to fatigue effects in a matrices test and a spatial abilities test (hypothesis 3). Testees who had already completed some tests scored .44 (matrices test) and .42 (spatial abilities test) standard deviations lower (Hedges *g*) – a difference that indeed could be relevant. This is inconsistent with previous results, where no (relevant) fatigue effects (e.g. Liu et al., 2004) or even an increase of performance (Jensen et al., 2013) in longer test-taking times were found. It also does not go along with the assumption of Khorramdel and Frebort (2011) that fatigue effects might show rather in simple than complex tasks. However, the test-taking situation in the given study differs from previous studies: In high-stakes testing situations, testees might realize their feelings of fatigue and compensate it with even more effort, which might be one reason for the dissociation between performance and subjective fatigue (Ackerman & Kanfer, 2009). In contrast, testees in self-assessments might not feel urged to put in compensatory effort to counterbalance fatigue effects.

Limitations and Implications for Future Research

As limitation, is has to be noted that the researched population in the present study is very specific, consisting of users who underwent an online counseling for architectural studies. Though not being probable, it cannot be ruled out that testees interested in completely other fields of study might show a different motivation or dropout behavior.

Another limitation lies within the fact that demographic information of dropouts is not available, since demographic information is only asked in the very end of the test battery. So both for the fatigue effects as well as for the effects on dropout rates, it remains open whether the observed effects do apply equally to men and women. Previous results suggest that this might not be the case: Attali, Neeman and Schlosser (2011) found that men show a larger differential performance than women between low and high stakes assessments. The authors assumed that at least a part of the differential performance was due to the fact that men invested less effort in low stakes assessment. Therefore it might also be possible that in the given study the performance differences or the effects on dropouts might not apply equally to men and women. A detailed examination of possible sex differences in the dropout process could be a matter for future studies. Putting a demographic questionnaire up front, it would be interesting to research whether demographic information as warming-up leads to a similar decrease of dropouts like the expectancy questionnaire, and whether a combination of both can decrease dropouts even further.

A major restriction lies within the necessary exclusion of users with technical problems, which might have caused a small bias: Not all tests caused technical difficulties when interrupted. Especially the learning test could not be completed after taking a break, so that exclusions of users with incomplete test batteries consist mainly of users who interrupted the learning test. Users who interrupted other tests did not always have to be excluded and were then added either to the dropouts or to the finishers. However, it would also be possible,

that interrupters represent a separate group regarding their test taking effort. This could also be a matter for future studies.

Regardless of these limitations the present study gives some suggestions on how to design online test batteries. When it comes to dropout rates, the findings confirm suggestions that have already been made by Reips (2002). Additionally, they underline the necessity for further research of test order effects in settings that are not high-stakes: Since the given results in self-assessments do indeed find fatigue effects, such effects should not automatically be ignored when implementing test batteries.

Acknowledgements

All data refers to the "Viennese Self-Assessment" project (Kubinger, Frebort, Khorramdel & Weitensfelder, 2012, 2013).

Special thanks for the good collaboration to the *Technische Universität Wien*, represented by then vice chancellor Prof. Adalbert Prechtl, Ph.D., MSc., Ilona Herbst, MSc., and Franz Reichl, Ph.D.

I also want to show my gratitude to the former staff of the *Centre for Testing and Consulting* of the University of Vienna, especially Prof. Klaus D. Kubinger, Ph.D., MSc., Lale Khorramdel, Ph.D., and Martina Frebort, Ph.D., as well as Sandra Ztrivas, MSc., Fabian Becker, MSc., Thomas Scheck, MSc. and then-intern Leonard Schünemann, MSc.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology, 15,* 163–181.

Arkes, H. R., & Blumer, C. (1985). The Psychology of Sunk Cost. *Organizational Behavior and Human Decision Processes, 35,* 124-140.

Attali, Y., Neeman, Z., & Schlosser, A. (2011). *Rise to the Challenge or Not Give a Damn: Differential Performance in High vs. Low Stakes Tests.* IZA Discussion Paper No. 5693. Institute for the Study of Labor (IZA), Bonn. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1842090 [Feb 20, 2016]

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do Examinees Have Similar Test-Taking Effort? A High-Stakes Question for Low-Stakes Testing. *International Journal of Testing, 10,* 342-363.

Berndl, G., Steinfeld, J., & Poinstingl, H. (2012). Schlussfolgendes Denken numerisch: Der Wiener Zahlenreihentest [Numeric reasoning: The Viennese number series test]. In: K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment:Theorie und Konzepte* (pp. 161-169). Lengerich: Pabst.

Brown, A. R., & Finney, S. J. (2011). Low-Stakes Testing and Psychological Reactance: Using the Hong Psychological Reactance Scale to Better Understand Compliant and Non-Compliant Examinees. *International Journal of Testing, 11*, 248-170.

Buchanan, T. (2001). Online Personality Assessment. In U.-D. Reips & M. Bosnjak (Eds.), Dimensions of Internet science (pp. 57–74). Lengerich: Pabst.

Buchanan, T. (2002). Online Assessment: Desirable or Dangerous? *Professional Psychology: Research and Practice (33),* 148-154.

Jensen, J. L., Berry, D. A., & Kummer, T. A. (2013). Investigating the Effects of Exam Length on Performance and Cognitive Fatigue. *PLoS ONE, 8:* e70270. Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0070270 [March 4, 2016].

Davis, D. R. (1946). The disorganization of behaviour in fatigue. *Journal of Neurology, Neurosurgery and Psychiatry, 9,* 23-29.

Finn, B. (2015). *Measuring Motivation in Low-Stakes Assessments*. ETS Research Reports Series ISSN 2330-8516. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/ets2.12067/epdf [Feb 20, 2016].

Frebort, M., Gleeson, R., & Weitensfelder, L. (2012). Wissenstest zur Erfassung des bereichspezifischen Vorwissens [Knowledge test to assess field-specific pre-knowledge]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment: Theorie und Konzepte* (pp. 83-91). Lengerich: Pabst.

Franke, G. H. (1997). “The Whole is More than the Sum of its Parts”: The Effects of Grouping and Randomizing Items on the Reliability and Validity of Questionnaires. *European Journal of Psychological Assessment, 13,* 67-74.

Frick, A., Bächtiger, M.-T., & Reips, U. (2001). Financial incentives, personal information and dropout in online studies. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet Science* (pp. 209-219). Lengerich: Pabst.

Göritz, A. S., & Stieger, S. (2008). The high-hurdle technique put to the test: Failure to find evidence that increasing loading times enhaces data quality in Web-based studies. *Behavior Research Methods, 40,* 322-327.

Hornke, L. F., Wosnitza, M., & Bürger, K. (2013). SelfAssessment: Ideen, Hintergründe, Praxis und Evaluation [SelfAssessment: Ideas, Backgrounds, Practical Experience and Evaluation]. *Wirtschaftspsychologie, 15,* 5-16.

Khorramdel, L., & Frebort, M. (2011). Context Effects on Test Performance. What About Test Order? *European Journal of Psychological Assessment, 27,* 103-110.

Khorramdel, L., & Maurer, M. (2012). Das Wiener Studieneignungs-Persönlichkeitspotenzial [The Viennese personality potential inventory for study aptitude]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment:Theorie und Konzepte* (pp. 103-118). Lengerich: Pabst.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*, 312-320.

Kubinger, K. D. (2015). Kritische Reflexionen zu Self-Assessments im Rahmen der Studienberatung [Critical Reflection of Self Assessments in the Context of Study Counselling]. *Das Hochschulwesen, 63,* 76-80.

Kubinger, K. D., & Ebenhöh, J. (2011). *Arbeitshaltungen (AHA)* [Work styles], Version 27. Mödling: Schuhfried.

Kubinger, K. D., & Ebenhöh, J. (2012). Experimentalpsychologische Verhaltensdiagnostik des Anspruchsniveaus: Der Unterstest Symbole Kodieren aus den „Arbeitshaltungen“ [Experimental-based behavior assessment of the aspiration level: The symbol-coding subtest of the „Arbeitshaltungen“]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment:Theorie und Konzepte* (pp. 93-101). Lengerich: Pabst.

Kubinger, K. D., Frebort, M., Khorramdel, L., & Weitensfelder, L. (Eds.) „Wiener Autorenkollektiv Studienberatungstests“ (2012). *Self-Assessment: Theorie und Konzepte* [Self-assessment: Theory and concepts]. Lengerich: Pabst.

Kubinger, K.D., Frebort, M., Khorramdel, L. & Weitensfelder, L. (2013). Prinzipien und Verfahren der Self-Assessments vom „Wiener Autorenkollektiv Studienberatungstests“

[Principles and Inventories of the Self-Assessments from the „Viennese Author Collevtive for Study Aptitude Testing"]. *Wirtschaftspsychologie, 15*, 17-24.

Kubinger, K. D., Haiden, A., Karolyi, M., & Maryschka, C. (2012). Diagnostik des Lernstils: Der Lerntest LAMBDA [Assessment of the learning style: The learning test LAMBDA]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests") (Eds.), *Self-Assessment:Theorie und Konzepte* (pp. 71-82). Lengerich: Pabst.

Liu, J., Allspach, J. R., Feigenbaum, M., Oh, H.-J., & Burton, N. (2004). A Study of Fatigue Effects from the New SAT®. College Board Research Report No. 2004-5. New York: The College Board. Retrieved from http://www.ets.org/Media/Research/pdf/RR-04-46.pdf [Feb 20, 2016].

Ortner, T. M. (2008) Effects of changed item order: A cautionary note to practitioners on jumping to computerized adaptive testing for personality assessment. *International Journal of Selection and Assessment, 16*, 249-257.

Penk, C. (2015). *Effekte von Testteilnahmemotivation auf Testleistung im Kontext von Large-Scale-Assessments* [Effects of test participation motivation on test performance in the context of large scale assessments]. Unpublished doctoral thesis, Humboldt-Universität zu Berlin.

Reips, U.-D. (2002). Standards for Internet-Based Experimenting. *Experimental Psychology, 49*, 243-256.

Ryan, J. J., Glass, L. A., Hinds, R. M., & Brown, C. N. (2010) Administration Order Effects on the Test of Memory Malingering. *Applied Neuropsychology, 17*, 246-250.

Schünemann, A. L. (2013). *Testreihenfolgeneffekte bei Persönlichkeits- und Leistungsdiagnostik in realen Auswahlsituationen: Ein Experiment in der Personalauswahl* [Test order effects regarding personality and ability assessment in

genuine selection situations: A personal selection experiment]. Unpublished diploma thesis, University of Vienna.

Testkuratorium (1986). Beschreibung der einzelnen Kriterien für die Testbeurteilung [Description of of particular criteria for test evaluation]. *Diagnostica, 32,* 358-360.

Tulsky, D. S., & Zhu, J. (2000). Could Test Length or Order Affect Scores on Letter Number Sequencing of the WAIS-III and WMS-III? Ruling Out Effects of Fatigue. *The Clinical Neuropsychologist, 14,* 474-478.

Undeutsch, N. (2012). Schlussfolgerndes Denken figural: Der Färbige Matrizentest [Figural reasoning: The colourful matrices test]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment: Theorie und Konzepte* (pp. 143-152). Lengerich: Pabst.

Wagner-Menghin, M. M. (2004). *Der Lexikon-Wissen-Test (LEWITE)* [The lexicon knowledge test]. Mödling: Schuhfried.

Weitensfelder, L. (2012). Test zur Angewandten Raumvorstellung [test for applied spatial abilities]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment: Theorie und Konzepte* (pp. 181-195). Lengerich: Pabst.

Weitensfelder, L., Frebort, M., Müller, C. E., & Mitschek, B. (2011). *Processing Characteristics and Reasonableness regarding self-assessments.* 11th European Conference on Psychological Assessment, Riga, August 31st - September 3rd 2011.

Weitensfelder, L., Undeutsch, N., Khorramdel, L., & Useini, C. (2012). Intrinische Studienmotivation: Interesse, Erwartungen und Selbstkonzept eigener Fähigkeiten [Intrinsic study motivation: interest, expectancies and self-concept of own abilities]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment: Theorie und Konzepte* (pp. 119-142). Lengerich: Pabst Science Publishers.

Wise, S. L., & DeMars, C. E. (2005). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment, 10,* 1-17.

Zhu, J., & Tulsky, D. S. (2000). Co-norming the WAIS-III and WMS-III: Is There a Test-Order Effect on IQ and Memory Scores? *The Clinical Neuropsychologist, 14,* 461-467.

Table 1
Contained tests and short test descriptions

	Version A	Version B	Version C
First	**Matrices Test** (Undeutsch, 2012)	Learning Test	Expectancy Questionnaire
Second	**Coding Test:** An adapted short version of one subtest of the experimental-based behavior task "Work Styles" (Kubinger & Ebenhöh, 2011; see also Kubinger & Ebenhöh, 2012), where symbols have to be coded in a simple task and testees have to estimate how much they will be able to do	Number series	Learning Test
Third	**Spatial Abilities Test:** Comparing 2D-plans with 3D-images for accuracy (Weitensfelder, 2012)	Coding Test	Number series
Fourth	**Interest Questionnaire:** Contains study-specific interest questions (Weitensfelder, Undeutsch, Khorramdel & Useini, 2012), reporting two different interest scores	Spatial Abilities Test	Coding Test
Fifth	**Learning Test:** A short version of the learning test LAMBDA (Kubinger, Haiden, Karolyi & Maryschka, 2012), where testees have to learn a company's organization chart	Interest Questionnaire	Spatial Abilities Test
Sixth	**Number series** (Poinstingl, in prep.; see Berndl, Steinfeld & Poinstingl, 2012), assessing numeric reasoning, also serves as filling test in the learning test. In the second half of the test, a strain condition is added.	Matrices Test	Interest Questionnaire
Seventh	**Personality Questionnaire,** assessing several study-relevant traits (Khorramdel & Maurer, 2012)	Personality Questionnaire	Matrices Test
Eight	**Knowledge Test:** Based on the concept of the LEWITE (Wagner-Menghin, 2004), already existing pre-knowledge in the field of architecture is assessed (see Frebort, Gleeson & Weitensfelder, 2012)	Knowledge Test	Personality Questionnaire
Ninth	**Expectancy Questionnaire,** assessing relevant expectancies and frequent false expectations for the aspired field of study (see Weitensfelder et al., 2012)	Expectancy Questionnaire	Knowledge Test
Tenth	**Questions of Surrounding Conditions**, also for self-reflection of the testees	Questions of Surrounding Conditions	Questions of Surrounding Conditions

Table 2
Dropouts in all three versions

	Version A		**Version B**		**Version C**		**Total**
	%	***n***	**%**	***n***	**%**	***n***	
Dropout	68.3%	157	72.7%	128	64.4%	134	419 (68.2%)
Completed	31.7%	73	27.3%	48	35.6%	74	195 (31.8%)
Total	230 (100%)		176 (100%)		208 (100%)		614 (100%)

Chi-square test for versions A vs. B (*n*=406): Pearson chi-square = .951, *p* (1-sided) = 165: chi-square test for versions B vs. C (*n*=384): Pearson chi-square = 3.033, *p* (1-sided) = .041.

Table 3
Multivariate analysis of variances and descriptive statistics for versions A and B

	MANOVA			**Version A (*n* = 67)**		**Version B (*n* = 44)**	
	F	***p***	**part. η^2**	***M***	***SD***	***M***	***SD***
Matrices test: items solved	5.092	.026	.045	5.537	3.076	4.205	2.993
Coding test: Proficiency level	.003	.958	.000	46.24	11.960	46.11	12.429
Coding test: Aspiration level	.409	.524	.004	-.165	.187	-.189	.203
Spatial abilities test: items solved	4.628	.034	.041	2.42	1.281	1.89	1.262
Interest questionnaire: Learning interest (items approved)	.805	.372	.007	13.90	2.432	14.30	2.075
Interest questionnaire: Topic interest (items approved)	1.122	.292	.010	12.90	2.856	13.48	2.791
Learning test: learning time needed	2.886	.092	.026	280.67	157.944	331.27	146.451
Numerical abilities: Items solved (part 1)	1.497	.224	.014	4.85	1.769	4.43	1.757

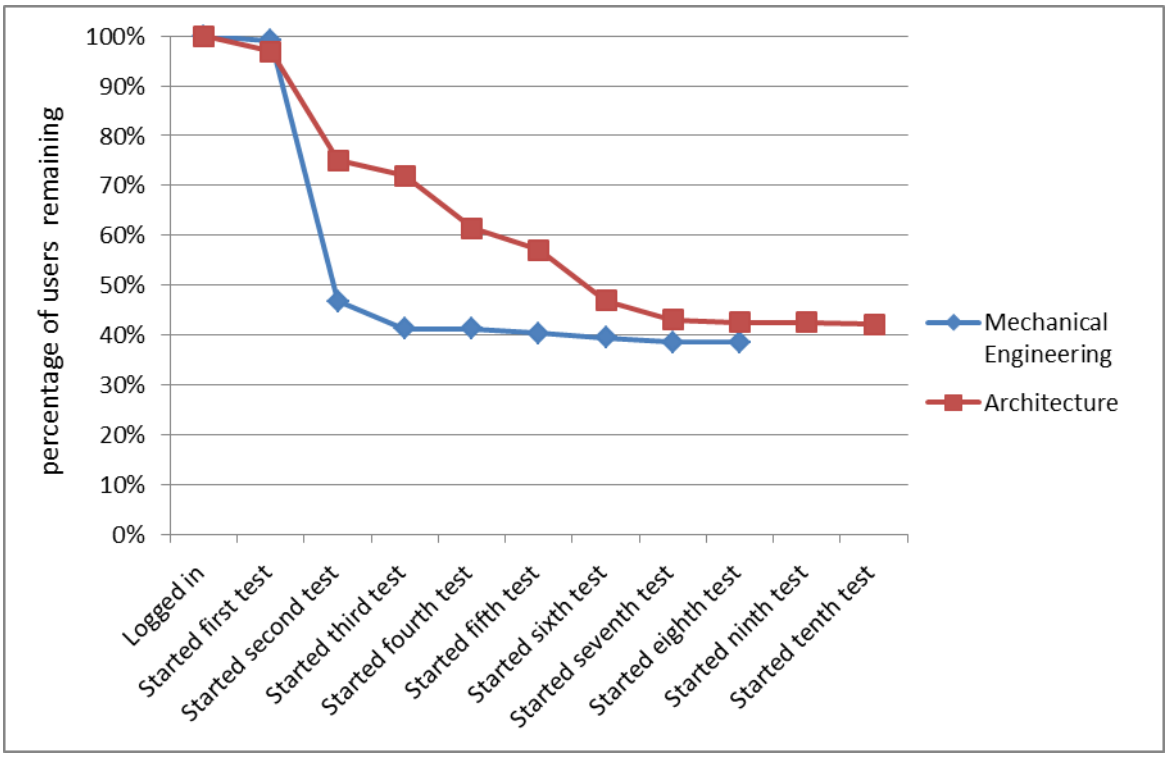


Figure 1. The dropout process of two different test batteries in a pre-study (modified from Weitensfelder, Frebort, Müller & Mitschek, 2011)

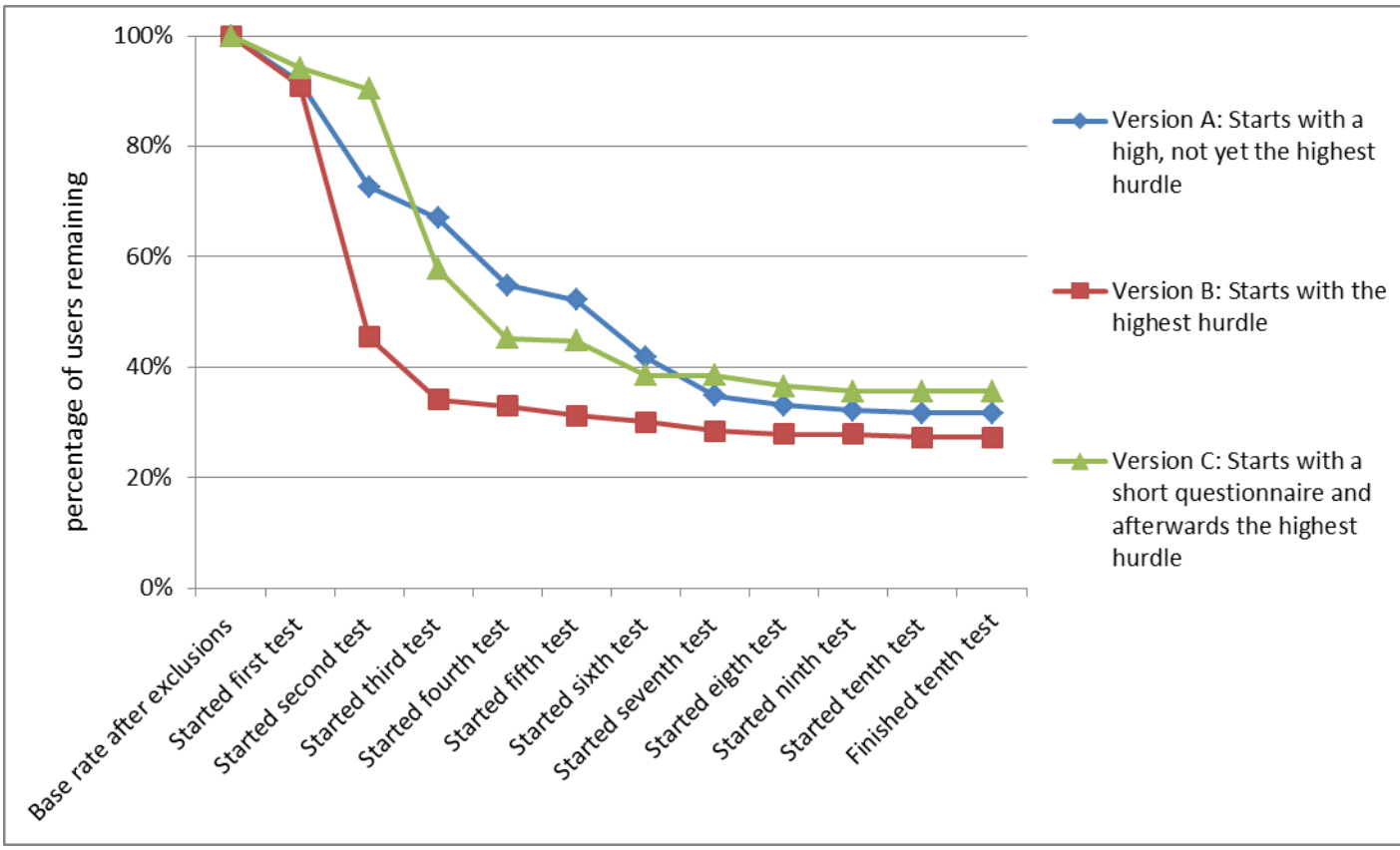


Figure 2. The dropout process in all three experimental versions

Implementation of a Field-Specific Interest Questionnaire for Mechanical Engineering: Lessons Learned from Psychometric Analyses

Author:

MMag. Lisbeth Weitensfelder

Doctoral Student at the University of Vienna, Faculty of Psychology (Division of Psychological Assessment and Applied Psychometrics), Liebiggasse 5, 1010 Vienna, Austria

lisbeth.weitensfelder@univie.ac.at

Acknowledgements

All data refers to the “Viennese Self-Assessment” project (Kubinger, Frebort, Khorramdel & Weitensfelder, 2012, 2013).

