Test-wiseness: ein unterschätztes Konstrukt?

Empirische Befunde zur Überprüfung und Erlernbarkeit von test-wiseness

Gun-Brit Thoma · Olaf Köller

Angenommen: 17. Januar 2018 / Online publiziert: 8. Februar 2018
© Der/die Autor(en) 2018. Dieser Artikel ist eine Open-Access-Publikation.

Zusammenfassung Test-wiseness gilt bereits seit langem als wichtiger Einflussfaktor auf die Ergebnisse von Multiple-Choice Tests. So scheinen Personen mit Wissen über test-wiseness besser in Multiple-Choice Tests abzuschneiden als Personen ohne dieses Wissen. Darüber hinaus konnte gezeigt werden, dass test-wiseness trainiert bzw. erlernt werden kann. Obwohl test-wiseness für das Abschneiden in einem Multiple-Choice Test von entscheidender Bedeutung zu sein scheint, stammen die vorhandenen Erkenntnisse und Befunde fast ausschließlich aus dem amerikanischen Sprachraum. Im deutschsprachigen Raum dagegen liegen kaum Befunde zu test-wiseness vor. Vor diesem Hintergrund wurde anhand einer englischen Testversion ein deutschsprachiger Test entwickelt. Ziel der vorliegenden Studie war es, den Einfluss thematischen Wissens sowie den Einfluss einer Schulung zu test-wiseness auf das Testergebnis zu untersuchen. Hierzu beantworteten 252 Studierende 24 deutschsprachige Multiple-Choice Aufgaben zur Erfassung ihres Wissens über test-wiseness. Es wurde ein 2 (mit vs. ohne Expertise) $\times$ 2 (mit vs. ohne Schulung) Versuchsdesign realisiert. Die Ergebnisse zeigen, dass Personen mit thematischem Wissen bessere Ergebnisse erzielen als Personen ohne thematisches Wissen und Personen mit Schulung bessere Ergebnisse als Personen ohne Schulung. Insgesamt machen die Befunde deutlich, dass ein deutschsprachiger Test zur Erfassung von test-wiseness entstanden ist, der der Güte vorhandener internationaler Tests entspricht. Zusätzlich wird deutlich, dass auch im deutschsprachigen Sprachraum zukünftig ein stärkerer Fokus auf die Kontrolle von test-wiseness gerichtet werden sollte.

Zusatzmaterial online Zusätzliche Informationen sind in der Online-Version dieses Artikels (<https://doi.org/10.1007/s35834-018-0204-0>) enthalten.

G.-B. Thoma (✉) · O. Köller
Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN), Universität Kiel,
Olshausenstraße 62, 24098 Kiel, Deutschland
E-Mail: thoma@ipn.uni-kiel.de

Schlüsselwörter Test-wiseness · Multiple-Choice Aufgaben · Testergebnis · Testtraining

Test-wiseness: A neglected construct?

Empirical findings on testing and trainability of test-wiseness

Abstract Test-wiseness is regarded as an important factor influencing the outcome of multiple-choice tests. Findings from existing studies show that people with test-wiseness knowledge get better results in a multiple-choice test than people without this knowledge. Further, it seems that test-wiseness can be trained. Although test-wiseness is highly important for test performance, most of the existing theoretical assumptions and empirical findings on test-wiseness originate from Anglo-American regions. Few findings about test-wiseness exist in Germany. Against this background, a test in German was developed on the basis of an existing test in English to examine peoples' knowledge about test-wiseness. The aim of the present study was to investigate the influence of thematic knowledge as well as the impact of a test-wiseness training on the test result of the developed test. $N=252$ students answered 24 multiple-choice questions which were developed to determine their knowledge about test-wiseness. A 2 (with vs. without thematic knowledge) $\times$ 2 (with vs. without a test-wiseness training) factorial experimental design was realized. The results show that persons with thematic knowledge answer more multiple-choice questions correctly than persons without this knowledge. At the same time the findings point out that test-wiseness can be trained. Overall, the findings indicate that a German version of a test was developed that matches the psychometric quality of existing international tests. Furthermore, they suggest that a stronger focus should be directed to the control of test-wiseness in multiple-choice tests in German-speaking regions in the future.

Keywords Test-wiseness · Multiple-choice questions · Test result · Test coaching

1 Einleitung

Multiple-Choice Aufgaben (MCA) spielen schon lange überall dort eine bedeutende Rolle, wo es darum geht, Wissen, Lernergebnisse, Fähigkeiten und Kompetenzen zu messen (Ben-Simon et al. 1997; Chang et al. 2007). Die Bandbreite an Anwendungen ist groß. So kommen sie in internationalen Vergleichsstudien in der Bildungsforschung genauso zum Einsatz wie bei der theoretischen Führerscheinprüfung. Gründe dafür sind, dass MCA in der Umsetzung leicht handzuhaben sind und ihnen eine hohe Ökonomie und Objektivität bei der Auswertung zugeschrieben wird. Nichtsdestotrotz geht die Anwendung mit bestimmten Problemen einher, die mit der Konstruktion und Auswertung von MCA zusammenhängen, wie das Formulieren von „guten“ Distraktoren bei der Konstruktion oder das Raten bei der Auswertung. Ein wichtiger Aspekt im Zusammenhang mit MCA stellt „test-wiseness“ (TW) dar. Hierauf wird im Folgenden näher eingegangen.