Special thanks for the good collaboration to the *Technische Universität TU Wien*, represented by then vice chancellor Univ.-Prof. Dr. Adalbert Prechtl, Mag. Ilona Herbst and Dr. Franz Reichl.

Implementation of a Field-Specific Interest Questionnaire for Mechanical Engineering: Lessons Learned from Psychometric Analyses

Abstract

While interest questionnaires based on Holland's RIASEC-model face the challenge of finding an optimal congruence measure to match a person's interests with the environment, tailor-made specific interest questionnaires avoid this problem in a pragmatic way: Instead of resulting in a general interest profile, only interests for contexts that are needed in the particular field are assessed. The particular field can then be assessed more detailed, while other areas and therefore a typological interest are not looked at. The article presents construction, implementation and psychometric analyses (esp. Rasch model validations) of a field-specific interest questionnaire for *mechanical engineering*, focusing on the struggle of how to create a unidimensional and therefore fair measure.

Key words: field-specific interests; interest assessment; questionnaire design; Rasch model

There is a broad variety of useful and approved interest assessments for career counseling, so that at a first glimpse the invention of new inventories seems unnecessary. But as stated below, there might be cases where other than the already existing inventories are needed, focusing on more narrow aspects of interest. The present paper introduces an interest questionnaire that is tailor-made for a specific field of study (in the following: "field-specific" interest questionnaire) and discusses it from a psychometric point of view, focusing on Rasch model analyses and its benefits.

An uncommon approach to interest questionnaires

Though the application of interest assessments is indisputable in modern career assessments, no consistent definition of interest exists until today. Therefore it's not surprising that different research tendencies exist, at a first sight investigating completely different topics under the label "interest". Applied psychology researches (longer-lasting) vocational interest, while educational psychology focuses on both the longer-lasting topic interest as well as on the situational perspective (e.g. Schraw & Lehman, 2001, for an educational overview). While some authors combine the two educational perspectives pretty well, e.g. by showing that situational interest might lead to longer-lasting topic interest (Hidi & Renninger, 2006; Krapp, 2002; Krapp, 2007), the two different psychological subfields with vocational vs. topic interest usually remain separated from each other. However, when a person wants to study at university, both subfields provide valuable information. Getting educated for a future profession might put university studies somewhere in the transitional field between educational and applied psychology, so that the two disciplines meet for field-specific interest questionnaires.

The Common Vocational Perspective

In general, the benefit of vocational interest questionnaires can be seen as granted, providing a key source for counselors (e.g. Sodano, 2015). Most vocational interest research bases on the empirically well-researched RIASEC-model of Holland (1959, 1997), which has

been a “surpassing achievement in vocational psychology” (Savickas & Gottfredson, 1999). Holland’s theory constitutes six different interest domains, with the initial letters naming the model: Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E) and Conventional (C) interests. Apart from one general factor (Tracey, 2012), the six domains are seen to base on two different factors (e.g. Prediger, 1982), usually called data/ideas and people/things, though other interpretations of underlying factors might also fit the data (e.g. sociability and conformity, see Hogan, 1983). Inventories following Holland’s approach offer the benefit that environments (like fields of study or professions) can also be classified regarding these six domains, making it possible to match people’s dispositions with environments (e.g. for reporting their “fit” with a specific environment or for finding suitable environments).

Limits of vocational interest questionnaires. Interest questionnaires are a useful source for career counseling, usually offering concrete examples for future career paths (e.g. the German *AIST-R*, Bergmann & Eder, 2005) by matching the resulting interest type with environment codes. Nevertheless, when matching a dispositional interest type with an environment, several problems might occur: First, finding an environment code can be problematic and the code might differ depending on the method used (Kaub, Stoll, Biermann, Spinath & Brünken, 2014; Lent & Lopez, 1996; Rolfs & Schuler, 2002). But even if the environment code is already stated, another problem occurs when calculating the congruence of a testee’s dispositional code with the environment code. Usually, the calculation of congruence happens in a typological manner, which means, the comparison between a person’s dispositional code and the environment is done based on the person’s intraindividually dominant interest areas. The absolute value of the testee’s interest is not important – the rank order of the interest areas is the only criterion. Because of this, Rolfs and Schuler (2002) suggested a dimensional conception of congruence, where the absolute distinctions of interest areas are (interindividually) aligned with the environment. Following

this conception, it is important that a person is interested in an area that's requested from the environment, but it's not important whether the person would also be interested in something else. The authors could show at a sample of university students that a dimensional conception is superior to a typological conception in matters of well-being in academic studies. Moreover, the authors even assume that considering irrelevant interest areas might lead to a systematic underestimation of congruence and well-being. Another study of Kaub and her colleagues (2014) lead to similar results, leading them to enunciate against the use of typological congruence indices: Testing teacher candidates, they found low to moderate correlations according to expectations between dimensional interest congruence and various criteria of study motivation and success. However, with a typological determination method, they partly even found correlations *against* their expectations, especially when it came to belief in own abilities.

Another Approach as a Solution

The results above lead to the conclusion that interest is important – but the determination of a "classical" typological interest type might lead to several disadvantages. Since the main difference between dimensional and typological conception lies in the fact that a dimensional conception includes just the interests needed, question arises, whether other interest areas should be determined in all cases. In other words: If known, that for a future profession only a person's value of the interest domains R (Realistic) and I (Investigative) would be important, why assess the other interest areas at all, instead of focusing on the needed domains more closely? And if just the interests needed are assessed, it might not even be important, to which underlying domain the assessed interests would belong to. This is where the present study comes in: It presents an interest questionnaire that just focuses on what is needed.

While a general overview like in Holland's model is needed for vocational orientation, there might be several testees who can already narrow a future profession or field of study

down – especially in those cases it could be more recommendable, economic and reasonable to focus on just the topics needed.

So far, only very few questionnaires exist that focus on more narrow areas of interest (e.g. Bollschweiler & Bernath, 1998; Bollschweiler & Toggweiler, 2009), but there is not much research about narrow assessments. A well-researched (e.g. Glavin, Richard & Porfeli, 2009; Porfeli, Richard & Savickas, 2010; Sodano & Richard, 2009) specific inventory is the *Medical Specialty Preference Inventory MSPI* (Zimny, 1979) resp. its revision *MSPI-R* (Richard, 2011) – a questionnaire which should help American medical students in making the right specialty choice. The *MSPI-R* shows satisfying predictive hit rates in medical specialty choice, while broad interest inventories reached their limits for such a specific question (Burns, 2016).

As a downside for narrow questionnaires, it has to be mentioned that the existence of a general factor for interest can be seen as a disadvantage, since it biases narrow interest-scales (Tracey, 2012). Given this, it might be especially important that narrow questionnaires prove to be valid and helpful for test-takers. But before they finally can be validated, their construction should withstand psychometric requirements.

Construction of Questionnaires According to the Rasch Model

Out of all item response theory (IRT) models, only the dichotomous logistic model of Rasch (1960) offers the benefit of empiric verifiability (Kubinger & Draxler, 2007). In essence the so-called “Rasch model” indicates that the probability of solving an item (in case of a questionnaire: the probability for answering in the affirmative) depends on the ability of the testee and the difficulty of the item. Other parameters (e.g. other personality traits than the aimed one, guessing parameters, learning effects) should not represent systematic influences on the solving probability.

To calculate a single (overall) score without paying attention, which of the included items were answered in the affirmative and which were not, the questionnaire has to be

unidimensional regarding the Rasch model. The model allows testing, whether a questionnaire score is a sufficient statistic. If the Rasch model does not hold, the questionnaire score cannot be seen as a fair measure of a testee's achievement (e.g. Kubinger, 2005). If the Rasch model however is valid, it can be seen as proof that all items base on the same underlying dimension. Furthermore, the Rasch model also allows testing whether "differential item functioning" (DIF) for different groups – e.g. men and women – is an issue. If an item shows DIF, it would mean, that men and women interpret this item differently – even if the underlying interest dimension is the same. Einarsdóttir and Rounds (2009) pointed out that even two thirds of the famous *Strong Interest Inventory SII* in its 1994-version (Harmon, Hansen, Borgen & Hammer, 1994) show gender-specific DIF. DIF could also be found on each RIASEC-scale of the German *AIST-R* (Bergmann & Eder, 2005), most distinctive at the realistic scale (Wetzel & Hell, 2013).

Main Aims of the Study and Study Questions

Field-specific interest questionnaires offer an opportunity to avoid several of disadvantages mentioned above: If a questionnaire just focusses on the topics that might be relevant in an aspired major, a typological calculation is neither possible nor necessary. However, it is not clear how such field-specific interest questionnaires should be classified in the field of given knowledge about interests. Unlike general vocational interest questionnaires, they do not give a broad overview over several domains, but it's probable that a whole field of study consists of more than just one topic (e.g. Sodano & Richard, 2009, found that 18 factors fit the data best for the medicine-specific *MSPI*).

The article presents the construction and implementation of a field-specific interest questionnaire for *mechanical engineering*. While the aims and benefit of the questionnaire can already be deducted from the theoretical background, the subsequent research focuses on the question whether such a tailor-made construction meets psychometric requirements at all. Before presenting the study questions, the questionnaire should be introduced.

Construction of a Field-Specific Interest Questionnaire for *Mechanical Engineering*

The interest questionnaire presented is part of an extensive self-assessment for study applicants at the *Vienna University of Technology* [Technische Universität Wien] (the *Viennese Self-Assessments* for technical majors are accessible at studienwahl.tuwien.ac.at, see also Kubinger, Frebort, Khorramdel & Weitensfelder, 2012, 2013). The self-assessment, where study applicants can test themselves via internet without presence of a test instructor, bases on a specially conducted requirement analysis and includes a broad variety of personality questionnaires, ability tests and experimental-based behavior tasks. Being part of these test batteries, a field-specific interest questionnaire is applied.

The interest questionnaire aims to ask (just) the topics that are relevant for the field of study (bachelor's programme). Furthermore, the questionnaire should have an informative character about the study contents (see Weitensfelder, Undeutsch, Khorramdel & Useini, 2012). The item construction leans on the curriculum of the fields of study and contains actual or at least potential study contents. First, the course catalogue of the curriculum was looked at, formulating questionnaire items about what is done in the courses. After this first step based on the curriculum, the items contained many technical terms and appeared to be abstract and hard to understand, therefore they were revised and reformulated in consultation with advanced *mechanical-engineering*-students (e.g. via asking them how they would explain to a high-school graduate, what is done in the course the item aims at).

The resulting questionnaire consists of 35 questions (example: "I am looking forward to learn, how processes in nature can be abstracted through mathematical models.") at a 4-point dichotomized Likert scale (the answer categories "agree" and "rather agree" both get one point, the answer categories "rather disagree" and "disagree" none). The dichotomization makes the Rasch model applicable, while the 4-point scale should still give some choice to the test-takers.

Considering that it is doubtable whether the interest for a whole field of study is a unitary construct, testees get their overall score reported, but they also get their interest regarding the four widest subjects reported criterion-based either as “strength” or “weakness”. 30 of the 35 questions in the questionnaire base on the four widest subjects in an unequal distribution, five items assess other contents (see tab. 1).

Study Questions

When implementing new questionnaires, numerous aspects can be looked upon. The present article focuses on psychometric aspects, including evaluations regarding IRT, namely the Rasch model: Apart from general fairness of the overall score, questionnaires should furthermore (not at least for sociopolitical reasons) be gender-fair and not include items that disadvantage study applicants with another native language.

Hence the first questions (1., 2.) to be answered in this article are:

1.) Is the calculation of an overall score, as it was assumed in the construction process, justified?

2.) Are the items fair regarding gender (2a) and language (2b)?

Since fairness at item level does not necessarily mean, that there are no differences at a global level, this should also be looked upon with a simple Welch-test. There are empirically well-confirmed differences between men and women in the people-things-dimension (Su, Rounds & Armstrong, 2009), though of course the present population may already be self-selected according to their interest (the sample consists of the target population: testees who want to find out about their fit for studying *mechanical engineering*, see below). That might diminish gender differences, but eventual gender differences will precautionary be looked at.

Methods

Participants

The sample consists of anonymous users of the *Viennese Self-Assessments* for *Mechanical Engineering* who completed the test battery. Of users, who passed the test battery

several times during the inquiry period, only the first results were considered (though a multiple user registration can't be controlled). Known registries for test or research purposes were excluded from the sample.

The resulting sample is a comprehensive survey of these users completing the test battery during a survey period in late 2011 and 2012.

Not being a convenience or laboratory sample, the present sample offers the benefit that it consists of the target population for which the questionnaires was constructed: Namely (future) study applicants and other testees who seek advice for the particular field of study. On the downside, there is no control over test conditions or the sincerity of the assessment process, though the completion of the self-assessment (a test battery of several hours) itself could serve as a sincerity criterion.

Altogether, there is data from 320 people. The test battery is anonymous and demographic information is not asked. Only for a part of the survey period, an additional questionnaire was provided, where demographic data was asked from the testees. The additional questionnaire was administered to 219 testees. 212 of these testees gave demographic information that seemed to be both plausible as well as it was self-reported to be honest. 52 of the 212 testees reported to be women, 31 of the 212 had another mother-language than German; mean age of the sample was 21.45 years (SD: 4.184).

Measures

To answer the two study questions and hypothesis 3, the interest questionnaires described above are the only measures needed.

Procedure

Apart from Rasch model calculations, standard software (SPSS) was used.

Rasch model calculations. To test, whether interest for the study contents can be seen (and reported) as unidimensional, the Rasch model is applied, using the program package *eRm* (Version 0.15-5, Mair, Hatzinger & Maier, 2015; see also Mair & Hatzinger, 2007) for the

statistical software *R* (Version 3.2.2, R Core Team, 2015). As a model test, Andersen's likelihood ratio-test (Andersen, 1973) is applied, testing, whether the same item difficulty parameters are valid for different subgroups: If an assumed item "B" is twice as hard for a subgroup of testees as item "A", this must also be the case for another subgroup of testees. Apart from the common split criterion "score", where testees with higher reported interest are opposed to testees with lower reported interest, the split criteria "gender" and "language" (German vs. non-German) are checked to ensure that the questionnaire is fair for these groups.

While no testees are excluded from the data set for the split criterion "score", for the split criteria "gender" and "language" only those testees remain in the sample, who completed the demographic declarations plausible and who reported to have done so honestly.

Gender differences at scale level. To see, whether men and women differ in their magnitude of reported interests, a Welch test (t test for heterogenous variances) is applied, which is more advisable than a "regular" t test for homogenous variances including precondition tests (Kubinger, Rasch & Moder, 2009). Alpha is set to the common .05.

Results

Classical test theory analyses show a good reliability for the overall score (Cronbach alpha: $r = .828$), but not for the subscales with exception of the biggest one (see tab. 1). Mean comparisons (tab. 2) do not show any difference regarding "gender" or "language", so further examinations will happen on item level.

Rasch model calcuations. To avoid an elevation of the alpha error (e.g. Kubinger & Draxler, 2007), level of significance is corrected (including the split criterion "score", where the major part of the data set overlaps with the data set for the other split criteria). As suggested by Koller, Alexandrowicz & Hatzinger (2012) for testing subgroup invariance, alpha is set to α/q, resulting in a rounded alpha of .017 for an initial α of .05 and three split criteria (= q) in Andersen's likelihood ratio-test. The subsequent Rasch model calculations

show the following: Given the split criterion “language” over all items, the Rasch model is valid. Also regarding the split criterion “gender”, the Rasch model is valid, but the graphical model check (GMC) shows two peculiar items (fig. 1), so that an exclusion of these items could be advisable – there would also be an *ex post* model fit of the Rasch model after exclusion (retested only for the split criterion “gender”). Results of Andersen’s likelihood ratio-tests are shown in table 3.

When it comes to the split criterion “score”, the items are not unidimensional. This might originate from the versatile curriculum. Separate calculations with the comprised subjects show the following (as for all items combined, alpha remains at .017): Due to response patterns (high endorsement with only a few items within every area) only one group (systems science subjects) could be calculated separately - showing unidimensionality regarding the split criterion score ($n = 320$, $p = 0.974$, *Andersen* $\chi^2 = 1.71$, crit. $\chi^2 = 17.062$, $df = 7$). Two more subjects seem to have some commonalities (namely “engineering science subjects” with “construction science and production engineering subjects”), so their items were put together for an artificially combined “technical” subject group. Put together, the Rasch model could be calculated for most items, showing a model fit also for this artificially combined group ($n = 320$, $p = 0.564$, *Andersen* $\chi^2 = 13.491$, crit. $\chi^2 = 28.813$, $df = 15$). The Rasch model would even hold, when “mathematical and natural science subjects” were added to that “technical” subject group ($n = 320$, $p = 0.333$, *Andersen* $\chi^2 = 21.072$, crit. $\chi^2 = 34.286$, $df = 19$). Due to the random ex-post approach, the result has to be interpreted carefully, but it could be a sign that the questionnaire’s non-existent unidimensionality might be caused by the versatile curriculum, where some subjects are more similar than others. Figure 2 shows the GMCs for the split criterion “score” both for all items combined as well as just for systems science subjects separately.

Discussion

Out of the theoretical basis that a certain interest pattern is not as important as a dimensional fit of interests (Kaub et al., 2014; Rolfs & Schuler, 2002), field-specific interest questionnaires offer a pragmatic alternative for specific counseling: When only the topics needed are asked, the problematic (Lent & Lopez, 1996) determination of a congruence index is not necessary any more. Furthermore, tailor-made items can be constructed corresponding to very specific and detailed matters, making it possible to give study applicants already an idea about the curriculum. The item construction of the questionnaire presented bases on the curriculum and was revised with advanced students from that field of study, so the questionnaire can probably be called valid, but of course validation studies (which were not part of the present research) with outside criteria still need to be done and the questionnaire's benefit still needs to be proven. In a validation context, it would be especially interesting to investigate the incremental validity of field-specific questionnaires in combination with general interest questionnaires, since for counselling purposes they could perfectly complement each other: Where one type of questionnaires reaches its limits, the other one starts working.

When focusing on the psychometric criteria of the presented questionnaire, especially on IRT results, the questionnaire shows potential, but also aspects that might need revision: A current disadvantage is that an overall score is reported, though the Rasch model does not hold for the split criterion "score", meaning that the items in the questionnaire represent more than one dimension. An explanation might be that the study contents are just too versatile – the calculations for the subjects seem to support that assumption. However, due to response patterns, only one of the subjects (being unidimensional) could be calculated, and similar-seeming subjects showed a model fit after combining their items to an artificial "technical" subject group. This shows, how Rasch model analyses could help to create ex-post fair (sub-)scores of a questionnaire.

Regarding the validity of the Rasch model for the split criterion “language”, the questionnaire is unidimensional. Also regarding the split criterion “gender”, the items can be seen as gender-fair, since the Rasch model is valid. At an overall level, no gender differences could be found for users who wanted to test their fit for mechanical engineering studies.

Summing up the Rasch model evaluations (and answering study question 1), the calculation of the overall score is not justified. When implementing assessments that should meet both scientific and practical requirements, compromises are often inevitable. However, the findings show how it could be possible to make adjustments via just reporting (aggregated) subscales instead of an overall score to meet psychometric requirements. Of course such a solution (when supported by a later cross-validation) might have negative effects on reliability. Thereby, it needs to be weighted whether the requirement that subscales might represent a certain curriculum order or psychometric requirements (validity of the Rasch model, reliability) are seen as more important. The correspondence of subscales with curriculum contents seems promising for counseling at a first glimpse, but apart from reliability concerns frequent curricula changes or renamings represent a downside of this approach.

Answering study question 2, the questionnaire can be seen as fair regarding both gender (2a) and language (2b).

As a limitation to the previous results, it has to be mentioned that all results base on a preselected sample: Namely the (anonymous) data of (probable future) study applicants for a technical major, who completed an online *Self-Assessment* for their aspired field of study. Though this sample is exactly the population that the questionnaire was constructed for, no conclusions can be made for people who are not interested in the field of study. It is possible that overall gender effects might occur, when the questionnaire is applied to other than that target population.

Several questions still remain open for future examinations: When still focusing on the psychometric point of view, it might be worth questioning, whether the dichotomization of single items - though being the basis for calculating the Rasch model – might also lead to disadvantages. The famous research article of MacCallum, Zhang, Preacher and Rucker (2002), which criticizes the technique of variable dichotomization, does not target on the dichotomization of single items, but one of the authors' major concerns – the loss of information – also applies to item level, when dichotomized. It could be a matter of future examination, whether the questionnaire could maybe benefit from a polytomous data model, like a Partial Credit Model (Masters, 1982). Especially, since the overall score is very screwed, showing a high endorsement (see average scores in tab. 2) , it could be a benefit to discriminate a bit more regarding the four different answer categories rather than also giving testees with "rather agreement" full points for an item.

Concluding, the given findings can give some ideas or guidelines for the construction of field-specific interest questionnaires. The results underline the necessity of Rasch model analyses, since even narrow fields might be too versatile to be unidimensional, causing overall scores not to be a fair measure. In such a case, a fragmentation to subscales as discussed above could be a solution.

Acknowledgements

Please see separate file due to blindness.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123-140. doi:10.1007/BF02291180

Bergmann, C., & Eder, F. (2005). *AIST-R. Allgemeiner Interessen-Struktur-Test mit Umwelt-Struktur-Test (UST-R) – Revision* [GIST-R. General Interest Structure Test with Environment Structure Test (EST-R) – Revision]. Göttingen: Beltz Test GmbH.

Bollschweiler, R., & Bernath, W. (1998). *SIT: Sozial-Interessentest* [SIT: Social Interest Test]. Dübendorf: SVB.

Bollschweiler, R., & Toggweiler, S. (2009). *KV-Interessentest R* [KV Interest Test R] *(2. Ed.).* Bern: SDBB.

Burns, S. T. (2016). History of medical specialty interest assessment. *Cogent education, 3,* 1131062. [Retrieved from: https://www.cogentoa.com/article/10.1080/2331186X.2015.1131062.pdf, Feb. 8th 2017]

Einarsdóttir, S., & Rounds, J. (2009). Gender bias and construct validity in vocational interest measurement: Differential item functioning in the Strong Interest Inventory. *Journal of Vocational Behavior, 74,* 295-307. doi:10.1016/j.jvb.2009.01.003

Glavin, K. W., Richard, G. V., & Porfeli, E. J. (2009). Predictive validity of the medical speciality preference inventory. *Journal of Vocational Behavior, 74,* 128-133. doi:10.1016/j.jvb.2008.11.004

Harmon, L. W., Hansen, J. C., Borgen, F. H., & Hammer, A. L. (1994). *Strong Interest Inventory: Applications and technical guide.* Palo Alto, CA: Consulting Psychologists Press, Inc.

Hidi, S., & Renninger, K. A. (2006). The Four-Phase Model of Interest Development. *Educational Psychologist, 41,* 111-127. doi:10.1207/s15326985ep4102_4

Hogan, R. (1983). A socioanalytic theory of personality. In M. M. Page (Ed.), *Nebraska symposium on motivation 1982. Personality: Current theory and research* (pp. 55–89). Lincoln, NE: University of Nebraska Press.

Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology, 6,* 35-45.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Kaub, K., Stoll, G., Biermann, A., Spinath, F. M., & Brünken, R. (2014). Interessenkongruenz, Belastungserleben und motivationale Orientierung bei Einsteigern im Lehramtsstudium [Congruence, stress experience, and motivational orientation of first-year teacher candidates]. *Zeitschrift für Arbeits- u. Organisationspsychologie, 58,* 125-139. doi:10.1026/0932-4089/a000149

Koller, I., Alexandrowicz, R., & Hatzinger, R. (2012). *Das Rasch Modell in der Praxis: Eine Einfhrung mit eRm* [The Rasch model in practical applications: An introduction with eRm]. Vienna: facultas.wuv, UTB.

Krapp, A. (2002). Structural and dynamic aspects of interest development: theoretical considerations from an ontogenetic perspective. *Learning and Instruction, 12,* 383-409. doi:10.1016/S0959-4752(01)00011-1

Krapp, A. (2007). An educational-psychological conceptualization of interest. *International Journal of Educational and Vocational Guidance, 7,* 5-21. doi:10.1007/s10775-007-9113-9

Kubinger, K. D. (2005). Psychological Test Calibration Using the Rasch Model – Some Critical Suggestions on Traditional Approaches. *International Journal of Testing, 5,* 377-394. doi:10.1207/s15327574ijt0504_3

Kubinger, K. D., & Draxler, C. (2007). Probleme bei der Testkonstruktion nach dem *Rasch*-Modell [Some problems in calibrating an item pool according to the Rasch model]. *Diagnostica, 53,* 131-143. doi:10.1026/0012-1924.53.3.131

Kubinger, K. D., Frebort, M., Khorramdel, L., & Weitensfelder, L. (Eds.) „Wiener Autorenkollektiv Studienberatungstests“ (2012). *Self-Assessment: Theorie und Konzepte* [Self-Assessment: Theory and concepts]. Lengerich: Pabst.

Kubinger, K.D., Frebort, M., Khorramdel, L. & Weitensfelder, L. (2013). Prinzipien und Verfahren der Self-Assessments vom „Wiener Autorenkollektiv Studienberatungstests“ [Principles and Inventories of the Self-Assessments from the „Viennese Author Collective for Study Aptitude Testing”]. *Wirtschaftspsychologie, 15,* 17-24.

Kubinger, K. D., Rasch, D., & Moder, K. (2009). Zur Legende der Voraussetzungen des *t*-Tests für unabhängige Stichproben [About the legend of t-test requirements for independent samples]. *Psychologische Rundschau, 60,* 26-27. doi:10.1026/0033-3042.60.1.26

Lent, E. B., & Lopez, F. G. (1996). Congruence from Many Angles: Relations of Multiple Congruence Indices to Job Satisfaction among Adult Workers. *Journal of Vocational Behavior, 49,* 24-37. doi:10.1006/jvbe.1996.0031

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the Practice of Dichotomization of Quantitative Variables. *Psychological Methods, 7,* 19-40. doi:10.1037//1082-989X.7.1.19

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20,* 1-20. doi:10.18637/jss.v020.i09

Mair, P., Hatzinger, R., & Maier M. J. (2015). *eRm: Extended Rasch Modeling (Version 0.15-5).* Retrieved from http://erm.r-forge.r-project.org/

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika, 47,* 149-74. doi:10.1007/BF02296272

Porfeli, E. J., Richard, G. V., & Savickas, M. L. (2010). Development of specialization scales for the MSPI: A comparison of empirical and inductive strategies. *Journal of Vocational Behavior, 77,* 227-237. doi:10.1016/j.jvb.2010.04.011

Prediger, D. J. (1982). Dimensions underlying Holland's hexagon: Missing link between interests and occupations? *Journal of Vocational Behavior, 21,* 259-287. doi:10.1016/0001-8791(82)90036-7

R Core Team (2015). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute of Education Research. (Expanded Edition, 1980. Chicago: University of Chicago Press.)

Richard, G. (2011). *Medical Speciality Preference Inventory, Revised Edition.* Washington DC: Association of American Medical Colleges.

Rolfs, H., & Schuler, H. (2002). Berufliche Interessenkongruenz und das Erleben im Studium. [Interest congruence and the experience of academic studies]. *Zeitschrift für Arbeits- und Organisationspsychologie, 46,* 137-149. doi:10.1026//0932-4089.46.3.137

Savickas, M. L., & Gottfredson, G. D. (1999): Holland's Theory (1959-1999): 40 Years of Research and Application. *Journal of Vocational Behavior, 55,* 1-4. doi:10.1006/jvbe.1999.1694

Schraw, G., & Lehman, S. (2001). Situational interest: A Review of the Literature and Directions for Future Research. *Educational Psychology Review, 13,* 23-52. doi:10.1023/A:1009004801455

Sodano, S. M. (2015). Meaning, measurement, and assessment of vocational interests for career intervention. In Paul J. Hartung, Mark L. Savickas, Bruce W. Walsh (Eds.), *APA handbook of career intervention, Volume 1: Foundations. APA handbooks in psychology* (pp. 281-301). Washington, DC: American Psychological Association.

Sodano, S. M., & Richard, G. V. (2009). Construct validity of the medical specialty preference inventory: A critical analysis. *Journal of Vocational Behavior, 74,* 30-37. doi:10.1016/j.jvb.2008.10.001

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and Things, Women and People: A Meta-Analysis of Sex Differences in Interests. *Psychological Bulletin, 135,* 859-884. doi:10.1037/a0017364

Tracey, T. J. G. (2012). Problems with single interest scales: Implications of the general factor. *Journal of Vocational Behavior, 81, 378-384.* doi:10.1016/j.jvb.2012.10.001

Weitensfelder, L., Undeutsch, N., Khorramdel, L., & Useini, C. (2012). Intrinsische Studienmotivation: Interesse, Erwartungen und Selbstkonzept eigener Fähigkeiten [Intrinsic study motivation: Interest, expectations and self-concept of own abilities]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment: Theorie und Konzepte* [Self-assessment: Theory and concepts] (pp. 119-142). Lengerich: Pabst.

Wetzel, E., & Hell, B. (2013). Gender-Related Differential Item Functioning in Vocational Interest Measurement. An Analysis of the AIST-R. *Journal of Individual Differences, 34,* 170-183. doi:10.1027/1614-0001/a000112

Zimny, G. H. (1979). *Manual for the medical specialty preference inventory.* St. Louis, MO: St. Louis University School of Medicine.

Table 1
Scales and reliabilities

	Number of items	**Cronbach's Alpha**
Total	**35**	**.828**
Engineering science subjects	12	.708
Systems science subjects	8	.605
Construction science & production engineering subjects	7	.503
Mathematical and natural science subjects	3	.465
Other (diverse and economics)	5	-

Notes: All results based on the complete sample (n = 320). All translations of subjects by the author. Naming and translation based on the curriculum that was valid when the questionnaire was constructed.

Table 2
Welch-tests regarding gender and language differences

Gender					**Language**				
Women		Men			German		non-German		
M	*SD*	*M*	*SD*	*p*	*M*	*SD*	*M*	*SD*	*p*
28.60	5.948	28.17	5.015	.321	28.39	5.254	27.61	5.245	.226

Notes: All p-values one-sided.
n = 212: 52 women; 181 German-speaking

Table 3
Andersen's likelihood ratio-tests

Split Criterion	***n***	**Andersen χ^2**	***df***	$\chi^2_{\alpha=.017}$	***p***
Score	320	124.372	33	52.471	< 0.001
Language	212	42.918	34	53.731	0.14
Sex	212	51.561	34	53.731	0.027
Sex after item exclusion[a]	212	29.139	32	51.206	0.612

Note: [a]Item exclusion advisable due to the graphical model check

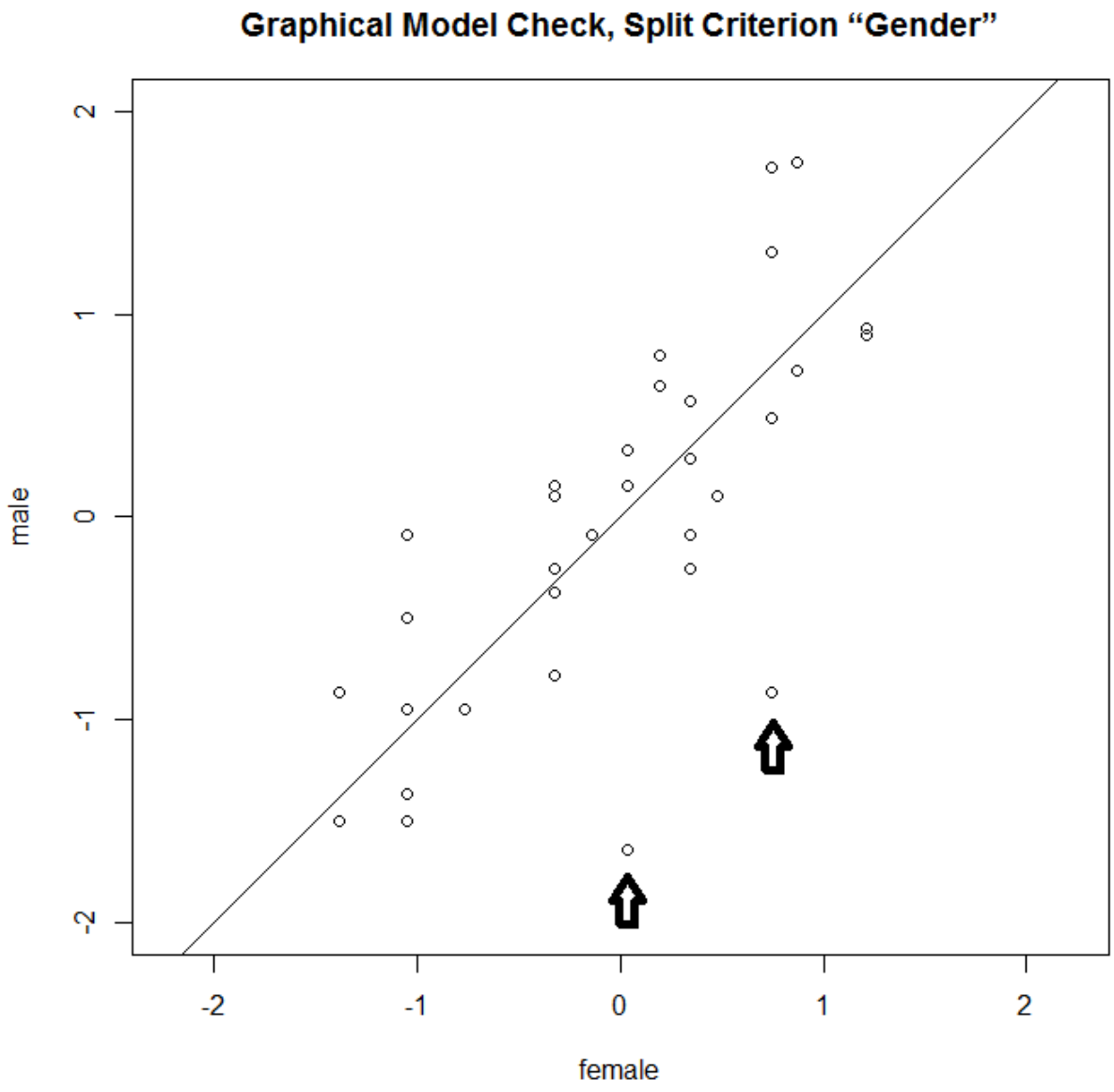


Figure 1. GMC for the split criterion "gender" with two graphically peculiar items (arrows)

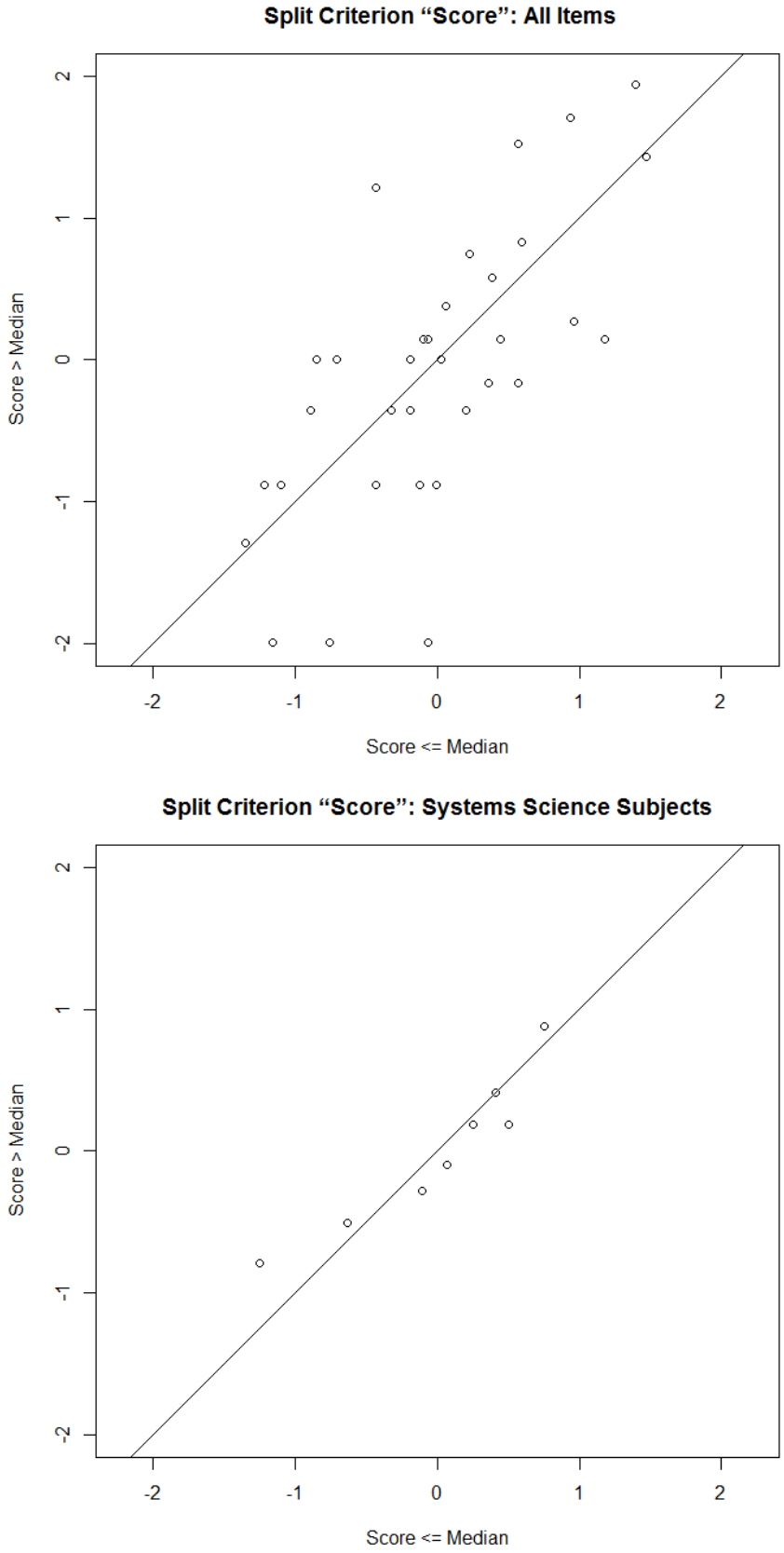


Figure 2. GMCs (split criterion "score") for all subjects combined (above) and "systems science subjects" separately

Implementation, Psychometric Analyses and Gender-Fairness of A Field-Specific Interest Questionnaire for Architecture

Author:

MMag. Lisbeth Weitensfelder

Doctoral Student at the University of Vienna, Faculty of Psychology (Division of Psychological Assessment and Applied Psychometrics), Liebiggasse 5, 1010 Vienna, Austria

lisbeth.weitensfelder@univie.ac.at

Acknowledgements

All data refers to the "Viennese Self-Assessment" project (e.g. Kubinger, Frebort, Khorramdel & Weitensfelder, 2012).

Special thanks for the good collaboration to the *Technische Universität Wien*, represented by then vice chancellor Univ.-Prof. Dr. Adalbert Prechtl, Mag. Ilona Herbst and Dr. Franz Reichl.

Implementation, Psychometric Analyses and Gender-Fairness of A Field-Specific Interest Questionnaire for Architecture

Abstract

When interests for future studies or professions are already known or can be narrowed down, field-specific interest questionnaires might be more appropriate for counseling than common general interest questionnaires. The present study presents implementation and psychometric evaluation of a field-specific interest questionnaire for architectural studies, focusing on Rasch model analyses and gender-fairness. Though the questionnaire might need some revisions, one of the two evaluated questionnaire scales seems to fit well into a unidimensional model. The results of the study raise the question whether more specific interest questionnaires show more gender-fairness than common ones.

Key words: Interest Assessment; Rasch Model; Test Construction; Gender-Fairness; Differential Item Functioning

When it comes to interest questionnaires, most base on the well-researched RIASEC-model of Holland (1959, 1997), consisting of six different interest areas (R - Realistic, I - Investigative, A - Artistic, S - Social, E - Enterprising and C - Conventional) or at least provide a general overview over different interest areas or professions (e.g. the German *Generelle Interessen-Skala GIS*, Brickenkamp, 1990). Specific interest questionnaires as an alternative are much less common, though there might be cases where more specific questionnaires could offer more of the information a testee needs by being able to report more details in an already intended field (e.g. if a testtaker wants to study in a technical field of study, but does not know whether a specific technical major would fit). In vocational psychology, specific interest questionnaires are rarely dealt with, while in educational psychology, several of the research actually focuses on interest from a rather specific point of view.