Das in der Praxis am häufigsten vertretene MCA-Format ist das Single-Choice Format (Haladyna und Rodriguez 2013), das typischerweise aus einem Itemstamm, einer richtigen Antwort (Attraktor) und mindestens drei Antwortalternativen (Distraktoren) besteht. Das Testergebnis, das eine Person erzielt, soll ihren Wissensstand hinsichtlich eines bestimmten Themas (Rogers und Yang 1996) oder ihre Fähigkeiten und Kompetenzen widerspiegeln. Es wird davon ausgegangen, dass je höher die Anzahl richtig gelöster MCA dieser Person ist, desto höher der Grad ihres Wissens oder ihrer Kompetenzen. Seit Mitte des letzten Jahrhunderts wird TW immer wieder als Einflussfaktor auf die Leistung einer Person in MC-Tests betont (Edwards 2003; Millman et al. 1965; Tomkowicz und Rogers 2005). TW bezeichnet dabei die Fähigkeit einer Person, eine hohe Punktzahl in einem MC-Test zu erzielen, indem die richtige Antwort durch Nutzung spezifischer Hinweise und Charakteristika der Items identifiziert wird (Edwards 2003; Gibb 1964; Millman et al. 1965). Besonders in den USA finden sich zahlreiche Studien, in denen der Grad an TW bei Studierenden erhoben wurde, u. a. zur Überprüfung der Studierfähigkeit. Obwohl die Bedeutung von TW bei der Bearbeitung von MCA mehrfach belegt wurde, und die Durchführung weiterer Studien über TW gefordert wird (Cohen 2006), scheint die Erfassung von TW im deutschsprachigen Raum eine untergeordnete Rolle zu spielen (vgl. Brunner et al. 2007). Unseres Wissens existiert kein deutschsprachiger Test, mit dem TW erfasst werden kann. Ebenfalls ist unklar, inwieweit deutschsprachige Studierende durch Training von TW ihre Testleistungen in einem TW-Test positiv beeinflussen können. Ziel dieser Studie ist es daher, diesen Fragen mit einem neu entwickelten deutschsprachigen TW-Test nachzugehen. Vor diesem Hintergrund werden zuerst die theoretischen Erkenntnisse und empirische Ausgangslage zu TW dargestellt, um anschließend näher auf die vorliegende Studie mit ihren Befunden einzugehen.

2 Was ist TW?

Die verbreitetste Definition von TW stammt von Millman et al. (1965). Sie definieren TW als „a subject’s capacity to utilize characteristics and formats of the test and/or test taking situation to receive a high score. Test-wiseness is logically independent of the examinee’s knowledge of the subject matter for which the items are supposedly measures“ (Millman et al. 1965, S. 707). TW bezeichnet demnach die Fähigkeit, gewisse Strategien zu nutzen, um unabhängig von dem thematischen Wissen aber durch „ungewollte“ Hinweise (seitens der Testkonstruktion) in den Aufgaben möglichst viele Testaufgaben richtig zu beantworten. Inwieweit die Strategien wirklich unabhängig sind, ist weitestgehend ungeklärt. Einige Befunde zeigen, dass TW schwach bis moderat mit der kognitiven Fähigkeit korreliert (vgl. Rogers und Yang 1996). Aufbauend auf der Definition und den bis dahin vorhandenen Erkenntnissen und Befunden entwickelten Millman et al. (1965) eine „Taxonomy of Test-wiseness Principles“. Laut dieser können die verschiedenen TW-Prinzipien zwei Kategorien mit ihren jeweiligen Unterkategorien zugeordnet werden:

- I. Zweck- und konstruktionsunabhängige Prinzipien
 - A. Zeitznutzungsstrategien (z. B. bei Zeitvorgabe erst die Fragen beantworten, die wirklich gewusst werden)
 - B. Fehlervermeidungsstrategien (z. B. erst nach sorgfältigem Lesen aller Antwortmöglichkeiten eine Antwort auswählen)
 - C. Ratestrategien (z. B. raten, wenn es keine Minuspunkte für falsche Antworten gibt)
 - D. Schlussfolgerungsstrategien (z. B. Antwortmöglichkeiten, die sicher falsch sind, ausschließen)
- II. Zweck- und konstruktionsabhängige Prinzipien
 - A. Intention betreffende Strategien (z. B. den Zweck des Tests bei der Beantwortung der Fragen beachten)
 - B. Hinweisnutzungsstrategien (z. B. bestimmte Hinweiswörter bei Beantwortung nutzen)