Topic Interests from the Educational Perspective

While in vocational psychology, interest is often dealt with from a broad perspective, several researchers in educational psychology (e.g. Schiefele & Krapp, 1996) focus on a rather narrow aspect of interest: Namely topic interest as being "a relatively enduring evaluative orientation towards certain topics" (Schiefele & Krapp, 1996, p. 141). Educational psychology has shown the importance of interest for learning: Interest does not only influence the process of learning (Hidi, 1990), but it also increases engagement (Flowerday & Shell, 2015). Topic interest influences learning outcome (mean correlation with grades: $r = .31$, $SD_r = .133$; see Schiefele, Krapp & Winteler, 1992), enhancing both quantity and quality of learning (Schiefele & Krapp, 1996). For students at university, it could be shown that interest is connected to the use of learning strategies (Schiefele, Wild & Krapp, 1995, quoted in Krapp, 1999): Interested students use more elaborative learning strategies ($r = .33$) for a specific exam, while they do not rehearse more often ($r = .05$). To conclude, it can be noted that topic interest plays an important role for learning (both for the process as well as for the

outcome) and therefore is essential for studying. Even though its importance seems proven, the assessment for topic interests happens in a very narrow way, limiting its use for broader counseling: e.g. testees have to rate their interest for text titles (Flowerday & Shell, 2015) or rate feeling-related and value-related adjectives anticipating the lecture of a specific text (Schiefele & Krapp, 1996).

Specific Interest Questionnaires from a Vocational Perspective

Published questionnaires on specific contexts are rare, but they do exist, usually assessing just one vocational interest type, such as interest solely for (just) economic or social professional fields (e.g. Bollschweiler & Bernath, 1998; Bollschweiler & Toggweiler, 2009). But research about narrower assessments is rare. A well-researched exception is the *Medical Specialty Preference Inventory MSPI* (Zimny, 1979) respective its revision *MSPI-R* (Richard, 2011), which is constructed for American medical students to help them finding the right specialty (for research about the instrument see e.g. Glavin, Richard & Porfeli, 2009; Porfeli, Richard & Savickas, 2010; Sodano & Richard, 2009). In such cases, common or broader inventories (e.g. inventories basing on Holland's interest model) have rather failed to be helpful (Burns, 2016).

One more approach for the assessment of specific interests can be found for users who undergo a (self-assessment) test battery to gain counseling for a specific field of study: For these users it can be argued that they already have a general interest for the field of study they are aspiring (e.g. when they want to study a specific technical field, they probably do have technical interest), but it might need a further examination, whether the needed (specific) interest is indeed at a sufficient level and the idea about the contents of the field of study is realistic (Weitensfelder, Undeutsch, Khorramdel & Useini, 2012). In such cases, the questions of specific interest questionnaires can also help to give an insight in the contents of the field of study.

Gender Differences in Interest Questionnaires

Several studies reveal gender differences in interest research (e.g. Hell & Päßler, 2011; Su, Rounds & Armstrong, 2009; Woods & Hampson, 2010). When it comes to vocational interests, differences regarding both the magnitude as well as the implication of interest parameters could be found. A prevalent metaanalysis of Su, Rounds and Armstrong (2009) came to the result that men prefer working with things, while women prefer people-lasting activities. In the RIASEC model, women scored higher in A, S and C, while men scored higher in I and R. With an effect size of Cohen d = .93, the difference on the things-people-dimension is remarkable. The cause for this difference is not completely resolved yet; explanatory approaches vary from hormonal causes (Hell & Päßler, 2011) to childhood personality traits (Woods & Hampson, 2010). Recently it could also be shown for science fields that gender-biased feedback might influence interest, with women reporting lower and men higher interests after getting feedback with a pro-male bias (Thoman & Sansone, 2016).

Regardless of their origin, gender differences raise the question of gender fairness in the construction of interest questionnaires – a matter that has rather been insufficiently reflected in the German-speaking area (Wetzel & Hell, 2012). Nevertheless it is necessary: Pässler, Beinicke and Hell (2014) showed that the famous *AIST-R* (Bergmann & Eder, 2005) produces differential validities for men and women. In terms of fairness, a different predictive validity is nothing else than a disadvantage for one group.

It is not clear, whether gender differences also exist in other than general vocational interest questionnaires. Burns (2014) investigated the above-mentioned medicine-specific interest questionnaire *MSPI-R* (Richard, 2011) and could only find insignificant differences regarding the predictive validity. No gender difference could also be found in a specific interest questionnaire for mechanical engineering (“Authors”, in prep.), though the sample for mechanical engineering might already have been self-selected by the volunteering test-takers themselves.

If gender differences are looked upon, usually they are investigated at scale level. Only a few studies (Einarsdóttir & Rounds, 2009; Pässler, Beinicke & Hell, 2014; Wetzel & Hell, 2013) focus on item level. These studies show that "differential item functioning" (DIF) regarding gender is an issue – which means that men and women interpret single items differently, even if the underlying interest dimension is the same. Einarsdóttir and Rounds (2009) pointed out that even two thirds of the famous *Strong Interest Inventory SII* in its 1994-version (Harmon, Hansen, Borgen & Hammer, 1994) show gender-specific DIF. DIF could also be found on each RIASEC-scale of the German *AIST-R*, most distinctive at the realistic scale (Wetzel & Hell, 2013). Validity of the Rasch model and therefore not hints on DIF on a global level however could be found for a field-specific questionnaire for mechanical engineering studies ("Authors", in prep.). Eliminating DIF-items does not necessarily diminish gender differences at scale level, though in the realistic scale the item-exclusion lead at least to a decrease of the differences (Wetzel & Hell, 2013). Certainly an elimination of DIF-items shows the positive effect of reducing different predictive validities for men and women (Pässler et al., 2014).

Main Aims of the Study and Study Questions

The article presents the evaluation of a field-specific interest questionnaire for *architecture*. The research focuses on the psychometric investigation of the developed questionnaire to make sure that the reported score is a fair measure overall and does not disadvantage study applicants regarding their gender or native language. Therefore the questions to be answered in this manuscript are

- whether the calculation of scores, as assumed in the construction process, is justified (study question 1),
- whether the items are fair regarding gender (study question 2a) and language (study question 2b).

Fairness on item level does not mean that there are no differences on a global level, so group differences will additionally be looked upon with a simple Welch-test.

Before focusing on the study questions mentioned above, the evaluated questionnaire shall be introduced.

The Field-Specific Interest Questionnaire for *Architecture*

The interest questionnaire presented is part of an extensive self-assessment for study applicants in *architecture* at the *Technische Universität Wien* (the *Viennese Self-Assessments* for technical majors are accessible at studienwahl.tuwien.ac.at, see also Kubinger, Frebort, Khorramdel & Weitensfelder, 2012), where several requirements for this field of study are assessed, with interest being one part of it. (The test battery takes most testees between 1.5 and 3 hours to complete.)

The questionnaire bases on the bachelor curriculum and contains actual or at least potential study contents. For item construction, the course catalogue of the field of study was looked through, with the course descriptions serving as basis for item formulations: Testees are asked whether they would be interested in doing what is (or at least could be) a topic in the courses (e.g. making free-hand drawings). Additionally, two different interest types are differentiated, inspired by Brickenkamp (1990): In Brickenkamp's *Generelle Interessen-Skala GIS*, three behavioral classes in a hierarchical order are distinguished: receptive, reproductive and creative interests: In a first phase, people have an increased receptiveness in a certain interest area (receptive), which leads to the intent to perform or reproduce activities in this area (reproductive), and finally, people want to become inventive or bring in own ideas (creative). Though the behavioral classes in Brickenkamp's model are in a hierarchical order, this was not adopted for the presented interest questionnaire after a lecturer's information that in architecture many students want to be creative, but sometimes their wish for acquiring the fundamental knowledge is poor. Therefore, two different sorts of interest resulted in the questionnaire: Namely "learning interest" (resp. the interest to gain knowledge, but also to

listen to things or look at things) and the interest on working with the contents in a reproductive or even creative way ("content interest", corresponds with Brickenkamp's reproductive and creative interest). The following items serve as an example: "I would like to look at the blueprint of a building with a computer simulation program" (learning interest) and "I would like to design a building with a computer simulation program by myself" (content interest). Testees get both feedback about their total score for their learning interest as well as about their total score for content interest. Regarding content interest, they also get a criterion-based ("strength"/"weakness") feedback about the six subjects included in the questionnaire (see tab. 1), while the number of items within each subject is very small (two to four items). The questionnaire contains 17x2 items (17 matters are asked twice, both as "learning interest" and "content interest") at a 4-point dichotomized Likert scale. The 4-point scale should offer some choice to the users, but the answers are calculated in a dichotomized way to make the Rasch model applicable (see later section).

Methods

Participants

The sample consists of anonymous users of the *Viennese Self-Assessment* for *Architecture* who completed the test battery during the survey period. Of users who passed the test battery during the inquiry period several times, only the first results were considered (though a multiple user registration or a passing before the survey period started wasn't controlled). Known registries for test or research purposes were excluded from the sample.

The sample consists of users of a survey period from March till December 2012, but the survey period for this sample is fragmented: During the survey period, data was also collected for an experimental study that lead to setting changes of the test battery for delimited sub-periods of time. Data from sub-periods with changed settings were removed from the present sample.

The sample offers the benefit that it consists of the target population for which the questionnaire was constructed: Namely (future) study applicants and other testees who seek advice for the particular field of study. On the downside, there is no control over test conditions or the sincerity of the assessment process, so that the completion of the self-assessment (a test battery of several hours) itself has to serve as a sincerity criterion.

Altogether, there is data from 441 people. Apart from an e-mail-adress, no data is needed to access the test battery, and demographic information is not asked. For a major part of the survey period, an additional questionnaire was provided though, where demographic data was asked: 411 users got questions about demographic data, which was answered in a plausible and (self-reported) honest way by 399 users, out of which 74 reported to have another mother-language than German, 239 reported to be female. Reported mean age was 21.14 years (MIN: 13, MAX: 50, SD: 4.962).

Measures

To answer the study questions, the interest questionnaire described is the only measure needed.

Procedure

Apart from Rasch model calculations, standard software (SPSS, version 20) was used. For Rasch model calculations, the program package *eRm* (Version 0.15-5, Mair, Hatzinger & Maier, 2015; see also Mair & Hatzinger, 2007) for the statistical software *R* (Version 3.2.2, R Core Team, 2015) is used.

Rasch model calculations. Being a model of the item response theory (IRT), the dichotomous logistic model of Rasch (1960) is the only one which offers the benefit of empiric verifiability (Kubinger & Draxler, 2007). In a nutshell the so-called “Rasch model” indicates that the probability of solving an item depends only on the ability of the testee and the difficulty of the item, but not on any systematic influences of other parameters (e.g. other personality traits, guessing parameters, learning effects). Therefore, if the Rasch model does

not hold, the questionnaire score cannot be seen as a fair measure of a testee's achievement (e.g. Kubinger, 2005), since other parameters (such as gender or learning effects) might influence the overall score. The Rasch model is only applicable for dichotomized data, which is why the questionnaire's answer categories are dichotomized up front. As a model test, Andersen's likelihood ratio-test (Andersen, 1973) is applied, testing whether the same item difficulty parameters are valid for different subgroups. Apart from the common split criterion "score" (median split), where testees with higher reported interest are opposed to testees with lower reported interest, the split criteria "gender[1]" and "language" (German vs. non-German) are checked to ensure that the questionnaire is fair for these groups. To test for subgroup invariance on item level, z-tests of Fischer and Scheiblechner (1970) with estimation of Wald (1943; often referred to as "Wald test") and graphical model checks (scatterplots and plottings of confidence parameter intervals) are applied.

While no testees are excluded from the data set for the split criterion "score", for the split criteria "gender" and "language" only those testees remain in the sample, who completed the demographic declarations plausible and who reported to have done so honestly. To avoid an elevation of the alpha error (e.g. Kubinger & Draxler, 2007), level of significance is corrected for Andersen's likelihood ratio test (including the split criterion "score", where the major part of the data set overlaps with the data set for the other split criteria). As suggested by Koller, Alexandrowicz & Hatzinger (2012) for testing subgroup invariance, alpha is set to α/q, resulting in a rounded alpha of .017 for an initial α of .05 and three split criteria (= q) in Andersen's likelihood ratio-test. For tests on item level (z-tests, graphical model checks), alpha is set to .01 instead of correcting it to $\alpha/(q \times k)$ as suggested by Koller et al. (2012), since with 17 items (= k), model violations would have to be huge to be significant.

[1] The split criterion is referred to as "gender", since sex is not measured in a biological way: Testtakers were asked in a simple question whether they were male or female. In contrast to English language, where sex and gender are different words, the same word refers to both in German, and it can't be controlled whether testtakers answered the question regarding to their sex or gender.

Gender differences at scale level. To see, whether males and females differ in their magnitude of reported interests, a Welch test (t test for heterogenous variances) is applied, which is more advisable than a “regular” t test for homogenous variances including precondition tests (Kubinger, Rasch & Moder, 2009). For the questionnaire, where two variables are tested on the same sample, alpha is Bonferroni Holm corrected (Holm, 1979), based on an initial alpha of 5%.

Results

Classical test theory analyses show that reliability is sufficient for both overall scores (r = .729 for the overall “content interest” scale, r = .705 for the “learning interest” scale), but not for the subscales (see tab. 1); especially with only two items included. Regarding gender or language, no overall differences could be found in the Welch tests for “learning interest” or “content interest” (tab. 2): Males reach the same average scores as females, and German-speaking testtakers reach the same average scores as non-German-speaking testtakers.

Insert table 1 about here

Insert table 2 about here

Rasch model calculations. Regarding the scale “learning interest“, the Rasch model is not valid for any of the three split criteria. Since the items for the “learning interest” scale are designed to correspond with the “content interest” scale, the “content interest” scale is looked at first to check, whether item removals are necessary on that scale, before the “learning interest” scale is checked again.

Regarding the scale “content interest“, the Rasch model is primarily valid for the split criteria “gender“ and “language“, but not for the split criterion “score“, done by median split (see tab. 3 for all split criteria).

Insert table 3 about here

For the median split, z-tests of Fischer and Scheiblechner with estimation of Wald are applied on item level, and parameter confidence intervals are plotted to check for subgroup invariance. Neither the z-tests (tab. 4) nor the parameter confidence interval plotting shows a significant misfit, but one item (item number 5) shows only a small overlap in the plotting of parameter confidence intervals (see figure 1).

Insert figure 1 about here

Insert table 4 about here

Due to the results, item number 5 is removed in a first step. After doing so, the model is overall valid for all three split criteria, tested with Andersen’s likelihood ratio test (see tab. 3). This could be a step where the calculation ends, but to ensure that no items with DIF remain in the questionnaire graphical checks via scatterplots are applied. Figure 2 (first column) shows the graphical model checks (GMCs) for all three split criteria after removal of the first item. One more item (see arrow in the first column) looks peculiar for the split criterion “language”, which is a hint on DIF for this item, so that a second item is removed. Afterwards the Rasch model is still valid for all three split criteria in the likelihood ratio tests (tab. 3), but another GMC for the split criterion “score” shows one more peculiar item (figure

2, second column). Therefore, a third item is removed. The model is now still valid for all three split criteria for the “content interest” scale, and additionally GMCs show no more misfits at item level (figure 3).

Since the “content interest” scale is now fair regarding all split criteria, the corresponding items of the “learning interest” scale are looked at again with just those items remaining that fit the model for the “content interest” scale. But still there is still no validity of the Rasch model for all three split criteria in the “learning interest” scale (tab. 3). Investigating the model at item level (z-tests, see tab. 4) shows that several items show misfits with the model, with different items for every split criterion. Since so many and further item removals do not seem reasonable, the “learning interest” scale cannot be considered as a fair measure.

Insert figure 2 about here

Insert figure 3 about here

Discussion

Though they are not very common, field-specific interest questionnaires could provide useful information for test-takers: When it comes to studying, tailor-made items can be constructed corresponding to very specific and detailed matters, making it possible to give study applicants already an idea about the curriculum (e.g. Weitensfelder et al. 2012), or to focus on aspects where broad inventories fail to be helpful (e.g. Burns, 2016).

One tailor-made questionnaire was looked at especially regarding its psychometric criteria, where the questionnaire showed several aspects that might need revision: Regarding the “content interest” scale, up to three items have to be excluded to make the scale

unidimensional concerning all three split criteria. After item removals, the scale is unidimensional and therefore its overall score represents a fair measure. (Nevertheless, cross validations would be necessary to assure unidimensionality.) It is especially worth mentioning that the scale is fair regarding the split criterion "gender", while previous results about less specific interest questionnaires rather showed gender differences at item level (e.g. Einarsdóttir & Rounds, 2009; Wetzel & Hell, 2013). The gender-fairness of the "content interest" scale raises the question, whether more specific questions could prevent from DIF in interest questionnaires: It might be possible, that due to its narrowness questions in specific questionnaires are formulated in a more precise way. Not only the scale presented, but also another field-specific interest questionnaire constructed for mechanical engineering basically showed Rasch model validity for the split criterion gender ("Authors", in prep.). Therefore, it could be a matter of future examinations, whether more scope of interpretation leads to different interpretations for men and women.

Also at scale level, no overall gender differences could be found for the questionnaire scores (see tab. 2), but it might be possible that users who underwent the questionnaire were already "self-selected" by their own interest. It should not be forgotten that the results base on a very specific sample, namely (probable future) study applicants for a major in *architecture*, who voluntarily completed an online *Self-Assessment* for their aspired field of study. Though this is exactly the population the questionnaire was constructed for, it would be possible that overall gender effects occur, when the questionnaire is applied to another than that goal population.

While the "content interest" scale seems overall satisfactory, the "learning interest" scale shows several psychometric disadvantages, not showing conformity with the Rasch model in any of the three split criteria. However, the learning interest scale seems problematic anyways: Test-takers are asked, whether they are interested in gaining knowledge about a specific topic, which probably leads to an intermixture of interest for learning and interest for

the contents. This might represent the everyday-life of studying, but the intermixture leads to a higher linguistic complexity. It cannot be ruled out that the linguistic complexity might influence different subgroups differently – which might be one of the reasons why this part of the questionnaire shows several misfitting items in the Rasch model. Due to the psychometric handicaps of the "learning interest" scale, it should be considered whether the benefit of a separate learning scale outbalances the psychometric handicap at all. From a psychometric point of view, it rather seems advisable to remove the scale completely and assess the interest for learning in another way (e.g. via a questionnaire that's unrelated to content interest).

Summing up the Rasch model evaluations (and answering study question 1), the calculation of overall scores is not justified without revisions: The "learning interest" scale cannot be seen as a fair measure, while for the "content interest" scale excluding up to three items would make the scale a fair measure. Regarding the overall results of the "content interest" scale, but also regarding reliability, it cannot be recommended to interpret results of single subscales, though the correspondence of subscales with curriculum contents seems promising for counseling at a first glimpse. But low subscale reliabilities and frequent curricula changes represent a down side of this approach - especially with the Rasch model being valid, an overall score alone would perfectly do anyways.

Answering study question 2, the "content interest" scale, but not the "learning interest" scale, can be seen as fair regarding gender (2a) and mother-language (2b). For both scales, no gender difference is visible at a global level in the given population.

Several questions still remain open for future examinations: From a psychometric point of view, it could be worth questioning, whether the questionnaire (which showed skewed overall scores, see tab. 1) could rather benefit from a polytomous data model like a Partial Credit Model (Masters, 1982). The questionnaire was designed after the Rasch model up front and all scores and feedbacks base on a dichotomized data set, but it might be worth examining, whether a polytomous calculation and a polytomous model might fit even better.

But there is also room for research on other than psychometric examinations: Especially research on the predictive validity would be needed. For specialty interest inventories, therefore also for the presented field-specific questionnaires, a so-called “empirical model” is recommended (Porfeli, Richard & Savickas, 2010): In an empirical model, an outside criterion should be predicted – as opposed to an inductive model, which focuses on the internal structure of the questionnaire. Since the questionnaire items aim at concrete study contents, it would be surprising if the questionnaire was not valid. However, it would be interesting to investigate the incremental validity of the field-specific questionnaire combined with a general interest questionnaire to see whether the specificity of the questionnaire indeed is an additional benefit for test-takers.

As an implication for future research that exceeds the presented questionnaire, the issue of gender-specific DIF might also need further examination: It seems that overall differences and gender-specific DIF might occur in general interest questionnaires, but that did not seem to be an issue in the “content interest” scale of the questionnaire presented or in another specific interest questionnaire (“Authors”, in prep.). Therefore it might be worth examining whether it is the scope of interpretation that leads to different item interpretations for men and women. If indeed a more specific item formulation proves to diminish gender-specific DIF, this finding could be a major improvement for the development of gender-fair questionnaires in general.

Acknowledgements

Please see separate file due to blindness.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123-140. doi:10.1007/BF02291180

Bergmann, C., & Eder, F. (2005). *AIST-R. Allgemeiner Interessen-Struktur-Test mit Umwelt-Struktur-Test (UST-R) – Revision* [GIST-R. General Interest Structure Test with Environment Structure Test (EST-R) – Revision]. Göttingen: Beltz Test GmbH.

Bollschweiler, R., & Bernath, W. (1998). *SIT: Sozial-Interessentest* [SIT: Social Interest Test]. Dübendorf: SVB.

Bollschweiler, R., & Toggweiler, S. (2009). *KV-Interessentest R* [KV Interest Test R] *(2. Ed.).* Bern: SDBB.

Brickenkamp, R. (1990). *GIS: Die Generelle Interessen-Skala [GIS: The General Interest-Scale]*. Göttingen: Hogrefe.

Burns, S. T. (2014). Gender Differences in the Validity of Career Interest Inventories. *Psychology, 5,* 785-797. doi:10.4236/psych.2014.58089

Burns, S. T. (2016). History of medical specialty interest assessment. *Cogent education, 3,* 1131062. [Retrieved from: https://www.cogentoa.com/article/10.1080/2331186X.2015.1131062.pdf, Feb. 8th 2017]

Einarsdóttir, S., & Rounds, J. (2009). Gender bias and construct validity in vocational interest measurement: Differential item functioning in the Strong Interest Inventory. *Journal of Vocational Behavior, 74,* 295-307. doi:10.1016/j.jvb.2009.01.003

Fischer, G. H., & Scheiblechner H. H. (1970). Algorithmen und Programme für das probabilistische Testmodell von Rasch [Algorithms and programs for Rasch's probabilistic test model]. *Psychologische Beiträge, 12,* 23-51.

Flowerday, T., & Shell, D. F. (2015). Distangling the effects of interest and choice on learning, engagement, and attitude. *Learning and Individual Differences, 40,* 134-140. doi:10.1016/j.lindif.2015.05.003

Glavin, K. W., Richard, G. V., & Porfeli, E. J. (2009). Predictive validity of the medical speciality preference inventory. *Journal of Vocational Behavior, 74,* 128-133. doi:10.1016/j.jvb.2008.11.004

Harmon, L. W., Hansen, J. C., Borgen, F. H., & Hammer, A. L. (1994). *Strong Interest Inventory: Applications and technical guide.* Palo Alto, CA: Consulting Psychologists Press, Inc.

Hell, B., & Päßler, K. (2011). Are occupational interests hormonally influenced? The 2D:4D-interest nexus. *Personality and Individual Differences, 51,* 367-380. doi:10.1016/j.paid.2010.05.033

Hidi, S. (1990). Interest and Its Contribution as a Mental Resource for Learning. *Review of Educational Research, 60,* 549-571. doi:10.3102/00346543060004549

Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology, 6,* 35-45.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Holm, S. (1979): A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics, 6,* 65-70.

Koller, I., Alexandrowicz, R., & Hatzinger, R. (2012). *Das Rasch Modell in der Praxis: Eine Einführung mit eRm* [The Rasch model in practical applications: An introduction with eRm]. Vienna: facultas.wuv, UTB.

Krapp, A. (1999). Interest, motivation and learning: An educational-psychological perspective. *European Journal of Psychology of Education, 14,* 23-40.

Kubinger, K. D. (2005). Psychological Test Calibration Using the Rasch Model – Some Critical Suggestions on Traditional Approaches. *International Journal of Testing, 5,* 377-394. doi:10.1207/s15327574ijt0504_3

Kubinger, K. D., & Draxler, C. (2007). Probleme bei der Testkonstruktion nach dem *Rasch*-Modell [Some problems in calibrating an item pool according to the Rasch model]. *Diagnostica, 53,* 131-143. doi:10.1026/0012-1924.53.3.131

Kubinger, K. D., Frebort, M., Khorramdel, L., & Weitensfelder, L. (Eds.) „Wiener Autorenkollektiv Studienberatungstests“ (2012). *Self-Assessment: Theorie und Konzepte* [Self-Assessment: Theory and concepts]. Lengerich: Pabst.

Kubinger, K. D., Rasch, D., & Moder, K. (2009). Zur Legende der Voraussetzungen des *t*-Tests für unabhängige Stichproben [About the legend of t-test requirements for independent samples]. *Psychologische Rundschau, 60,* 26-27. doi:10.1026/0033-3042.60.1.26

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20,* 1-20. doi:10.18637/jss.v020.i09

Mair, P., Hatzinger, R., & Maier M. J. (2015). *eRm: Extended Rasch Modeling (Version 0.15-5).* Retrieved from http://erm.r-forge.r-project.org/

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika, 47,* 149-74. doi:10.1007/BF02296272

Pässler, K., Beinicke, A., & Hell, B. (2014). Gender-Related Differential Validity and Differential Prediction in Interest Inventories. *Journal of Career Assessment, 22,* 138-152. doi:10.1177/1069072713492934

Porfeli, E. J., Richard, G. V., & Savickas, M. L. (2010). Development of specialization scales for the MSPI: A comparison of empirical and inductive strategies. *Journal of Vocational Behavior, 77,* 227-237. doi:10.1016/j.jvb.2010.04.011

R Core Team (2015). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute of Education Research. (Expanded Edition, 1980. Chicago: University of Chicago Press.)

Richard, G. (2011). *Medical Speciality Preference Inventory, Revised Edition.* Washington DC: Association of American Medical Colleges.

Schiefele, U., & Krapp, A. (1996). Topic Interest and Free Recall of Expository Text. *Learning and Individual Differences, 8,* 141-160. doi:10.1016/S1041-6080(96)90030-8

Schiefele, U., Krapp, A., & Winteler, A. (1992). Interest as a Predictor of Academic Achievement: A Meta-Analysis of Research. In K. A. Renninger, S. Hidi & A. Krapp (Eds.), *The Role of Interest in Learning and Development* (pp. 183-212). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schiefele, U., Wild, K. P., & Krapp, A. (1995). Course-specific interest and extrinsic motivation as predictors of specific learning strategies and course grades. Paper presented at the 6th EARLI Conference in Nijmegen.

Sodano, S. M. (2015). Meaning, measurement, and assessment of vocational interests for career intervention. In Paul J. Hartung, Mark L. Savickas, Bruce W. Walsh (Eds.), *APA handbook of career intervention, Volume 1: Foundations. APA handbooks in psychology* (pp. 281-301). Washington, DC: American Psychological Association.

Sodano, S. M., & Richard, G. V. (2009). Construct validity of the medical specialty preference inventory: A critical analysis. *Journal of Vocational Behavior, 74,* 30-37. doi:10.1016/j.jvb.2008.10.001

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and Things, Women and People: A Meta-Analysis of Sex Differences in Interests. *Psychological Bulletin, 135,* 859-884. doi:10.1037/a0017364

Thoman, D. B., & Sansone, C. (2016). Gender bias triggers diverging science interests between women and men: The role of activity interest appraisals. *Motivation and Emotion, 40,* 464-477. doi:10.1007/s11031-016-9550-1

Tracey, T. J. G. (2012). Problems with single interest scales: Implications of the general factor. *Journal of Vocational Behavior, 81,* 378-384. doi:10.1016/j.jvb.2012.10.001

Wald, A. (1943). Tests of statistical hypothesis concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society, 54,* 426–482.

Weitensfelder, L., Undeutsch, N., Khorramdel, L., & Useini, C. (2012). Intrinsische Studienmotivation: Interesse, Erwartungen und Selbstkonzept eigener Fähigkeiten [Intrinsic study motivation: Interest, expectations and self-concept of own abilities]. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Eds.), *Self-Assessment: Theorie und Konzepte* [Self-assessment: Theory and concepts] (pp. 119-142). Lengerich: Pabst.

Wetzel, E., & Hell, B. (2012). NIE-Richtlinien zur Gender Fairness von Interessentests. Sind deutschsprachige Interessentests gender-fair? Eine qualitative Analyse [Are German interest inventories gender-fair? – A qualitative analysis]. *Zeitschrift für Arbeits- und Organisationspsychologie, 56,* 37-47. doi:10.1026/0932-4089/a000065

Wetzel, E., & Hell, B. (2013). Gender-Related Differential Item Functioning in Vocational Interest Measurement. An Analysis of the AIST-R. *Journal of Individual Differences, 34,* 170-183. doi:10.1027/1614-0001/a000112

Woods, S. A., & Hampson, S. E. (2010). Predicting Adult Occupational Environments From Gender and Childhood Personality Traits. *Journal of Applied Psychology, 95,* 1045-1057. doi:10.1037/a0020600

Zimny, G. H. (1979). *Manual for the medical specialty preference inventory.* St. Louis, MO: St. Louis University School of Medicine.

"Authors" (in prep.). Implementation of a Field-Specific Interest Questionnaire for Mechanical Engineering: Lessons Learned from Psychometric Analyses.

Table 1
Scales and reliabilities for the interest questionnaire in Architecture

	Number of items (mean)	**Cronbach's Alpha**
Learning interest: total	**17 (13.72)**	**.705**
Content interest: total	**17 (12.94)**	**.729**
Content: Building construction, construction and realization	4 (2.61)	.497
Content: Design	4 (3.52)	.346
Content: Urban design	3 (2.62)	.481
Content: Art	2 (1.80)	.147
Content: Theory and history	2 (1.25)	.301
Content: Technology	2 (1.14)	.561

Notes: All results based on the complete sample (*architecture*: n = 441). All translations of subjects by the author. Naming and translation based on the curriculum that was valid when the questionnaire was constructed.

Table 2
Welch-tests regarding gender and language differences

	Gender					**Language**				
	Females		Males			German		non-German		
	M	*SD*	*M*	*SD*	*p*	*M*	*SD*	*M*	*SD*	*p*
"Learning Interest"	13.51	2.693	13.98	2.610	.043	13.68	2.658	13.81	2.718	.351
"Content Interest"	12.82	2.914	13.13	2.992	.161	12.95	2.875	12.92	3.259	.469

Notes: All p-values one-sided.
n = 399: 239 women; 325 German-speaking

Table 3
Andersen's likelihood ratio-tests for "learning interest" and "content interest"

Split Critierion	***n***	**Andersen χ^2**	***df***	$\chi^2_{\alpha=.017}$	***p***
"Learning Interest"					
Score	441	35.4	15	28.813	.002
Language	399	30.013	15	28.813	.012
Gender	399	55.695	16	30.198	<.001
"Content Interest"					
Score	441	31.015	14	27.414	.006
Language	399	25.291	16	30.198	.065
Gender	399	27.776	16	30.198	.034
"Content Interest" after one Item Removal					
Score	441	24.625	14	27.414	.038
Language	399	25.873	15	28.813	.039
Gender	399	27.23	15	28.813	.027
"Content Interest" after two Item Removals					
Score	441	24.555	12	24.569	.017
Language	399	17.027	14	27.414	.255
Gender	399	25.102	14	27.414	.034
"Content Interest" after three Item Removals					
Score	441	18.401	11	23.120	.073
Language	399	17.378	13	26.000	.183
Gender	399	25.253	13	26.000	.021
"Learning Interest" after three Item Removals					
Score	441	23.272	11	23.120	.016
Language	399	30.246	13	26.000	.004
Gender	399	44.932	13	26.000	<.001

Table 4:

z-Tests of Fischer and Scheiblechner with estimation of Wald

Scale: Content Interest			**Scale: Learning Interest (after removal of three items)**						
	Split criterion: Score			Split criterion: Score		Split criterion: Language		Split criterion: Gender	
	z-statistic	*p*-value		*z*-statistic	*p*-value	*z*-statistic	*p*-value	*z*-statistic	*p*-value
i01	2.046	.041	**rem.**	-	-	-	-	-	-
i02	-.712	.476	**L02**	-.621	.534	-1.033	.302	-.379	.704
i03	.905	.365	**rem.**	-	-	-	-	-	-
i04	1.671	.095	**L04**	1.651	.099	-3.324	.001	-.703	.482
i05	2.356	.018	**rem.**	-	-	-	-	-	-
i06	-1.764	.078	**L06**	-.603	.547	-1.264	.206	-2.680	.007
-	n.a.	n.a.	**L07**	n.a.	n.a.	-.757	.449	-.318	.751
i08	.900	.368	**L08**	1.223	.221	2.080	.038	1.797	.072
i09	-1.544	.123	**L09**	-.785	.432	-2.348	.019	-.013	.989
i10	1.449	.147	**L10**	3.428	.001	1.007	.314	2.009	.045
i11	.382	.702	**L11**	1.744	.081	1.333	.182	2.022	.043
i12	-1.776	.076	**L12**	-1.188	.235	.790	.429	-1.966	.049
i13	-.828	.408	**L13**	-.335	.738	1.108	.268	1.018	.309
i14	-1.775	.076	**L14**	-.470	.638	.675	.499	-3.086	.002
i15	.155	.876	**L15**	n.a.	n.a.	.777	.437	2.852	.004
-	n.a.	n.a.	**L16**	-.138	.890	.782	.434	.719	.472
i17	-1.655	.098	**L17**	-1.954	.051	-.914	.361	-1.830	.067

Notes: “n.a.”: item not available (inappropriate response patterns); “rem.”: item removed due to peculiarities in the “content interest” scale

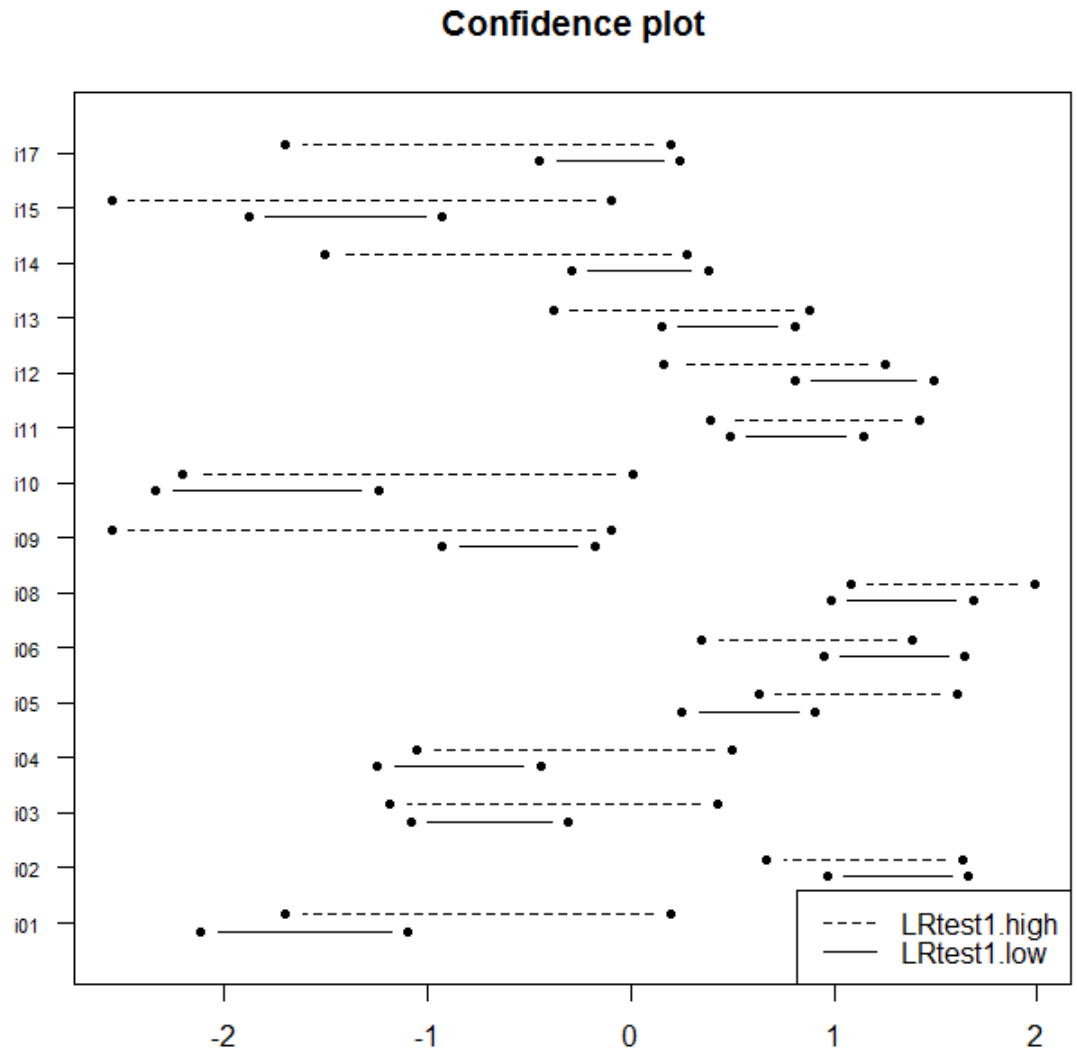


Figure 1. Plotting of parameter confidence intervals (plotDIF) for the scale “content interest” (median split); alpha is set to .01.

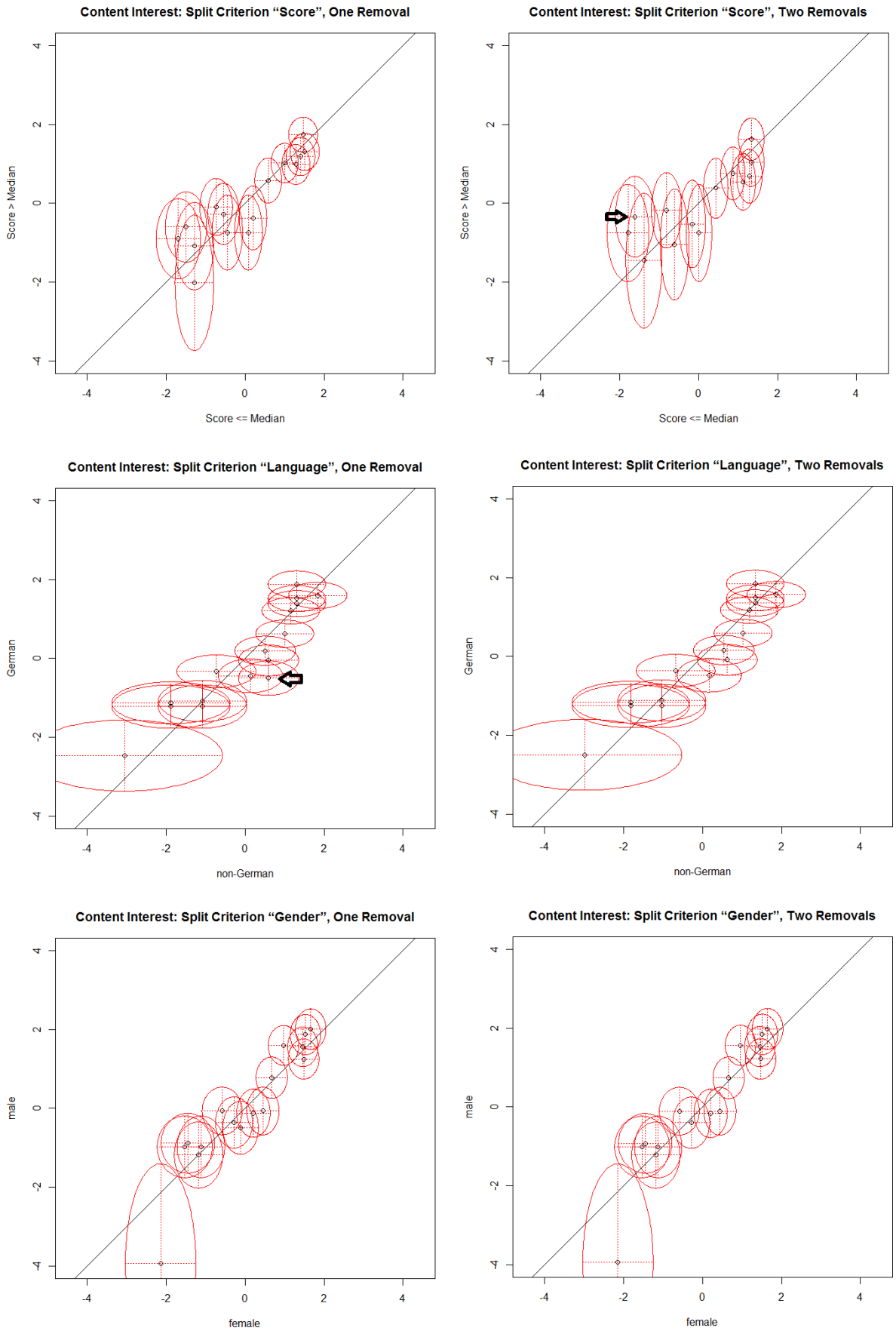


Figure 2. Graphical model checks (GMCs) for all split criteria after one and two item removals at the "content interest" scale

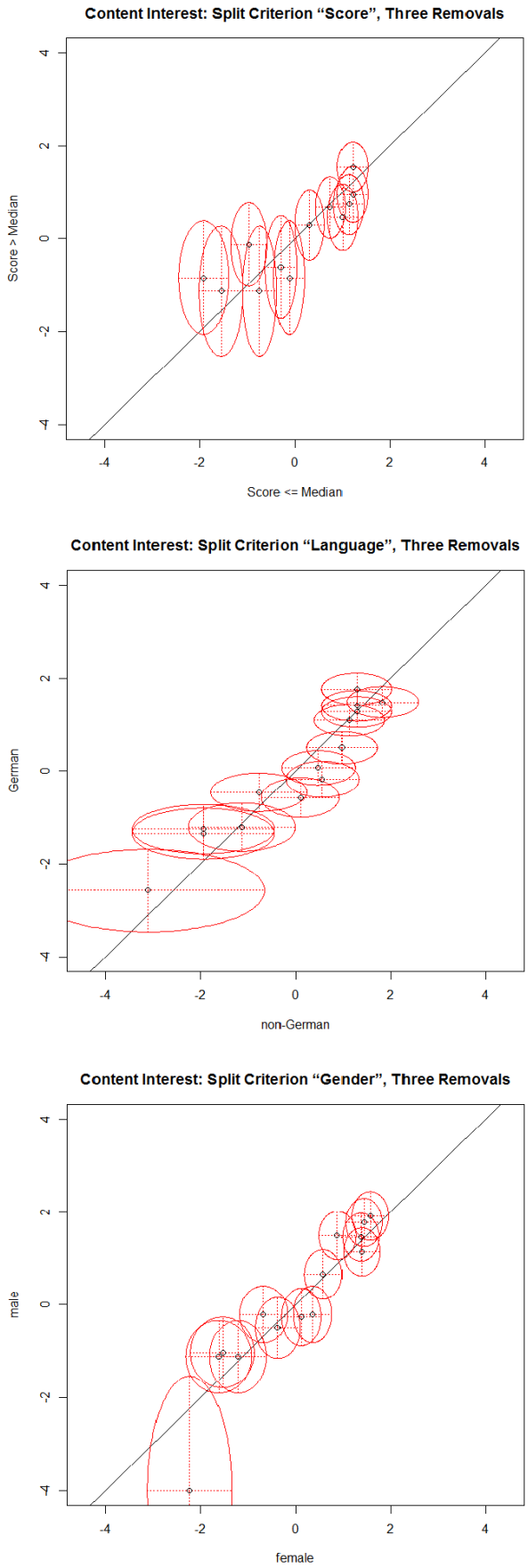


Figure 3. Graphical model checks (GMCs) for all split criteria after three item removals at the "content interest" scale

Anhang B

Voträge, auf die in der Klammer Bezug genommen wurde

1) Kubinger, K. D., Weitensfelder, L., Frebort, M., & Sonnleitner, P. (2012). *Non-cognitive vs. cognitive predictors for academic success – opportunities and limits of self-assessments illustrated by the Viennese products.* Paper presented at the AERA [American Educational and Research Association] Annual Meeting, Vancouver, April 13th - April 17th 2012.
   [Paper zum Vortrag]
2) Weitensfelder, L. (2011). *Conception of a facet-oriented spatial ability test (TARV – Test of Applied Relations and Visuo-spatial abilities).* Vortrag auf der 11th European Conference on Psychological Assessment, Riga, August 31st - September 3rd 2011.
   [Folien zum Votrag]
3) Weitensfelder, L., Frebort, M., Müller, C. E., & Mitschek, B. (2011). *Processing Characteristics and Reasonableness regarding self-assessments.* Vortrag auf der 11th European Conference on Psychological Assessment, Riga, August 31st - September 3rd 2011.
   [Folien zum Votrag]

Noncognitive Versus Cognitive Predictors for Academic Success: Opportunities and Limits of Self-Assessments Illustrated by the Viennese Products

Klaus D. KUBINGER, Lisbeth WEITENSFELDER, Martina FREBORT & Philipp SONNLEITNER

Division of Psychological Assessment and Applied Psychometrics including Centre of Testing and Consulting, University of Vienna, Austria

Address correspondence to Lisbeth Weitensfelder, Centre of Testing and Consulting, University of Vienna, Liebiggasse 5, 1010 Vienna, Austria (e-mail: lisbeth.weitensfelder@univie.ac.at).