Die erste Kategorie umfasst vom Testzweck und -konstruktion unabhängige Strategien, die in jeder beliebigen Testsituation anwendbar sind. Dagegen sind in der zweiten Kategorie Strategien zusammengefasst, die von dem Testzweck bzw. dem Testentwickler abhängig sind. Zeitznutzungs- (IA), Fehlervermeidungsstrategien (IB) und Strategien, die die Intention betreffen (IIA), helfen, kein schlechtes Testergebnis aus anderen Gründen als fehlendem thematischem Wissen zu erzielen. Rate- (IC), Schlussfolgerungs- (ID) und Hinweisnutzungsstrategien (IIB) dagegen dienen dazu, Fragen zusätzlich zu denen, die wirklich einzig aufgrund des thematischen Wissens richtig beantwortet werden können, richtig zu beantworten (Millman et al. 1965; Yang 2000). Während eines Tests können mehrere dieser Strategien angewendet werden. Beispielsweise kann eine Person zuerst einmal alle Fragen beantworten, bei denen die Person überzeugt ist, die richtige Antwort zu wissen. Anschließend, bei noch nicht beantworteten Fragen, kann die Testperson nach Hinweiswörtern in den Antworten suchen, um sie richtig zu beantworten. TW ist somit ein Faktor, der sich auf die Testergebnisse auswirkt (Sarnacki 1979), da die TW-Strategien zur richtigen Lösung führen können, obwohl die richtige Antwort nicht gewusst wird. Die Stärke dieses Einflusses ist jedoch nicht eindeutig geklärt. So finden sich Studien, bei denen TW 10 bzw. 13 % (Bruch 1981), 16 % (Fagley 1987) oder 46,4 % (Miller et al. 1990) der Varianz in Testleistungen aufklärt. Darüber hinaus wird kontrovers betrachtet, ob TW und das thematische Wissen einer Person wirklich unabhängig voneinander sind. Eine Unabhängigkeit von TW und thematischem Wissen wird nach Rogers und Bateson (1991) durch Befunde von Studien (Bajtelsmit 1975; Slakter et al. 1970; Woodley 1973) gestützt, in denen Personen mit TW gegenüber Personen ohne TW besser bei der Beantwortung von völlig inhaltsfreien MCA abschnitten. Zudem konnten Studien TW bei Tests in unterschiedlichen Themengebieten nachweisen (Diamond und Evans 1972; Pryczak 1973; Rowley 1974), was ebenfalls eine Unabhängigkeit unterstützen würde. Allerdings scheint es, dass TW zumindest zu einem gewissen Grad mit dem thematischen Wissen zusammenhängt. So weisen Millman et al. (1965) darauf hin, dass gerade für Schlussfolgerungs- (ID) und Hinweisnutzungsstrategien (IIB) teilweise thematisches Wissen nötig ist. Auch Yang (2000) weist darauf hin, dass es unmöglich sei, ohne thematisches Wis-

sen zu identifizieren, ob zwei Antworten dasselbe oder gegensätzliches aussagen (Schlussfolgerungsstrategie). Obwohl also die TW-Strategien theoretisch als voneinander getrennte Faktoren angesehen werden, treten sie in der Praxis vermutlich gemeinsam auf und beanspruchen wahrscheinlich teilweise die gleichen kognitiven Fähigkeiten. Inwieweit TW methodisch ein ein- oder mehrdimensionales Konstrukt ist, ist jedoch nicht eindeutig. TW ist somit ein Sammelbegriff für verschiedene Strategien, die durch die Anwendung das auf dem thematischen Wissen basierende Testergebnis verbessern können.

3 Wie wird TW gemessen?

Am häufigsten wird TW mit MCA erfasst, die überwiegend nur richtig beantwortet werden können, wenn TW angewendet wird (Edwards 2003; Gibb 1964; Woodley 1973). Der nicht nur erste, sondern auch am häufigsten verwendete und als Grundlage zur Entwicklung weiterer TW-Tests genutzte Test (Yang 2000), ist der „Experimental Test of Testwiseness“ von Gibb (1964). Er gilt als das umfassendste Instrument zur Erfassung von TW mit den besten psychometrischen Eigenschaften (Edwards 2003) und besteht aus insgesamt 70 MCA. Um die Möglichkeit zu reduzieren, dass die Items aufgrund vorhandenen Wissens beantwortet werden können, beinhalten sie sehr schwierige geschichtliche Fakten und Interpretationen. Sie können nicht mit Weltwissen gelöst werden. Als Folge sollten sie nur durch die Anwendung von TW-Strategien oder durch Raten richtig beantwortet werden können (Gibb 1964; Miller et al. 1988). Im Test von Gibb (1964) sind jeweils zehn Items zu einer Subskala zusammengefasst. Die Subskalen stehen für einzelne Hinweise in den MCA, die direkt auf die richtige Antwort hindeuten bzw. einer Testperson helfen können, falsche Antworten zu verwerfen. Tab. 1 zeigt die Subskalen des Tests, die nach einzelnen Hinweisen klassifiziert sind, sowie die entsprechende Strategie in der Taxonomie von Millman et al. (1965):

Nach Millman et al. (1965) entsprechen diese Subskalen den Schlussfolgerungs-(ID) und Hinweisnutzungsstrategien (IIB). Es scheint allerdings, dass die Anwendung bzw. die Identifizierung der verschiedenen Strategien nicht gleich schwer sind. So zeigte Morse (1998) in einer Studie, in der 243 Studierende dreier Universitäten den TW-Test von Gibb beantworteten, dass „specific determiners“ am schwersten und „grammatical cues“ am leichtesten von den sieben Strategien anzuwenden ist. Auch die Strategien „longer correct option“ und „absurd option“ scheinen eher einfach zu identifizieren zu sein, während „stem-option cues“ und „precise option“ zu den schwierigeren der sieben Strategien zählen.

Weiterhin gibt Gibb für die Gesamtskala eine Reliabilität (Kuder-Richardson Formula; KR-20) von 0,72 bei 193 Collegestudierenden an. Die Reliabilitäten der einzelnen Subskalen reichen von 0,16 bis 0,76. Ob TW wirklich aus mehreren Dimensionen besteht, scheint jedoch nicht ausreichend geklärt, da verschiedene Studien zu unterschiedlichen Ergebnissen kommen. So weisen die Befunde von Miller et al. (1988) darauf hin, dass anstelle der einzelnen Subskalen besser ein Gesamtwert verwendet werden sollte. Miller et al. (1990) dagegen kommen anhand ihrer Ergebnisse zu dem Schluss, dass TW nicht sieben, sondern zwei Dimensionen umfasst.

Tab. 1 Subskalen des TW-Tests von Gibb (1964) mit Einordnung in die Taxonomie von Millman et al. (1965)

TW-Strategien nach Gibb (1964)	Taxonomie von Millman et al. (1965)	Erläuterung
Sound-, Word-, Phrase-Repeat	Stem-option cues (IIB)	Ähnlichkeiten zwischen Itemstamm und Antwortmöglichkeiten sind eher in den richtigen als in falschen Antworten zu finden
Absurd Relationship	Absurd option (ID)	Nicht zur Frage passende oder völlig abwegige Antwortmöglichkeiten weisen auf falsche Antworten hin
Categorical Exclusive	Specific determiners (IIB)	Spezifische Wörter wie jegliche, nie oder völlig weisen auf falsche Antworten hin
Precise, Qualified, Specific	Precise option (IIB)	Die richtige Antwort ist meistens spezifischer und präziser als falsche Antworten
Length	Longer correct option (IIB)	Die längste Antwort ist häufig die richtige Antwort
Grammar	Grammatical cues (IIB)	Grammatikalische Fehler sind eher in falschen als in richtigen Antworten zu finden
Give-Away	Item giveaway (ID)	Anhand einer Aufgabe kann auf die richtige Antwort einer anderen Aufgabe im Test geschlossen werden

Schließlich zeigen Harmon et al. (1996), dass sowohl eine Ein-Faktorenlösung als auch eine Zwei-Faktorenlösung sinnvoll ist, wobei zwei TW-Strategien (stem-option und specific determiners) bei der Zwei-Faktorenlösung auf keinem der Faktoren laden. Einig sind sich die Autoren der drei Studien allerdings darin, dass, wenn von einer Ein- bzw. Zweidimensionalität ausgegangen wird, eine geringere Anzahl an MCA ausreichen würde. Eine solche Kürzung wurde von Edwards (2003) vorgenommen. Die ursprüngliche Anzahl an Items wurde von 70 auf 20 reduziert und somit die Beantwortungszeit von 25–30 min auf 10 min gesenkt. Die interne Konsistenz der gekürzten Testversion beträgt 0,57. Dies entspricht internen Konsistenzen, die auch in anderen Studien, die den Gesamttest von Gibb eingesetzt haben, erreicht wurden.

Es existieren einige weitere Tests, bei denen entweder jeweils ganz neue MCA entwickelt oder MCA unterschiedlicher Tests miteinander kombiniert wurden (Bajtelsmit 1975; Dodeen 2008; Hayati und Ghoghjogh 2008; Millman 1966; Slakter et al. 1970; Woodley 1973; Yang 2000). Werden die vorhandenen Tests in die Taxonomie von Millman et al. (1965) eingeordnet, so zeigt sich, dass die am häufigsten durch Tests erhobenen TW-Strategien Schlussfolgerungs- und Hinweisnutzungsstrategien sind.

4 Training von TW

Die Frage nach der Trainier- bzw. Erlernbarkeit von TW ist schon früh aufgeworfen worden. Eine Studie zu diesem Thema stammt von Houston (2005). In dieser Studie hatten die Probanden (86 angehende Feuerwehrfrauen und -männer) als Teil eines größeren Trainingsprogramms zuerst einige MCA zu TW zum Thema „Verhalten bei

Feuer“ beantwortet. Anschließend erhielten sie eine Schulung über TW-Strategien sowie Informationen zur Entwicklung von MCA. Nach dieser Schulung beantworteten die Probanden erneut die MCA. Die Ergebnisse im Posttest waren in fast allen Kategorien von TW signifikant besser als im Pretest. Für die Unterschiede in den Testergebnissen ergab sich eine Effektgröße von $d = -0,38$. In einer Metaanalyse von Bangert-Drowns et al. (1983) wurden Ergebnisse von 30 Studien zur Effektivität von Trainingsprogrammen bei Leistungstest, die alle TW beinhalteten, analysiert. Es wurden in diesen Studien unterschiedliche Stichproben von Klasse 2 bis zur Universität berücksichtigt. Die Befunde ergaben eine durchschnittliche Effektstärke von $d = 0,25$ für Leistungstestergebnisse, die nach einem TW-Trainingsprogramm durchgeführt wurden. In einer weiteren Metaanalyse von Samson (1985), in die Befunde von 24 Studien zur Trainierbarkeit von „test-taking skills“ in der Grundschule und Mittelstufe eingegangen sind, wird eine durchschnittliche Effektgröße von $d = 0,33$ erzielt. Dass Personen mit TW-Training gegenüber Personen ohne TW-Training bessere Ergebnisse erzielen, konnte auch in weiteren Studien gezeigt werden (z. B. Callenbach 1973; Dolly und Vick 1986; Dolly und Williams 1986; Oakland 1972; Omvig 1971).

Vor dem Hintergrund, dass MC-Tests den Anspruch haben, „echtes“ Wissen abzubilden, ist die Frage nach der Trainierbarkeit von TW von hoher Relevanz. Wenn das Erlernen von TW generell die Leistungen bei MC-Test erhöht, erfordert dies zukünftig einen stärkeren Fokus auf die Kontrolle von TW.

5 Fragestellungen

Vor dem Hintergrund berichteter Befunde ergeben sich für die vorliegende Untersuchung folgende Fragestellungen:

1. Zeigt ein entwickelter deutschsprachiger TW-Test zufriedenstellende psychometrische Kennwerte, die denen englischsprachiger Tests entsprechen?
2. Hat das thematische Wissen einen Effekt auf das Testergebnis im entwickelten TW-Test?
3. Hat eine Schulung einen Einfluss auf das Testergebnis im entwickelten TW-Test?