Abstract

Self-assessments recently gained a lot of attention in German-speaking countries, especially regarding economic efficiency of limited university admissions. Self-assessments enable university applicants to test their qualification for the intended study field – usually unproctored, anonymous and just for information without binding consequences. Usually the inventory of self-assessments includes cognitive tests and personality questionnaires. Beyond that the Viennese self-assessments include experimental-based behavior tasks (based on the principle of R.B. Cattell; see Kubinger, 2009) which are impossible or at least very hard to fake. Especially with the results of the Viennese self-assessments it could be shown, that non-cognitive traits and work styles show a noteworthy relation to academic success.

Keywords: self-assessment, aptitude, non-cognitive predictors, experimental-based behavior tasks, academic success

Introduction

Self-assessment denominates a psychological test setting where testees test themselves, mostly via internet, so that they pass the whole assessment process unproctored and on their own responsibility (cf. Kubinger, 2009). Though self-assessments contain appropriate psychological tests, the tests are usually not exactly the same ones like in traditional treatment situations to prevent items and test concepts from spreading (see Kubinger, Frebort & Müller, 2012). For several university fields of study in German-speaking countries self-assessments are implemented as an advisory service, e.g. for psychology (see Kubinger, Moosbrugger, Frebort, Jonkisz & Reiß, 2007). The main goal of such self-assessments, where university applicants can test their qualification for the intended field of study themselves, is an increase in information for those participants. Both the wish for a better self-selection of (future) students into their study fields as well as the intent to improve the university's image might be a reason for the recent boom of self-assessments in study advisory context (cf. Kubinger, Frebort & Müller, 2012).

In general self-assessments as study advisory service have to meet the usual quality criteria for psychological tests (for an overview see Kubinger, 2003), but some requirements might be different: A representative sample for normative comparision might not be a necessity, however there has to

be a comparision with the requirements for the intended study field (see Kubinger et al, 2007), and the feedback of results has to be especially careful and precise since testees don't get personal support from psychologists.

Practice examples: The *Viennese Self-Assessments*

A wrong choice of a study field can have devastating consequences both for individuals as well as for the economy: Zimmerhofer, Heukamp and Hornke (2006) estimate the costs of Germany's college dropouts at about one billion euro each year. Lack of accomplishment, lack of motivation and vocational reorientation account for the biggest parts of dropouts (cf. Heublein, Hutzsch, Schreiber, Sommer & Besuch, 2009). Those dropout reasons show the importance of both cognitive as well as non-cognitive predictors, even though correlations between non-cognitive predictors and academic success are usually quite modest (see O'Connor and Paunonen, 2007; Sandner, 2008). Nevertheless, some researchers underline the importance of personality for predicting academic performance (Chamorro-Premuzic & Furnham, 2003; Conard, 2006), and requirement analyses for study fields contain non-cognitive as well as cognitive requirements (for an example see Khorramdel, Maurer, Frebort & Kubinger, 2012).

Based on such specific requirement analyses for several study fields, the Centre of Testing and Consulting (Division of Psychological Assessment and Applied Psychometrics, University of Vienna) developed self-assessements for psychology and several technical studies (architecture, engineering, computer science and electrotechnology). The test batteries for the *Viennese Self-Assessments* contain a variety of cognitive tests, questionnaires and experimental-based behavior tasks *sensu* R.B. Cattell. Cognitive tests include a memory test, tests for measuring logical or numeric abilities or specific ability tests to meet specific requirements (e.g. spatial abilities, visual structuring abilities; for examples see Weitensfelder, 2012, or Weitensfelder and Becker, 2012). Non-cognitive abilities are either asked (through questionnaires, e.g. about interest or study-relevant personality criteria) or measured in experimental-based behavior tasks: A learning test measures learning styles, a simple coding task surveys whether the goal-setting of testees is appropriate or not.

After finishing the test battery, the participating study applicants get an elaborated feedback about their strengths and weaknesses compared to the requirements of the intended study field. As far as possible, in case of weaknesses they get hints on how to train or compensate weaknesses. That the feedback in fact addresses to the goal population (namely: future applicants for the study fields) can be seen in information from anonymous users of the *Viennese Self-Assessment Computer Science/Electrotechnology*: Only 23 out of a sample of 325 anonymous users stated to be already students in the field the self-assessment was developed for. Out of the remaining users, more than 30% (93 users) quoted that they intended to study in that field with a 100%-certainty. Overall, the

probability's median to begin the intended fields of not-yet-students is at 80%, which underlines the importance of an appropriate and careful feedback (Figure 1).

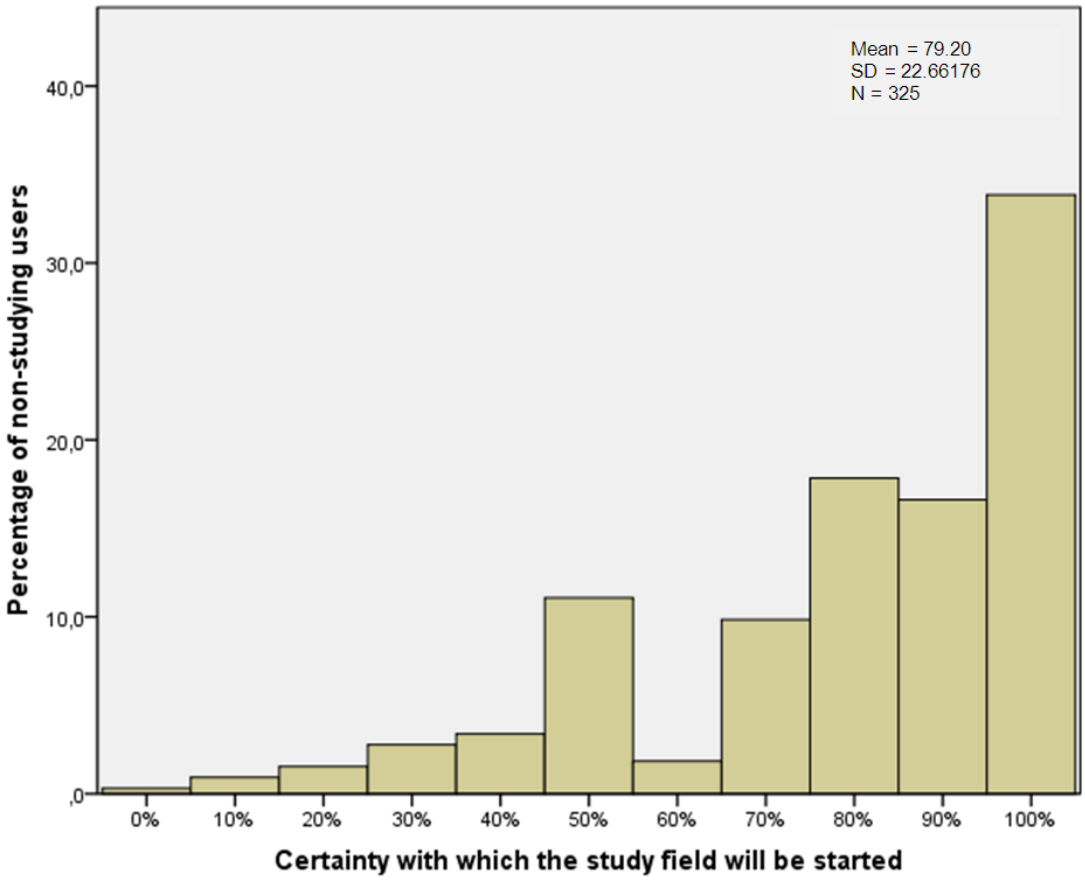


Figure 1: Certainty with which the study field will be started, sample of the Viennese Self-Assessment Computer Science/Electrotechnology

Cognitive vs. non-cognitive predictors: Method

For architecture (Müller, 2011) and engineering (Mitschek, 2011) university students from the Technical University of Vienna voluntarily passed through pilot phases[2] of the *Viennese Self-Assessment* for their field of study. Additionally in both samples students filled out a questionnaire about their study success. To take a closer look at a correlation between test results and study success, data from freshmen (in their first semester) as well as unrealistic or unclear statements about study success were reduced, so that 60 architecture and 59 engineering students remained in the overall sample (78 male; mean age 22.08 years, SD = 2.8). However, not all of them finished all tests and questionnaires.

[2] After the pilot phase self-assessments were reduced for crucial items or ones with inappropriate difficulties. Also for the present study, those items are reduced from the data set.

Since the record of the study success used in the different self-assessments was not identical, the collected data had to be z-transformed to combine the samples from architecture and engineering; outliners were set to the mean plus two standard deviations.

Apart from study success, scores from the used matrices tests and results from a learning test were z-transformed to meet different amounts of items. Regarding variables from a newly developed personality questionnaire (see Khorramdel & Maurer, 2012) the data was skewed, so that a z-transformation was not possible, and raw scores of identical items in both samples were used instead. Three subscales out of seven subscales of the questionnaire were examined: Frustration tolerance, motivation and organizational skills.

Regarding interest, students from those two branches got different questionnaires: To represent each field of study perfectly, two questionnaires with study-course-related items were developed (see Weitensfelder, Undeutsch, Khorramdel & Useini, 2012) which contained questions about architectural contents only or about engineering topics only. For comparision of the two fields, the overall scores of the questionnaires were z-transformed.

To measure reasoning, a matrices test was used (see Undeutsch, 2012). Memory was measured with the learning test LAMBDA (see Kubinger, Haiden, Karolyi & Maryschka, 2012) where students had to memorize 20 units of an organization chart to later correct mistakes in that chart. The learning test in the *Viennese Self-Assessments* cancels automatically if a time limit of 20 minutes is exceeded (that includes correction times). Since it is not clear whether and when students who didn't fulfill the quest in time would have succeeded, only data of succeeding students was used for correlations, whereas data of both succeeding and timed-out students was observed in crosstabs.

For all students, the relevant success criteria was the average studying speed. To see whether cognitive and non-cognitive variables are related to academic performance, Pearson correlations are calculated, while the following variables where included: study success (z-transformed), learning time in the learning test as an indicator for memory (z-transformed), reasoning (z-transformed), overall interest for their study field (z-transformed) and subscales of a newly developed personality questionnaire for students (raw scores). All correlations are expected to be one-sided and positive, except for the needed time to memorize, where negative correlations with studying speed are expected.

To take a closer look at experimental-based behavior tasks, the simple coding test was looked at more closely. In that test students have to do very simple coding tasks in several rounds using a coding key. To measure goal-setting, testees get a feedback on how they performed, while they are asked after how many symbols they might manage to code next. Goal-setting of students was divided in three categories: appropriate goal-setting, too low goal-setting (more than 25% underestimation compared of the own proficiency level) and too high goal setting (more than 15%

overestimation of the own proficiency level; see Kubinger and Ebenhöh, 2012). For these categories, crosstabs for three different success categories (high studying speed, average studying speed, low studying speed) were calculated. According to the requirement analyses for engineering, an accurate work style is very important for engineering students, so that with the same simple coding test the relative amount of "misclicks" with the mouse was measured as well. 2% or more misclicks were seen as inaccurate due to the simplicity of the task and the amount of symbols coded.

Cognitive vs. non-cognitive predictors: Results

Correlation coefficients show, that there are no significant correlations between the results in the cognitive tests and academic performance. However, there were some small correlations between non-cognitive traits and academic success: Highest correlations (r = .27) could be found for organizational skills, small correlations could be found for motivation (r = .26) and overall interest for the field (r = .26) as well as for frustration tolerance (r = .20). There were also some intercorrelations between the contemplated variables, especially regarding variables from the personality questionnaire for students. Furthermore, some correlations between overall interest and some personality scales as well as between reasoning and memorizing time could be found. Table 1 shows the correlation matrix for the examined variables.

Tab. 1: Correlation matrix for cognitive and non-cognitive predictors and success

		Success	Reas	Mem	Org	Mot	Frust	Interest
Success criteria								
		-						
Reasoning	*r*	.058						
	sig.	.283	-					
	n	99						
Memorizing time	*r*	-.103	.245*					
	sig.	.208	.039	-				
	n	65	53					
Organiziational skills	*r*	.273*	.211*	-.022				
	sig.	.004	.034	.435	-			
	n	91	76	58				
Motivation	*r*	.264*	-.006	-.182	.443*			
	sig.	.006	.478	.084	.000	-		
	n	90	81	59	82			
Frustration tolerance	*r*	.195*	-.156	-.201	.363*	.457*		
	sig.	.027	.071	.063	.000	.000	-	
	n	98	90	59	81	88		
Interest	*r*	.260*	-.057	-.057	.247*	.359*	.223*	
	sig.	.010	.317	.345	.019	.001	.026	-
	n	80	73	52	71	76	76	

As mentioned above, data from students where the learning-test cancelled due to time-out could not be incorporated in the correlation matrix. Table 2 shows the learning test data from all students in a crosstable, regardless whether they succeeded in the test or not.

It can be seen that while there is basically no difference in the time needed for memorizing regarding the success criteria, there are differences between those ones who timed out and those ones who completed the test in time.

Tab. 2: Results in the learning test and academic success

Learning test	**Success criteria: average studying speed**			**Overall**
	Low speed	Medium speed	High speed	
long memorizing time	5 (14%)	3 (9%)	5 (17%)	13
short memorizing time	16 (44%)	17 (49%)	19 (66%)	52
test timed out	15 (42%)	15 (43%)	5 (17%)	35
Overall	36 (100%)	35 (100%)	29 (100%)	100

Regarding goal-setting in the experimental-based behavior task, it can be seen in the architecture sample, that students with low studying speed more often show inappropriate goal-settings (Table 3). However, this could not be found for engineering students, where more and less successful students did not differ regarding this trait.

In engineering students showed at the coding test, that an imprecise work style rather seems to get along with lower studying speed (Table 4).

Tab. 3: Goal-setting of architectural students

Goal-setting	**Success criteria: average studying speed**			**Overall**
	Low speed	Medium speed	High speed	
Low goal-setting	6 (67%)	3 (21%)	6 (40%)	15
Appropriate goal-setting	3 (33%)	11 (79%)	8 (53%)	22
High goal-setting	0 (0%)	0 (0%)	1 (7%)	1
Overall	9 (100%)	14 (100%)	15 (100%)	38

Tab. 4: Accuracy of engineering students when doing simple coding tasks

Accuracy	**Success criteria: average studying speed**			**Overall**
	Low speed	Medium speed	High speed	
Rather inaccurate coding	12 (60%)	7 (50%)	4 (44%)	23
Rather accurate coding	8 (40%)	7 (50%)	5 (56%)	20
Overall	20 (100%)	14 (100%)	9 (100%)	43

Discussion

The present study shows the importance of taking a closer look at non-cognitive predictors for academic success: Out of the examined variables it have been especially personality traits and interest that showed moderate correlations with the success criteria studying speed. But that does not imply that cognitive predictors can be ignored: Intellectual abilities have been shown to be important for academic success in several contexts (e.g. Vásquez & de Anglat, 2009), even though correlations of intellectual abilities with academic success sometimes seem to be rather small (e.g. Busato, Prins, Elshout & Hamaker, 2000).

One explanation for the lack of correlations between the cognitive predictors and the success criteria in the present study might be, that the sample was already pre-selected with probably just high-ability-testees (all of them were already students in their field of study), and students with low abilities might already have dropped out or not started the study field at all. Furthermore, as the crosstab from the learning test showed, a certain level of ability seems useful: While there was basically no difference regarding the success criteria from testees who managed to pass the test, only a few of the testees who did not manage to finish the test at all were in the group of the more successful students, while most where in the group of students with low or average studying speed (see Table 2, “test timed out”). Also, the used learning test has proven itself in earlier contexts already (Kubinger, Haiden, Karolyi & Maryschka, 2012).

That leads to the conclusion, that a certain level of abilities might be needed; however, once a certain level is reached, it seems that especially non-cognitive traits are important for academic success. Nevertheless in practice, self-assessments as a study advisory service should contain cognitive tests as well as non-cognitive tests: There always might be study applicants whose abilities could be below a certain level and who therefore would benefit from hints on how to train deficits.

Also, work styles such as appropriate goal-setting or an exact work style even at simple tasks seem to have relevance for study success, though sample size is not very big and future research about this topic seems advising. Work style might also have had an influence on the results of the cognitive tests in the sample; a hint is given by the interrelation between memorizing time and the result in the matrices test: Since testees who needed longer time for memorizing by trend had better results in reasoning, one explanation might be that they just worked more precisely, and that correlation between reasoning and study success also could be impaired by an unwary work style.

For the interrelations between personality scales and interest however, that does not seem appropriate; here other explanations (e.g. social desirability when answering questions) seem more supposable.

To summarize it can be said, that both cognitive as well as non-cognitive predictors might play an important role for academic success. As the present study might indicate, there is not much variance regarding cognitive predictors, so that it is especially the non-cognitive traits (including work styles) that might be of importance. Since users of self-assessment have no need for faking good, questionnaires are as well applicable as experimental-based behavior tasks. However, that does not imply that cognitive predictors can be ignored, since users of self-assessment might benefit from useful feedback. When doing so, feedback about cognitive predictors should be extremely careful for not discouraging qualified applicants for a field of study.

References:

Busato, V.V., Prins, F.J., Elshout, J.J. & Hamaker, C. (2000). Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Personality and Individual Differences, 29,* 1057–1068.

Chamorro-Premuzic, T. & Furnham, A. (2003). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality, 37,* 319-338.

Conard, M.A. (2006). Aptitude is not enough: How personality and behavior predict academic performance. *Journal of Research in Personality, 40,* 339-346.

Heublein, U., Hutzsch, C., Schreiber, J., Sommer, D. & Besuch, G. (2009*). Ursachen des Studienabbruchs in Bachelor- und herkömmlichen Studiengängen. Ergebnisse einer bundesweiten Befragung von Exmatrikulierten des Studienjahres 2007/2008*. *[Reasons for study drop-outs in bachelor- and traditional courses. Results of a nationwide exmatriculates' survey of the study year 2007/2008.]* Projektbericht Dezember 2009. HIS: Hannover. Zuletzt zugegriffen am 6. Jänner 2012 unter http://www.his.de/pdf/21/studienabbruch_ursachen.pdf [Retrieved January 6, 2012 from http://www.his.de/pdf/21/studienabbruch_ursachen.pdf].

Khorramdel, L. & Maurer, M. (2012). Das Wiener Studieneignungs-Persönlichkeitsinventar. [Viennese Study-Qualification Personality Inventory.] In K.D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests") (Hrsg.), *Self-Assessment: Theorie und Konzepte* (in press). Lengerich: Pabst.

Kubinger, K.D. (2003). Gütekriterien. [Quality criteria.] In K.D. Kubinger & R.S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 195–204). Weinheim: Beltz/PVU.