6 Methode

6.1 Stichprobe und Design

Es wurden Daten von 252 Studierenden einer Universität und einer Fachhochschule erhoben. Die Stichprobe setzte sich aus 91 Geschichtsstudierenden und 161 Studierende anderer Fachbereiche wie Betriebswirtschaftslehre, Pädagogik, Biologie oder Sport zusammen. Die Geschichtsstudierenden befanden sich durchschnittlich im dritten Fachsemester, sodass davon ausgegangen werden konnte, dass sie über Geschichtswissen verfügen. Dies wurde berücksichtigt, um zu prüfen, inwieweit der Test von Personen mit Expertise im Fach Geschichte gelöst werden kann.

Tab. 2 Deskriptive Kennwerte der vier Stichproben

<i>N</i>	Keine Schulung & kein Vorwissen		Keine Schulung & Vorwissen		Schulung & kein Vorwissen		Schulung & Vorwissen	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Alter	23,22	3,63	22,78	5,26	21,62	2,76	22,69	5,23
Gesamtanzahl richtiger Antworten im KFT	15,59	4,06	15,73	4,45	16,24	4,50	16,44	3,33

Die Testung erfolgte in regulären Lehrveranstaltungen durch einen geschulten Testleiter mittels eines standardisierten Durchführungsmanuals. Um sicher zu gehen, dass sich die Studierenden nicht in ihren kognitiven Fähigkeiten unterscheiden, bearbeiteten sie den Untertest N2 des Kognitiven Fähigkeitstests KFT 4–12 R+ (Heller und Perleth 2000). Er bestand aus 25 sprachfreien MCA, die in acht Minuten zu lösen sind. Sie erforderten die Zuordnung von Figurenpaaren, die in einer bestimmten Relation zueinander standen und ähnelten dem Prinzip der TW-MCA. Die in der Untersuchung erreichte Reliabilität (Kuder-Richardson 20) der KFT-Skala lag bei 0,78. Im Anschluss an den KFT folgte der deutschsprachige TW-Test, der zwar wahrheitsgemäße Geschichtsfragen umfasst, jedoch nur mit spezifischem Wissen richtig beantwortet werden konnte. Der Test wurde als Kurzform konzipiert und wird als zentrales Element dieser Arbeit in Abschn. 6.2 näher erläutert. Schließlich beantworteten die Studierenden Fragen zu ihrer Person (Geschlecht und Alter). Für die Beantwortung des TW-Tests und der persönlichen Angaben gab es keine Zeitvorgabe. Die Testbearbeitung dauerte ca. 30 min. Die Instruktion lautete, alle MCA zu beantworten und, wenn die Antwort nicht gewusst wird, diejenige zu wählen, die am ehesten als richtig erachtet wird.

Um der Trainierbarkeit von TW nachgehen zu können, erhielt ein Teil der Studierenden eine Schulung. Dabei wurde zufällig ausgewählt, in welcher Lehrveranstaltung die Studierenden eine Schulung erhalten (Schulung vs. keine Schulung). Die Schulung bestand aus zwei DIN A4 Seiten (Onlinematerial 1), die vor der Beantwortung der MCA ausgehändigt wurden und eine Beschreibung beinhalteten, was TW ist sowie zur Verdeutlichung der TW-Strategien jeweils ein Beispiel, das im TW-Test keine Verwendung fand. Die Schulung erfolgte standardisiert und dauerte ungefähr 10 min. Insgesamt wurde damit ein 2 (Wissen vs. kein Wissen) $\times$ 2 (Schulung vs. nicht Schulung) faktorielles Design mit vier Teilstichproben realisiert. Der Anteil weiblicher Personen betrug in der Stichprobe „keine Schulung & kein Wissen“ 73,40 %, in der Stichprobe „keine Schulung & Wissen“ 48 %, in der Stichprobe „Schulung & kein Wissen“ 62,12 % und in der Stichprobe „Schulung und Wissen“ 50 %. Tab. 2 gibt einen Überblick über die Stichprobenkennwerte.

Es ergaben sich für die Anzahl richtig gelöster Aufgaben im KFT ($p > 0,008$ als adjustiertes α) keine signifikanten Unterschiede zwischen den Stichproben. Hinsichtlich des Alters unterscheiden sich nur die Gruppen „keine Schulung & kein Wissen“ und „Schulung & kein Wissen“ signifikant voneinander ($U = 2174$, $z = -3,26$, $p = 0,001$).

6.2 Material

Der entwickelte TW-Test umfasst insgesamt 24 Single Choice-MCA (Onlinematerial 2) mit einer Frage und vier Antwortmöglichkeiten. Grundlage für die Auswahl der Items ist die Kurzversion des „Experimental Test of Testwiseness“ (Gibb 1964) von Edwards (2003). Für die vorliegende Studie waren nicht alle Items geeignet. Entfernt wurden MCA, die direkt das Wort Deutschland oder einen Bezug zu Deutschland enthielten, um ausschließen zu können, dass diese Aufgaben aufgrund thematischen Wissens beantwortet werden können. Die deutsche Übersetzung wurde durch eine Muttersprachlerin (native speaker) mit gleichzeitiger Erfahrung in der Übersetzung empirischer Texte der Bildungsforschung auf ihre Richtigkeit überprüft. Um den Wegfall zu kompensieren, wurden Items der Originaltestversion (Gibb 1964) hinzugefügt. Die Auswahl erfolgte aufgrund der publizierten Itemkennwerte. In Anlehnung an die Taxonomie von Millman et al. (1965) beinhaltet der deutschsprachige TW-Test die folgenden sechs Subskalen mit je vier Items:

1. stem-option cues (IIB),
2. absurd option (ID),
3. specific determiners (IIB),
4. precise option (IIB),
5. longer correct option (IIB) und
6. grammatical cues (IIB).

6.3 Statistische Auswertungen

Wie in der Darstellung der Forschungsfragen ersichtlich, geht es im ersten Teil der Untersuchung um psychometrische Analysen des deutschsprachigen TW-Tests. Hierzu wurden theoriegeleitet Faktorenanalysen (FA) unter Vorgabe einer Sechsen und einer Zwei-Faktoren-Lösung durchgeführt. Darüber hinaus wurden die Daten unter der Annahme der Eindimensionalität im Hinblick auf ihre Schwierigkeit und Trennschärfe (Kelava und Moosbrugger 2012) mit dem Programm *ConQuest* (Wu et al. 2007) ausgewertet. Zusätzlich wurden der T-Wert und der Weighted Infit Mean Square (WMNSQ) als Prüfgrößen für die Passung der Items zu einer gemeinsamen Dimension herangezogen. Cronbachs Alpha wurde als Maß für die interne Konsistenz mit dem Softwarepaket *SPSS* (Statistical Package for the Social Sciences) berechnet. Zusätzlich wurde als Maß für die Reliabilität in *ConQuest* (Wu et al. 2007) die EAP/PV-Reliabilität berechnet (vgl. Rost 2004). Ergänzend wurde mit *ConQuest* die Dimensionalität des Tests geprüft. Zur Überprüfung der Güte wurden auf Basis der Likelihood jedes Modells und der Modellparameterzahl der BIC (Bayes Information Criterion) und CAIC (Consistent Akaike's Information Criterion) herangezogen. Verglichen wurde, ob es sich um ein ein-, zwei- oder sechsdimensionales Modell handelt. Bei allen Analysen mit *ConQuest* (Wu et al. 2007) wurden in der vorliegenden Studie im Hintergrundmodell die Variablen Schulung (Schulung vs. Keine Schulung), thematisches Wissen (Wissen vs. Kein Wissen) und die Anzahl richtig beantworteter MCA im KFT berücksichtigt, um möglichst messfehlerkorrigierte Aussagen machen zu können.

Um die Fragestellungen hinsichtlich des Einflusses der TW-Schulung und des thematischen Wissens auf das Gesamtergebnis im TW-Test zu untersuchen, wurden zuerst deskriptive Kennwerte für die Anzahl richtig beantworteter MCA pro Teilstichprobe ermittelt. Zur inferenzstatistischen Überprüfung wurden aufgrund unterschiedlicher Gruppengrößen und -varianzen ein Kruskal-Wallis-Test und anschließend U-Tests nach Mann und Whitney (mit Anpassung des α -Levels nach Bonferroni) berechnet. Zuletzt erfolgte eine Regressionsanalyse zur Untersuchung des Effekts des thematischen Wissens und der Schulung auf das Ergebnis im TW-Test.

7 Ergebnisse

Zunächst werden die Ergebnisse des Tests zur Erfassung von TW (Fragestellung 1), anschließend die Ergebnisse der Fragestellungen 2 und 3 berichtet.

7.1 Test zur Messung von TW

Um initial die Dimensionalität der 24 TW-Items zu prüfen, wurden zwei FA – mit der Vorgabe von sechs Faktoren bzw. zwei Faktoren – berechnet. Die Ergebnisse (siehe Tab. 5 und 6 im Anhang) beider FA zeigten, dass die jeweiligen Items nicht substanziell auf den theoretisch angenommenen Faktoren laden. Da sich die theoretisch postulierte Struktur nicht zeigen ließ, wurden vertiefende Itemanalysen durchgeführt und die Homogenisierung der Skala zu einem eindimensionalen Konstrukt angestrebt. Die Schwierigkeiten und Trennschärfen der 24 Items sowie ihre Infit-Maße (WMNSQ und t -Werte) bezogen auf ihre Homogenität im eindimensionalen Raschmodell wurden ermittelt. Für domänenbezogene Leistungstests werden Trennschärfen von $\geq 0,25$ als zufriedenstellend angesehen. Allgemein werden t -Werte über 2 und unter -2 (Smith 1995) als signifikant betrachtet, wobei Werte unter -2 tolerierbar sind, da hoch trennscharfe Items unter anderem zu einer besseren Reliabilität führen können als niedrig trennscharfe Items (Items $t > 2$) (Glug 2009; Rost 2004). Für den WMNSQ wurde ein Intervall zwischen 0,75 und 1,33 als akzeptabel betrachtet (Adams und Khoo 1996). Im vorliegenden Test schwankten die Itemschwierigkeiten (Logits aus Raschanalysen) zwischen 2,966 (Item 11) und $-0,453$ (Item 1). Die Trennschärfen der Items lagen zwischen 0,06 (Item 19) und 0,51 (Item 9). Die t -Werte für die Infit-Maße lagen zwischen $-2,8$ (Item 9) und 2,6 (Item 21). Aufgrund nicht ausreichender Werte für die Trennschärfe bei 10 Items sowie t -Werte > 2 bei zwei Items wurde der TW-Test um die folgenden Items reduziert: 7, 11, 13, 17, 18, 19, 21, 22, 23 und 24. Ohne diese zehn Items variierten die Item-Logits zwischen $-0,25$ (Item 1) und 3,16 (Item 16). Es wurde eine mittlere Itemschwierigkeit von 1,07 erreicht. Diese wich von der mittleren Personenfähigkeit von 0,49 ab. Der Test schien für die vorliegende Stichprobe demnach etwas zu schwer zu sein. Die Trennschärfen schwankten zwischen 0,26 (Item 1) und 0,53 (Item 9). Die t -Werte der verbliebenen Items lagen zwischen $-2,1$ (Item 9) und 1,8 (Item 12) und die dazu korrespondierenden WMNSQ-Werte zwischen 0,91 (Item 9) und 1,10 (Item 12). Damit erreichten die übrigen 14 Items die geforderten Kriterien. Tab. 3 zeigt die Itemkennwerte der 14 Items.