Kubinger, K.D. (2009a). *Psychologische Diagnostik – Theorie und Praxis psychologischen Diagnostizierens [Psychological Assessment – Theory and practice of psychological assessment]* (2., überarb. u. erweiterte Aufl.). Göttingen: Hogrefe.

Kubinger, K.D. & Ebenhöh, J. (2012). Experimentalpsychologische Verhaltensdiagnostik des Anspruchsniveaus: Der Untertest Symbole Kodieren aus den „Arbeitshaltungen". [Experimental-based behavior assessment of goal-setting: The subtest symbol coding from the „Arbeitshaltungen".] In K.D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests") (Hrsg.), *Self-Assessment: Theorie und Konzepte* (in press). Lengerich: Pabst.

Kubinger, K.D., Frebort, M. & Müller, C. (2012). Self-Assessment im Rahmen der Studienberatung: Möglichkeiten und Grenzen. [Self-assessment in context of study advice: Possibilities and risks.] In K.D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests") (Hrsg.), *Self-Assessment: Theorie und Konzepte* (in press). Lengerich: Pabst.

Kubinger, K.D., Haiden, A., Karolyi, M. & Maryschka, C. (2012). Diagnostik des Lernstils: Der Lerntest LAMBDA. [Assessment of learning style: The learning test LAMBDA.] In K.D. Kubinger, M.

Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (in press). Lengerich: Pabst.

Kubinger, K.D., Moosbrugger, H., Frebort, M., Jonkisz, E. & Reiß, S. (2007). Die Bedeutung von Self-Assessments für die Studienplatzbewerbung. [The importance of self-assessments for study application.] *Report Psychologie, 32,* 322–332.

Mitschek, B. (2011). *Ermittlung prototypischer Testkennwerte im Wiener Self-Assessment Maschinenbau anhand erfolgreicher Studierender. [Detection of prototypic test values in the Viennese Self-Assessment Engineering on the basis of successful students.]* Unveröffentlichte Diplomarbeit: Universität Wien.

Müller, C.E. (2011). *Ermittlung prototypischer Testkennwerte für das Wiener Self-Assessment für Architektur anhand erfolgreich und wenig(er) erfolgreich Studierender. [Detection of prototypic test values for the Viennese Self-Assessment Architecture on the basis of successful and less successful students.]* Unveröffentlichte Diplomarbeit: Universität Wien.

O’Connor, M.C. & Paunonen, S.V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences*, *43*, 971–990.

Sandner, N. (2008). Möglicher Nutzen nicht-kognitiver Merkmale bei der universitären Auswahl, Beratung und Profilbildung. [Possible benefits of non-cognitive traits at universitarian selection, advice and profile formation.] In H. Schuler & B. Hell (Hrsg.), *Studierendenauswahl und Studienentscheidung* (S. 129–138). Göttingen: Hogrefe.

Undeutsch, N. (2012). Schlussfolgerndes Denken figural: Der Färbige Matrizentest. [Figural reasoning: The colorful matrices test.] In K.D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (in press). Lengerich: Pabst.

Vásquez, S.M., de Anglat, H.D. (2009). Academic achievement and formal thought in engineering students. *Electronic Journal of Research in Educational Psychology, 7,* 653–672.

Weitensfelder, L. (2012). Test zur Angewandten Raumvorstellung. [Test of Applied Relations and Visuospatial Abilities.] In K.D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (in press). Lengerich: Pabst.

Weitensfelder, L. & Becker, F. (2012). Optische Differenzierungs- und Gliederungsfähigkeit: Der Test Figurales Synthetisieren. [Optical differentiation and visual structuring abilities: The test Figural Synthesizing.] In K.D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (in press). Lengerich: Pabst.

Weitensfelder, L., Undeutsch, N., Khorramdel, L. & Useini, C. (2012). Intrinsische Studienmotivation: Interesse, Erwartungen und Selbstkonzept eigener Fähigkeiten. [Intrinsic study motivation: Interest, expectations and self-concept of abilities.] In K.D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (in press). Lengerich: Pabst.

Zimmerhofer, A., Heukamp V. & Hornke, L. (2006). Ein Schritt zur fundierteren Studienfachwahl. Webbasierte Self-Assessments in der Praxis. [A step to a well-founded choice of a field of study. Web-based self-assessments in practice.] *Report Psychologie, 31*, 62-72.

Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Conception of a facet-oriented spatial ability test (TARV – Test of Applied Relations and Visuo-spatial abilities)

Lisbeth Weitensfelder

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology, University of Vienna



Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Content

1. Spatial abilites and spatial facets
2. A new idea:
   Conception of the TARV and it's facets
3. TARV-Items (examples)
4. Results



1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-Items
4. Results



Spatial abilities

„ability to generate, retain, retrieve and transform well-structured images" (Lohman, 1996)

- No unitary construct:
  - Thurstone: defined space 1938 as one factor, but subdivided it 1950 (Rost, 1976)
  - Subclassification as well by Lohman (1979), Linn & Petersen (1985)
- So far: No consistent definition, which and how many facets are necessary to define the construct of spatial abilities

Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-Items
4. Results



Examples for spatial abilities in literature

Thurstone's facets (1950)	Lohman's facets (1979)	Linn & Petersen's facets (1985)
Spatial Relations (Imagining a configuration that is moved into different positions)	Spatial Relations (Speeded Rotation) (speed at simple rotation tasks)	Mental Rotation (rather speed than accuracy of rotation)
Visualization (Imagination how parts of a configuration are adjourned)	Visualization (complex tasks without time pressure)	Spatial Visualization (complicated, multistep tasks, e.g. paper folding)
Spatial Orientation (Imagination where self in space would be placed)	Spatial Orientation (Imagination how an object looks from another side)	Spatial Perception (determation of spatial relationships with reference to the own body)

Examples for spatial abilities in literature

Thurstone's facets (1950)	Lohman's facets (1979)	Linn & Petersen's facets (1985)
Spatial Relations (Imagining a configuration that is moved into different positions)	Spatial Relations (Speeded Rotation) (speed at simple rotation tasks)	Mental Rotation (rather speed than accuracy of rotation)
Visualization (Imagination how parts of a configuration are adjourned)	Visualization (complex tasks without time pressure)	Spatial Visualization (complicated, multistep tasks, e.g. paper folding)
Spatial Orientation (Imagination where self in space would be placed)	Spatial Orientation (Imagination how an object looks from another side)	Spatial Perception (determation of spatial relationships with reference to the own body)

1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-Items
4. Results



→ even with similar labellings, the facets content different abilities and the demarcation between the facets is not always clear

Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Common tests for spatial abilities

- Usually tests base on a single facet of spatial ability, e.g.:
  - 3DW (Gittler, 1990), MRT (Vandenberg & Kuse, 1978): mental rotation
  - Schlauchfiguren (Stumpf & Fay, 1983): orientation
- Measurement of one facet is justifiable, if prognostic validity of the test is sufficiently high
- But: Some jobs require more than one facet of spatial abilities (e.g. pilots)
- Different facets show different gender effects

1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-Items
4. Results



Development of a new test: TARV

- Idea:
  - Developing a test that contains more than mental rotation
  - Clear demarcation between spatial facets

    → adapted definition of spatial facets

1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-items
4. Results



Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Development of a new test: TARV

- 3 different facets of spatial abilities
  (based on an adaption of Thurstone's and Lohman's subclassifications):
  - Relations: identification of size and distance proportions
  - Rotation: mental rotation of objects or parts of objects
  - Orientation: location (esp.: direction), where an object or self is placed

1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-items
4. Results



TARV-facets when parking a car

- **Relations** (of size and distance):
  Looking for parking slots: „Does the car fit in the parking slot?"
- **Rotation:**
  Parking: „I know the car fits in this slot, but to get in there, I have to move backwards and turn in correctly."
- **Orientation:**
  Getting back: „I know I parked the car about a hundred metres away… But in which direction?"

1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-items
4. Results



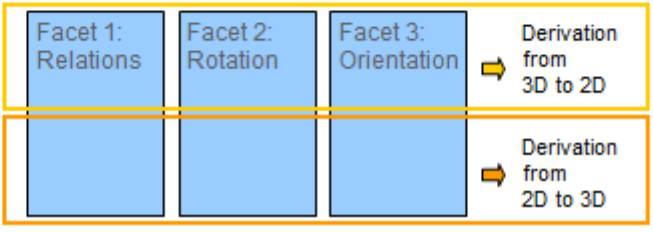
Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

TARV design: Facets and items

Subjects have to perform two different sorts of tasks:

- Derivation from 3D to 2D: 2dimensional plans have to be compared with a given 3D-object
- Derivation from 2D to 3D: A 3dimensional object has to be compared with given 2D-plans



1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-items
4. Results

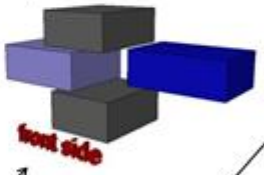


TARV-Items (examples): Derivation from 3D to 2D

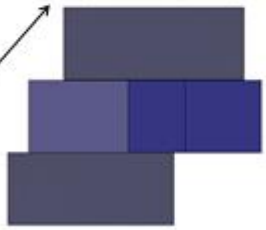
- Subjects have to compare a plan with an object and decide, whether the plan fits to the object or not. (Here: the given plan does fit.)

1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-Items
4. Results





perspective of the plan: right side



A 3dimensional lettering allows orientation.
An instruction phase clarifies sides.

- The given plan is correct.
- The given plan is not correct.

Centre of Testing and Consulting

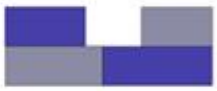
Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

TARV-Items (examples): Derivation from 2D to 3D

1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-Items
4. Results

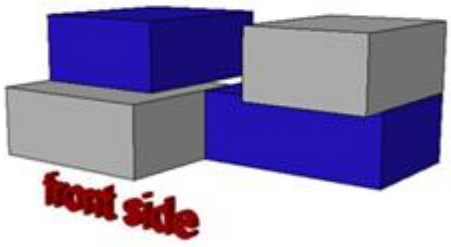


perspective of the plan: front side



perspective of the plan: top





- Compared to the two plans, the given object is correct.
- Compared to the two plans, the given object is not correct.

Sequential presentation of the items

- Original version: Subjects had to answer 3 plans for every object in a row
- An item is only considered correct if all 3 answers for every object are correct
- But: time consuming, bottom effects (difficult items)
- → now just 2 plans for every object that both have to be correct (guessing probability: .25)

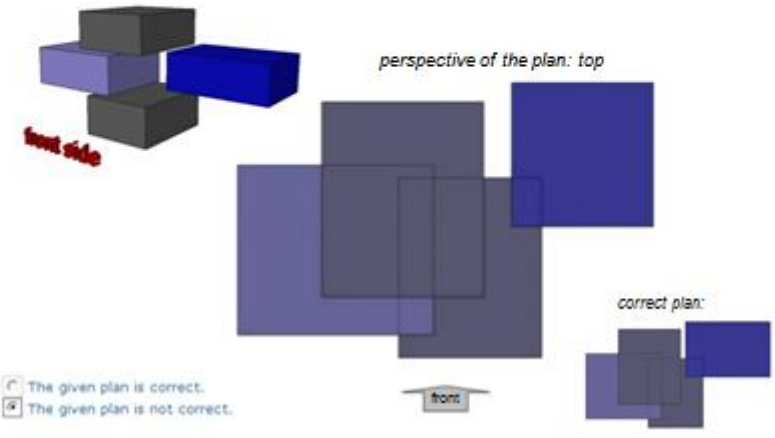
1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-items
4. Results



Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Item Example: Relations



1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-items
4. Results



Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

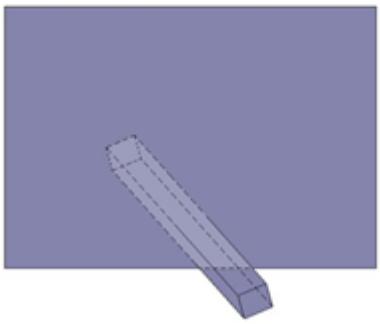
Item Example: Rotation

1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-Items
4. Results





perspective of the plan: bottom



(•) The given plan is correct.
( ) The given plan is not correct.

Incorrect perspectives for rotation items often content incorrect angles between one object and another. E.g. in the plan the green part could be tilted differently than in the object.

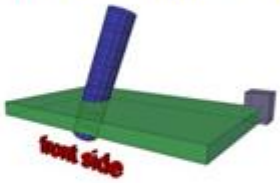
Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

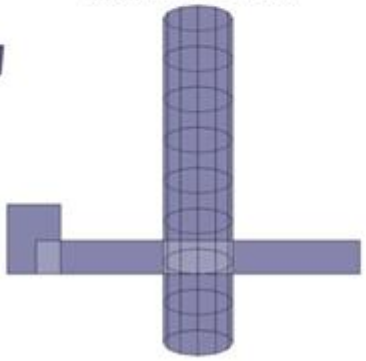
Item Example: Orientation

1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-Items
4. Results





perspective of the plan: right side



correct plan:



( ) The given plan is correct.
(•) The given plan is not correct.

Results: One or more dimension?

- A Rasch-model was calculated in an incomplete design of 3 different TARV-versions with several linking items:
  - Architecture version (9 items): n = 140 (unproctored)
  - Engineering version (10 items): n = 124 (unproctored)
  - Aviation version (10 items): n = 59 (proctored)

In total: 22 different items
(15 items without the aviation version)

1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-items
4. Results



Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Results: One dimension!

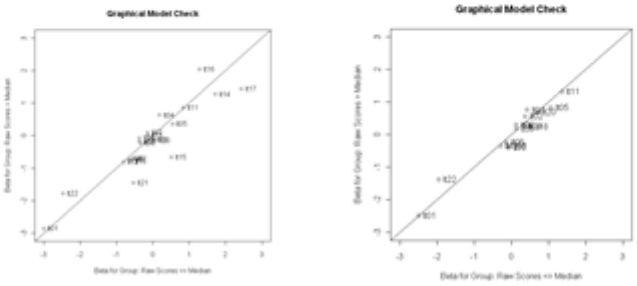
- All 3 facets as defined in the TARV belong to a single dimension according to the median split

  LR-value: 17.273, critical chi-square value: 32.671 (a=5%); p-value = .694 (with aviators);
  LR-value: 7.939, critical chi-square value: 23.685 (a=5%); p-value = .892 (without aviators)

- (Not enough demographic data for gender split.)

1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-items
4. Results





Median split for all items with (left) and without (right) aviators; the aviation version differed in the presentation mode of the items

Discussion

- Question, whether different facets contribute to different difficulty levels → future research with LLTM
  - evidence for different difficulties: coherent to literature, men scored higher in rotation of the TARV aviation version. Those effects could not be found in the other TARV versions.
- Regarding TARV results, the spatial facets form one dimension. → Spatial abilites as measured in the TARV consist of more than e.g. rotation, but it still seems to be one ability!
  - eventually comparable to mathematics: calculation is more than multiplications
- Time-consuming sequential items → currently: work on a multiple-choice version

1. Spatial abilities & spatial facets
2. Conception of the TARV
3. TARV-items
4. Results



Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Thank you!

http://psychologie.univie.ac.at/diagostik
www.testzentrum.at

lisbeth.weitensfelder@univie.ac.at



References

- Gittler, G. (1990). *Dreidimensionaler Würfeltest (3DW). Ein rasch-skalierter Test zur Messung des räumlichen Vorstellungsvermögens.* Weinheim: Beltz.
- Gittler, G. & Arendasy, M. (2002). Endlosschleifen. Psychometrische Grundlagen des Aufgabentyps **EF**. *Diagnostica, 48*, 164–175.
- Gruber, M. (2008). *Einsatz- und Anwendungsmöglichkeiten des 3DW und des EST bei der Selektion und Leistungsüberprüfung von Flugleitern und Lotsenpiloten.* Unpublished diploma thesis, University of Vienna.
- Grubesic, A. (2010). *Raumvorstellung aus Sicht des TAKV-F 4.0 und EST. Eine Validierung anhand von Berufsgruppen.* Unpublished diploma thesis, University of Vienna.
- Linn, M. C. & Petersen, A.C. (1985). Emergence and Characterization of Sex Differences in Spatial Ability: A Meta-Analysis. *Child Development, 56*, 1479–1498.
- Lohman, D.F. (1979). *Spatial Ability: A Review and Reanalysis of the Correlational Literature.* Technical Report No. 8, Aptitude Research Project, School of Education, Stanford University.
- Lohman, D. F. (1996). Spatial Ability and G. In Ian Dennis & Patrick Tapsfield (Eds), *Human Abilities: Their Nature and Measurement* (pp. 97–116). Hillsdale, NJ: Erlbaum.
- Rost, D. (1976). *Der Begabungsfaktor „Raumvorstellung": Theorie und Training.* Dissertation, University of Hamburg.
- Stumpf, H. & Fay, E. (1983). *Schwierigkeiten. Ein Test zur Messung des räumlichen Vorstellungsvermögens.* Göttingen: Hogrefe.
- Vandenberg, S.G. & Kuse, A.R. (1978). Mental Rotations, A Group Test of Three-Dimensional Spatial Visualization. *Perceptual & Motor Skills 47*, 599–604.
- Voyer, D., Voyer, S. & Bryden, M.P. (1995). Magnitude of Sex Differences in Spatial Abilities: A Meta-Analysis and Consideration of Critical Variables. *Psychological Bulletin, 117(2)*, 250–270.



Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Processing Characteristics and Reasonableness Regarding Self-Assessments

Lisbeth Weitensfelder
Martina Frebort
Christine E. Müller
Barbara Mitschek



Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology, University of Vienna

Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Content

- Responders vs. non-responders: Are they comparable at all?
- Processing self-assessments: Characteristics and cancelling data
- Cancelling self-assessments: A closer look at the cancellers
- Cancelling self-assessments: A closer look at the tests and resulting ideas on reasonableness



Responders vs. non-responders (1)

- Voluntary participants might differ in several apsects from the population in question (e.g. Rosenthal & Rosnow, 1975; Karner, 2002; Porter & Whitcomb, 2005)

  → Question, whether students using the offer of a self-assessment are comparable with students in general (e.g. Kubinger et al., 2007).

- Differences between responders and non-responders couldn't always be reproduced (Leitner, 2008; Unterhofer, 2009: comparable results of early and late responsers)
- Müller (in prep.) and Mitschek (2011) compared responders and non-responders for the Viennese self-assessments architecture and engineering

1. Responders vs. non-responders
2. Processing self assessments
3. Cancelling: the cancellers
4. Cancelling: the tests
5. Discussion

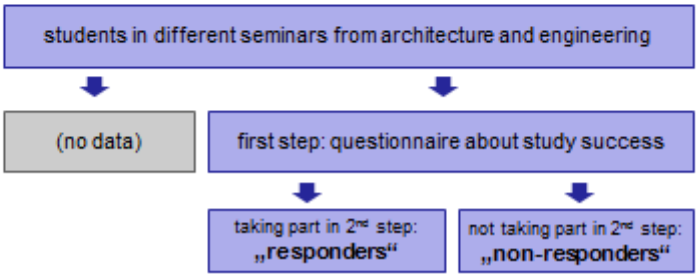


Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Responders vs. non-responders (2)

- Test design: Students in different seminars were asked to participate in a validation study for the self-assessments
  - Before attending the study a questionnaire about study success (and demographic data) had to be filled out



1. Responders vs. non-responders
2. Processing self assessments
3. Cancelling: the cancellers
4. Cancelling: the tests
5. Discussion



Responders vs. non-responders (3)

1. Responders vs. non-responders
2. Processing self assessments
3. Cancelling: the cancellers
4. Cancelling: the tests
5. Discussion



- Results: no differences regarding study succes (Welch-tests, α = .05); success criteria: average studying speed

	architecture – study success in ECTS/sem*	engineering – study success in hours/sem*
responders	n = 59; mean: 14.14 (SD: 5.2)	n = 68; mean: 12.73 (SD: 6.59)
non-responders	n = 67; mean: 13.47 (SD: 5.49)	n = 13; mean: 12.3 (SD: 4.02)

*) adjusted data; students in their first semester and students with unplausible answers were excluded from the success criteria

- No hint, that non-responders might be worse than responders
  → data of voluntary participants seems valid

Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Viennese self-assessments: Processing data and test abortions

1. Responders vs. non-responders
2. Processing self assessments
3. Cancelling: the cancellers
4. Cancelling: the tests
5. Discussion



Sample:

- Self-assessment psychology:
  n = 538 anonymous participants (61% completed)
- Self-assessment architecture:
  n = 228 anonymous participants (42% completed)*
- Self-assessment engineering:
  n = 109 anonymous participants (39% completed)*

*) abortion rates of volunteers (=„responders") who took part in the validation study were much lower (see Müller, in prep.; Mitschek, 2011).