Tab. 3 Itemkennwerte der TW-Items ($N=252$) unter Berücksichtigung des Hintergrundmodells

TW-Skalen	Item-Schwierigkeit (p)	Logit	r_{it}	WMNSQ	t -Wert
<i>Stem-option cues</i>					
MCA 12	0,33	1,758	0,29	1,10	1,8
MCA 14	0,54	0,804	0,39	1,02	0,5
MCA 15	0,46	1,142	0,36	1,04	1,0
Mittelwert	0,44	1,235	0,35	–	–
<i>Absurd options</i>					
MCA 10	0,62	0,497	0,32	1,08	1,7
Mittelwert	0,62	0,497	0,32	–	–
<i>Specific determiners</i>					
MCA 16	0,12	3,161	0,27	1,03	0,3
Mittelwert	0,12	3,161	0,27	–	–
<i>Precise option</i>					
MCA 4	0,45	1,232	0,37	1,02	0,6
MCA 5	0,62	0,460	0,49	0,93	–1,5
MCA 9	0,43	1,286	0,53	0,91	–2,1
Mittelwert	0,50	0,993	0,46	–	–
<i>Longer correct option</i>					
MCA 1	0,76	–0,250	0,26	1,05	0,6
MCA 3	0,33	1,791	0,33	1,02	0,3
Mittelwert	0,55	0,771	0,29	–	–
<i>Grammatical cues</i>					
MCA 2	0,62	0,429	0,52	0,95	–1,0
MCA 6	0,38	1,583	0,33	1,05	0,9
MCA 8	0,70	0,051	0,52	0,93	–1,2
MCA 20	0,50	1,000	0,48	0,93	–1,8
Mittelwert	0,55	0,766	0,46	–	–
Mittelwert	0,49	1,067	0,39	–	–
Gesamt					

WMNSQ Weighted Infit Mean Square; t -Wert: t -Wert des WMNSQ, Logit Itemschwierigkeit aus ConQuest-Analysen, r_{it} Itemtrennschärfe

Basierend auf vorherigen Befunden sowie der Tatsache, dass durch die Reduzierung der Itemzahl zwei der sechs Subskalen nur noch aus jeweils einem Item bestanden, wurde eine Ein- gegenüber einer Zweidimensionalität unter Berücksichtigung eines Hintergrundmodells (s. oben) geprüft. Die Überprüfung einer Sechsdimensionalität entfiel, da das Modell saturiert gewesen wäre, sodass keine zuverlässigen Kennwerte zu erwarten waren.

Die Analysen ergaben aufgrund des Itemsselektionsprozesses erwartungsgemäß für ein eindimensionales (BIC=4962,48, CAIC=4944,48) gegenüber einem zweidimensionalen (BIC=5545,32, CAIC=5568,32) Modell die beste Passung. Die ermittelte interne Konsistenz von $\alpha=0,55$ erreichte eine vergleichbare Reliabilität wie die englischsprachigen TW-Tests (z. B. Edwards 2003). Wurde unter Berücksichti-

Tab. 4 Unterschiede zwischen den mittleren Rangsummen der vier Gruppen

Gruppen- vergleich	Mittlerer Rang	Rang- summe	Mittlerer Rang	Rang- summe	<i>U</i>	<i>Z</i>	<i>p</i>
1–2	71,30	6702,00	102,17	7663,00	2237,00	–4,13	0,000
1–3	57,91	5443,50	113,40	7597,50	978,50	–7,49	0,000
1–4	48,45	4554,00	96,94	1551,00	89,00	–5,67	0,000
2–3	45,72	4104,00	90,28	6049,00	1254,00	–5,18	0,000
2–4	40,35	3026,00	72,50	1160,00	176,00	–4,47	0,000
3–4	41,45	2777,00	44,31	709,00	499,00	–0,43	0,667

1 keine Schulung & kein Wissen, 2 keine Schulung & Wissen, 3 Schulung & kein Wissen, 4 Schulung & Wissen, Adjustierung des Signifikanzlevels nach Bonferroni-Korrektur auf $\alpha = 0,0083$

gung des Hintergrundmodells die EAP/PV-Reliabilität geschätzt, so zeigte sich mit 0,72 ein zufriedenstellendes Ergebnis.

7.2 Auswirkung thematischen Wissens und einer Schulung auf TW

Die Ergebnisse zeigten, dass im Mittel die Studierenden der Gruppe 1 „keine Schulung & kein Wissen“ $M = 5,21$ ($SD = 1,9$), die Studierenden der Gruppe 2 „keine Schulung & Wissen“ durchschnittlich $M = 6,52$ ($SD = 1,83$), die Studierenden der Gruppe 3 „Schulung & kein Wissen“ $M = 8,70$ der MCA richtig ($SD = 2,74$) und die Studierenden der Gruppe 4 „Schulung & Wissen“ im Mittel $M = 9,19$ ($SD = 1,64$) der MCA richtig beantworteten. Die Unterschiede zwischen den Gruppen waren signifikant (Kruskal-Wallis-Test, $H = 82,72$, $df = 3$, $p < 0,001$).