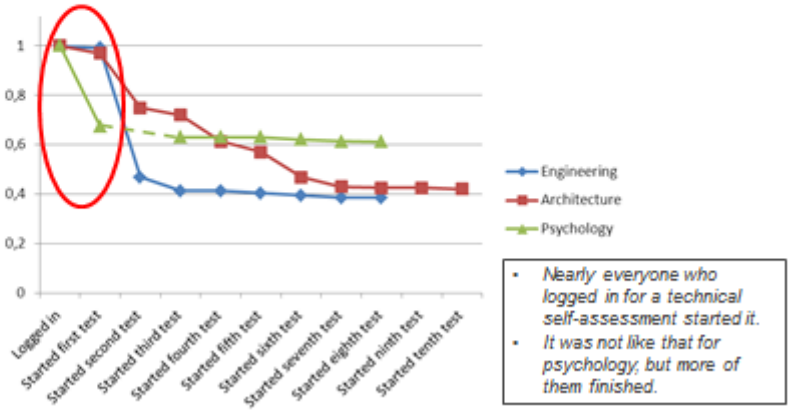
Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

1. Responders vs. non-responders
2. Processing self assessments
3. Cancelling: the cancellers
4. Cancelling: the tests
5. Discussion



Participation process in the Viennese self-assessments (1)



- *Nearly everyone who logged in for a technical self-assessment started it.*
- *It was not like that for psychology, but more of them finished.*

Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

1. Responders vs. non-responders
2. Processing self assessments
3. Cancelling: the cancellers
4. Cancelling: the tests
5. Discussion



Participation process in the Viennese self-assessments (2)



Starting page of the self-assessment psychology (left) and technical self-assessments (below)



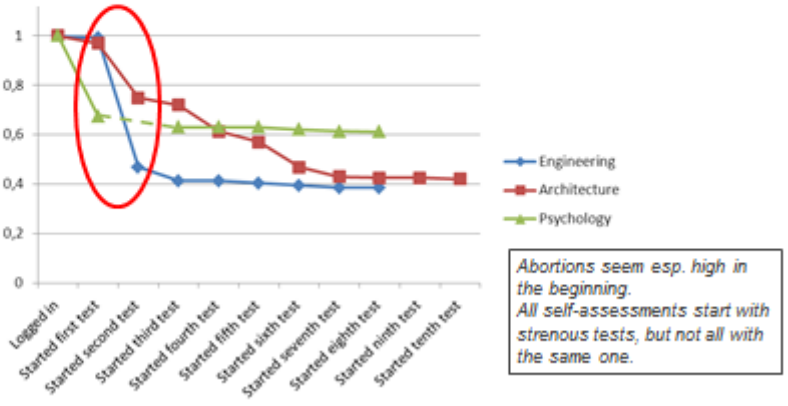
Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Participation process in the Viennese self-assessments (3)

1. Responders vs. non-responders
2. Processing self assessments
3. Cancelling: the cancellers
4. Cancelling: the tests
5. Discussion





Abortions seem esp. high in the beginning.
All self-assessments start with strenous tests, but not all with the same one.

Centre of Testing and Consulting

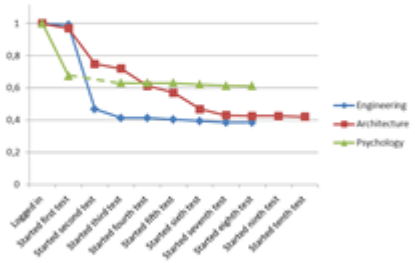
Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Cancelling self-assessments: A closer look at the cancellers

1. Responders vs. non-responders
2. Processing self assessments
3. Cancelling: the cancellers
4. Cancelling: the tests
5. Discussion



- As the graphs before (and right) showed, only between 39 and 61% of users who logged in finished the self-assessment.



- Question: Which users do abort?
- Hypothesis: Users who do abort a test do worse in it than users who don't.

Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Cancelling self-assessments: The users (1)

- In general: Test aborters do worse in that test.
- Table: cancellers in the matrices test (architecture) vs. non-cancellers; Welch-Test, α = .05)

Architecture – percentage of matrices items solved relative to items executed	
non-cancellers (n=175); mean = 42,39; SD = 26,83	**cancellers (n=35);** mean = 29,64; SD = 24,05

- Differences between cancellers an non-cancellers might lay within abilities and/or within effort.

1. Responders vs. non-responders
2. Processing self assessments
3. Cancelling: the cancellers
4. Cancelling: the tests
5. Discussion



Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Cancelling self-assessments: The users (2)

- People who abort the self-assessment in a later stage attain lower scores in the matrices test than people who do not cancel (Welch-Test, α = .05).

Architecture – number of matrices items solved (percentage of correct items)	
Self-assessment completed (n=96): mean = 7.64 (47.72% correct); SD = 4.33 (27.11%)	**Self-assessment cancelled later than in the matrices test (n=78):** mean = 5.78 (36.14% correct) SD = 4.05 (25.31%)

- → Even if subjects cancel in a later test, they score lower in the earlier matrices test.

  → Compared to their effort, the tests might be too hard for cancellers.

1. Responders vs. non-responders
2. Processing self assessments
3. Cancelling: the cancellers
4. Cancelling: the tests
5. Discussion



Cancelling self-assessments: A closer look at the tests

- Does abortion have to do with characteristics of the tests themselves…
- …and/or the position in the test battery?

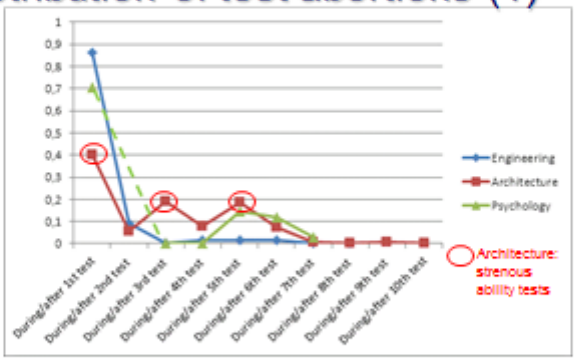
1. Responders vs. non-responders
2. Processing self assessments
3. Cancelling: the cancellers
4. Cancelling: the tests
5. Discussion



Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Distribution of test abortions (1)



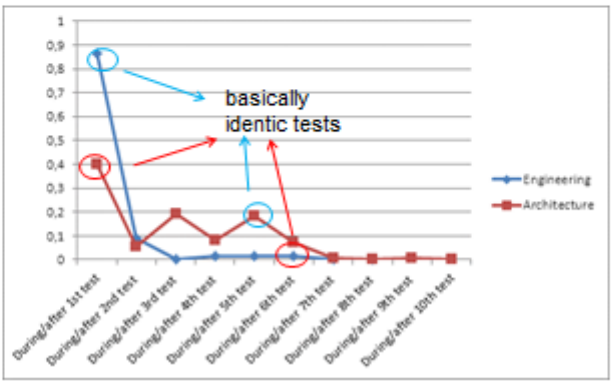
1. Responders vs. non-responders
2. Processing self assessments
3. Cancelling: the cancellers
4. Cancelling: the tests
5. Discussion



- Most abortions in architecture happen at strenous tests.
- Strenous ability tests also got the highest test-specific amout of critical feedback in a qualitative survey after the validation study (Müller, in prep.).
- Hardly any negative feedback for questionnaires, none for experimental behavior tasks.

Distribution of test abortions (2)

- Architecture and engineering share basically most of their tests, but in different test order.
- The first test in engineering (and psychology) is a learning test with high abortion rates. This test is the fifth one in architecture.



1. Responders vs. non-responders
2. Processing self assessments
3. Cancelling: the cancellers
4. Cancelling: the tests
5. Discussion

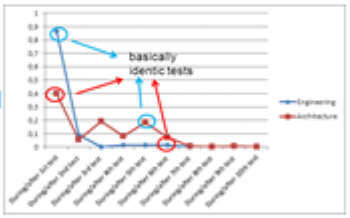


Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Test abortion positions and consequences on reasonableness

- Strenous procedures are more likely to be aborted.
- Tests in the beginning are aborted most often.
- If the first test is strenous, but not the hardest one: Later abortion rates are higher.



- In total, abortion rates in architecture and engineering seem to be the same (~60%).
- A „high hurdle" criteria seems more reasonable (least time-consuming for cancellers), unless cancellers can benefit from earlier test.

1. Responders vs. non-responders
2. Processing self assessments
3. Cancelling: the cancellers
4. Cancelling: the tests
5. Discussion



Discussion

- Responders and non-responders seem to be comparable regarding study success.
- However, there are differences between cancellers and non-cancellers: Cancellers do worse in the tests.
- Strenous procedures are cancelled more often and seem to be less reasonable than questionnaires or experimental behavior tasks sensu Cattell.
- First tests are cancelled most often. If the first test is very strenous, there are even more abortions at the first test, but hardly any abortions later.

1. Responders vs. non-responders
2. Processing self assessments
3. Cancelling: the cancellers
4. Cancelling: the tests
5. Discussion



Centre of Testing and Consulting

Division of Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna

Thank you!

http://psychologie.univie.ac.at/diagostik
www.testzentrum.at

lisbeth.weitensfelder@univie.ac.at



References

- Karner, T. (2002). The volunteer effect of answering personality questionnaires. *Psychologische Beiträge, 44*, 42–49 .
- Kubinger, K.D., Moosbrugger, H., Frebort, M., Jonkisz, E. & Reiß, S. (2007). Die Bedeutung von Self-Assessments für die Studienplatzbewerbung. *Report Psychologie, 32*, 322–332.
- Leitner, A. (2008). *Der Einfluss der Freiwilligkeit auf die Normierung/Eichtabellen*. Unpublished diploma thesis, University of Vienna.
- Mitschek, B. (2011). *Ermittlung prototypischer Testkennwerte im Wiener Self-Assessment Maschinenbau anhand erfolgreich Studierender*. Unpublished diploma thesis, University of Vienna.
- Müller, C.E. (In prep). *Ermittlung prototypischer Testkennwerte im Wiener Self-Assessment Architektur anhand erfolgreich Studierender*. Unpublished diploma thesis, University of Vienna.
- Porter, S.R. & Whitcomb, M.E. (2005). Non-Response in Student Surveys: The Role of Demographics, Engagement and Personality. *Research in Higher Education, 46/2*, 127–152.
- Rosenthal, R. & Rosnow, R.L. (1975). *The volunteer subject*. New York: Wiley.
- Unterhofer, E. (2009). *Eine Non-Responder-Analyse zur Erfassung des Einflusses der Freiwilligkeit auf Testeichung und -validierung*. Unpublished diploma thesis, University of Vienna.



Anhang C:
Abstract und Zusammenfassung

Abstract (English)

Facing limited resources, universities are confronted with the challenge to assign applicants in a best possible way. One of the means universities recently use are online self-assessments to provide both an information tool as well as a tool for self-selection to possible applicants. Based on such self-assessments for technical fields of study, the present dissertation asks which abilities and traits are actually necessary to study in technical fields, and how these traits should best be assessed. Already existing requirement profiles for selected technical fields of study are presented, where it can be seen, that a majority of requirements overlaps with general requirements for studying at a university; only a few requirements can be seen as specific (e.g. spatial abilities for architectural studies). For such specific requirements, new test concepts that are already in use for self-assessments of the TU Wien are presented: A new concept for a test for spatial abilities and two field-specific interest questionnaires for architecture and mechanical engineering, with the contribution of the interest questionnaires representing manuscripts for peer-reviewed journals. The interest questionnaires contain items that aim to already inform possible study applicants about study contents. The questionnaires offer the benefit of being tailor-made for the field of study, but some adaptions are needed to make them meet psychometric requirements.

Regarding the general design of such test batteries for study counseling purposes, it is investigated in a third manuscript, how test order affects dropout rates and fatigue. Based on the results of the third contribution, it can be recommended to start self-assessments with a short expectancy questionnaire as a warming-up phase and put the highest hurdle afterwards. In contrast to results from literature, which focus on test situations, where the results usually have a powerful ramification for the testees, fatigue effects in the investigated self-assessment for architecture could be found.

Zusammenfassung (Deutsch)

Im Zuge der Studienplatzbewirtschaftung stellt sich für Hochschulen die Frage, wie denn knappe Studienplätze bestmöglich zu vergeben sind. Als eines unter vielen Mitteln setzen Hochschulen verstärkt auch auf Self-Assessments zur Information und Selbstselektion von Studieninteressierten. Im Rahmen derartiger Self-Assessments für technische Studienrichtungen widmet sich die vorliegende Dissertation der Frage, welche Anforderungen für technische Studienrichtungen denn überhaupt benötigt werden, und wie diese bestmöglich zu erfassen seien. Am Beispiel von an der TU Wien eingesetzten Self-Assessment-Testbatterien zur Studienwahlberatung werden zunächst (bestehende) Anforderungsprofile für ausgewählte technische Studienrichtungen dargestellt. Dabei zeigt sich, dass – bis auf wenige Spezialfähigkeiten für einzelne Studienrichtungen, wie z.B. Raumvorstellung für das Architekturstudium – großteils die selben Anforderungen relevant sind wie für allgemeine Studierfähigkeit. Im Anschluss wird auf die Erfassung einiger der spezifischen Anforderungen eingegangen, indem neue Testkonzepte präsentiert werden: Ein neues Konzept zu einem Raumvorstellungstest sowie die Gestaltung maßgeschneiderter Interessenfragebögen für die intendierte Studienrichtung (hier an den Beispielen für Architektur bzw. Maschinenbau). Die Interessenfragebögen werden auch in eigenen Fachzeitschriftenartikeln näher beleuchtet, in denen dargestellt wird, wie Fragebögen mit derart „praktisch“ gewonnenen Iteminhalten, die gleichzeitig einen konkreten Einblick in die Studienrichtung vermitteln sollen, gestaltet werden können, um dennoch psychometrisch faire Ergebnisse rückzumelden: Letztlich sind Fragebögen, die maßgeschneidert konkrete Studieninhalte abbilden sollen, wohl ökologisch valide, es benötigt aber noch einiger Anpassungen, um die Erfassung der Interessen so zu gestalten, dass das Testergebnis ein gänzlich faires Maß darstellt.

In Hinblick auf die Gestaltung einer Testbatterie zu Studienberatungszwecken in ihrer Gesamtheit wird zudem mit einem dritten Fachzeitschriftenartikel abgehandelt, wie sich unterschiedliche Vorgabereihenfolgen auf Abbrecherquoten und Leistungseinbußen in Self-Assessments auswirken. Darauf basierend wird empfohlen, Self-Assessments mit einer Erwartungsabklärung als Aufwärmphase sowie im Anschluss mit der höchsten Hürde zu beginnen: Im Gegensatz zu aus der Literatur berichteten nicht-vorhandenen Ermüdungseffekten bei konsequenzbehafteten Testsituationen konnten bei dem untersuchten Self-Assessment für Architektur nämlich durchaus Leistungseinbußen im Verlauf beobachtet werden.

Lebenslauf

geb. am 11.05.1982 in Klagenfurt (Staatsbürgerschaft: Österreich)

Ausbildung

seit 2009	Doktoratsstudium der Naturwissenschaften (Psychologie)
2016	Zertifizierung als Arbeits- und Organisationspsychologin
2015	Eintragung in die Liste der Klinischen und GesundheitspsychologInnen
2011	Abschluss des Diplomstudiums Betriebswirtschaft an der WU Wien
2008	Abschluss des Diplomstudiums Psychologie an der Universität Wien
2008	Erasmussemester in Lissabon, Portugal

Berufslaufbahn

seit 2017	freie Dienstnehmerin an der TU Wien im Rahmen eines Self-Assessment-Projekts
seit 2015	wissenschaftliche Projektmitarbeiterin an der Universitätsklinik für Psychiatrie und Psychotherapie, Medizinische Universität Wien
seit 2014	Abhaltung von Gastvorträgen für die FH Technikum Kärnten (Studiengang Health Care IT)
2016-2017	freiberuflich in der Praxis für Neurorehabilitation, 1150 Wien
2014-2015	Fachausbildungsstelle für Klinische Psychologie an der Universitätsklinik für Psychiatrie und Psychotherapie, AKH Wien
2012-2014	Mutterschutz und Kinderbetreuungszeit
2009-2014	Lehrbeauftragte für Psychologische Diagnostik an der Universität Wien
2009-2012	Mitarbeiterin an der Test- und Beratungsstelle des Arbeitsbereichs Psychologische Diagnostik, Fakultät für Psychologie, Universität Wien
2009-2010	Vorstandsmitglied im Verein „Grenzenlos“, www.grenzenlos.or.at
2007-2008 (Sommer)	Mitarbeit und Coleitung bei einem Workcamp für einen Kinderbauernhof zur Betreuung obdachloser Kinder über „Grenzenlos“
2007 (Sommer)	Teilnahme an einem Umweltprojekt-Workcamp in Willits, CA (USA)
2007-2009	Studienassistentin am Arbeitsbereich Psychologische Diagnostik, Fakultät für Psychologie, Universität Wien (Unterbrechung während des Erasmussemesters)
2006-2009	Werkvertragsnehmerin an der Test- und Beratungsstelle des Arbeitsbereichs Psychologische Diagnostik, Fakultät für Psychologie, Universität Wien
2006 (Sommer)	Volunteertätigkeit bei der „Victoria Immigrant and Refugee Centre Society“ in Victoria, BC (Kanada)
2002-2012	Qualitätsbeauftragte für die Ordination Dr. Weitensfelder, Klagenfurt
1999,2002	Verfassen und Herausgabe zweier Lyrik-Bände

Wissenschaftliche Publikationen

Bücher und Buchbeiträge

Frebort, M., Gleeson, R., & Weitensfelder, L. (2012). Wissenstests zur Erfassung des bereichspezifischen Vorwissens. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 83-91). Lengerich: Pabst.

Hofer, S., & Weitensfelder, L. (2012). Herkömmliche Erhebungstechniken in Self-Assessments zur Studienberatung. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 63-69). Lengerich: Pabst.

Kubinger, K. D., Frebort, M., Khorramdel, L. & Weitensfelder, L. („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.) (2012). *Self-Assessment: Theorie und Konzepte.* Lengerich: Pabst.

Scheck, T., Becker, F., Weitensfelder, L., & Frebort, M. (2012). Rahmenbedingungen von Self-Assessments zur Studienwahlberatung. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 25-48). Lengerich: Pabst.

Weitensfelder, L. (2012). Test zur Angewandten Raumvorstellung. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 181-195). Lengerich: Pabst.

Weitensfelder, L., & Becker, F. (2012). Optische Differenzierungs- und Gliederungsfähigkeit: Der Test Figurales Synthetisieren. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 207-217). Lengerich: Pabst.

Weitensfelder, L., & Hofer, S. (2012). Verstehen von Texten: Leseverständnistest für Technik-Studierende. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 197-205). Lengerich: Pabst.

Weitensfelder, L., Undeutsch, N., Khorramdel, L., & Useini, C. (2012). Intrinsische Studienmotivation: Interesse, Erwartungen und Selbstkonzept eigener Fähigkeiten. In In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 119-142). Lengerich: Pabst.

Zeitschriften

Kubinger, K. D., Frebort, M., Khorramdel, L. & Weitensfelder, L. (2013). Prinzipien und Verfahren der Self-Assessments vom „Wiener Autorenkollektiv Studienberatungstests“. *Wirtschaftspsychologie, 15,* 17-24.

Kubinger, K. D., Frebort, M., Khorramdel, L., Weitensfelder, L., Sonnleitner, P., Hohensinn, C., Reif, M., Gruber K. & Holocher-Ertl, S. (2008). Large Scale Assessment at the Austrian Educational Standards: a Review. *Testing International, 19, 15-16*.

Kubinger, K. D., Frebort, M., Holocher-Ertl, S., Khorramdel, L., Sonnleitner, P., Weitensfelder, L., Hohensinn, C. & Reif, M. (2007). Large-Scale Assessment zu den Bildungsstandards in Österreich: Testkonzept, Testdurchführung und Ergebnisverwertung. *Erziehung und Unterricht, 7-8*, 588-599.

Weitensfelder, L. (in press). Test Order Effects in an Online Self-Assessment: An Experimental Study. *Psychological Test and Assessment Modeling.*

Vorträge

Kubinger, K. D., Weitensfelder, L., Frebort, M., & Sonnleitner, P. (2012). *Non-cognitive vs. cognitive predictors for academic success – opportunities and limits of self-assessments illustrated by the Viennese products.* AERA [American Educational and Research Association] Annual Meeting, Vancouver, April 13th - April 17th 2012.

Weitensfelder, L., Frebort, M., Müller, C. E., & Mitschek, B. (2011). *Processing Characteristics and Reasonableness regarding self-assessments.* 11th European Conference on Psychological Assessment, Riga, August 31st - September 3rd 2011.

Weitensfelder, L. (2011). *Conception of a facet-oriented spatial ability test (TARV – Test of Applied Relations and Visuo-spatial abilities).* 11th European Conference on Psychological Assessment, Riga, August 31st - September 3rd 2011.

Weitensfelder, L., Grubesic, A., & Kubinger, K.D. (2010). *Ein facettenorientierter Raumvorstellungstest: Gendereffekte.* 47. Kongress der Deutschen Gesellschaft für Psychologie, Bremen, 26.-30. September 2010.

Arbeitsberichte

Weitensfelder, L., Kubinger, K. D., Frebort, M. & Holocher-Ertl, S. (2007). *Forschungsbericht: Dokumentation der webbasierten Rückmeldung der Ergebnisse in den Standard-Tests zur Überprüfung der Bildungsstandards für Mathematik, Deutsch und Englisch am Ende der 4. und 8. Schulstufe.*Forschungsberichte der *Test- und Beratungsstelle* des Arbeitsbereich Psychologische Diagnostik, Heft 11. Wien: Fakultät für Psychologie, Universität Wien.

Tests

Weitensfelder, L. (2013). *FigS Figurales Synthethisieren.* Mattersburg: PT Verlag.

Weitensfelder, L. (2013). *TARV Test zur Angewandten Raumvorstellung.* Mattersburg: PT Verlag.

Weitensfelder, L. & Ztrivas, S. (2013). *LESE-VT Leseverständnistest für Technik-Studierende.* Mattersburg: PT Verlag.