Wurden nun die Befunde der Gruppenvergleiche mittels Mann-Whitney U-Test betrachtet (Tab. 4), so zeigte sich, dass sich die Gruppen jeweils mit einer Ausnahme, nämlich der Unterschied zwischen den Gruppen 3 und 4 ($p = 0,67$), signifikant voneinander unterschieden ($p < 0,008$ als adjustiertes α). Personen mit thematischem Wissen beantworteten somit mehr MCA im TW-Test richtig als Personen ohne thematisches Wissen und Personen mit Schulung beantworteten mehr MCA richtig als Personen ohne Schulung. Das thematische Wissen ($\beta_{\text{Wissen}} = 0,23$, $p < 0,001$) und die Schulung ($\beta_{\text{Schulung}} = 0,65$, $p < 0,001$) waren signifikante Prädiktoren für das Ergebnis im TW-Test. Das thematische Wissen hatte dabei einen geringeren Effekt als die Schulung.

8 Diskussion

In einem ersten Schritt ging es in der vorliegenden Studie darum, die Güte des TW-Tests zu ermitteln, um anschließend in einem zweiten Schritt zu untersuchen, ob eine Schulung und/oder thematisches Wissen in Bezug auf das Thema des TW-Tests einen Einfluss auf das Testergebnis haben.

Für die theoretisch begründeten mehrdimensionalen Modelle zeigten sich keine stützenden Befunde. Daher wurde eine Analyse der Itemkennwerte vorgenommen und eine eindimensionale Lösung angestrebt. Es wurden dann zehn Items ausgeschlossen. Die dann gefundene Eindimensionalität steht im Einklang mit englisch-

sprachigen Studien (Harmon et al. 1996; Miller et al. 1988). Auch die interne Konsistenz der letztendlichen Skala entspricht bisherigen Werten englischsprachiger Tests. Einige Schwachstellen ließen sich jedoch im vorliegenden Test feststellen. Durch den Wegfall von 10 Items besteht der TW-Test aus insgesamt 14 Items. Zwei der sechs TW-Strategien wurden dadurch von nur einem Item abgedeckt. Obwohl die einzelnen Items deutlich jeweils einer der TW-Strategien zugeordnet werden können, fällt in einigen wenigen MCA bzw. in deren Antwortmöglichkeiten eine Vermischung zweier TW-Strategien auf (MCA 5 „specific determiners“ und „precise option“). Die geringe Itemzahl sowie die vereinzeltere Vermischung von TW-Strategien können eine Eindimensionalität des TW-Tests begünstigt haben. Darüber hinaus schienen die Studierenden z. B. mit einigen der Items nicht so gut umgehen zu können bzw. waren nicht in der Lage, die geforderte TW-Strategie anzuwenden. So waren vor allem MCA, die die Strategie „specific determiners“ beinhalteten, schwierig zu lösen, was den Ergebnissen anderer Studien entspricht (Morse 1998). Möglicherweise sind die Formulierungen bzw. Hinweiswörter in den Antwortmöglichkeiten im Deutschen nicht so deutlich gewesen wie es im englischen Originaltest der Fall ist, sodass die Distraktoren attraktiver als die richtige Antwort erschienen. Eine falsche Antwort könnte so begünstigt worden sein. Verschiedene Autoren weisen in diesem Zusammenhang darauf hin, dass der erfolgreiche Einsatz einzelner TW-Strategien abhängig von der Erfahrung in der Anwendung dieser Strategien ist und sich diese über einen gewissen Zeitraum entwickelt (Hayati und Ghogh 2008, Sarnacki 1979). Drei der insgesamt vier Items sind u. a. aufgrund unzureichender Trennschärfen weggefallen. Dagegen waren die Strategien „absurd options“ und „grammatical cues“ für die Studierenden eher zu leicht (vgl. Morse 1998). Weitere Analysen hierzu wären empfehlenswert.

Insgesamt kann festgestellt werden, dass ein Test entstanden ist, der der Güte vorhandener internationaler Tests entspricht und bei kleinen sowie großen Studien im Rahmen der Bildungsforschung, bei denen es darum geht, Leistungen, Kompetenzen und Fähigkeiten zu messen, eingesetzt werden kann, da der Aufwand der Durchführung und Auswertung nicht sehr groß ist. Kritisch muss angemerkt werden, dass nicht weitere Variablen wie die Testmotivation oder die Testangst miterhoben wurden, die einen Einfluss auf das Ergebnis gehabt haben könnten. Gerade die Motivation wurde in vorherigen Studien als bedeutend für die Stabilität und die interne Konsistenz sowie für das Testergebnis hervorgehoben (Edwards 2003; Miller et al. 1988). Ein niedriges Testergebnis hat für die Studierenden keine Konsequenzen zur Folge. Ein hohes Testergebnis zu erzielen war also für sie nicht von großer Bedeutung. Somit ist es möglich, dass sie sich weniger angestrengt haben, als es bei einer Konsequenz für ein niedriges Testergebnis der Fall gewesen wäre. In zukünftigen Arbeiten sollten daher die Befunde durch weitere Indikatoren gestützt werden.

Ebenfalls kritisch zu erwähnen ist, dass die Studie von Edwards (2003), auf dessen Kurzversion des Tests von Gibb (1964) der TW-Test der vorliegenden Studie basiert, nicht publiziert wurde. Es handelt sich hierbei um eine unveröffentlichte Dissertation, sodass eine Verwendung und Validierung durch Andere erschwert wird.

Bezogen auf den zweiten Teil der Studie zeigen die Befunde, dass Personen mit thematischem Wissen mehr der TW-MCA richtig beantworten als Personen ohne thematisches Wissen. Dies spricht dafür, dass TW nicht unabhängig von dem

thematischen Wissen einer Person ist oder dass sich die Items sowohl mit Hilfe des Wissens als auch durch TW lösen lassen. Dies müsste ggf. nachgeprüft werden.

Alles in Allem war die Stichprobe für die Studie nicht sehr groß. Insbesondere die Gruppe „Schulung & Wissen“ ($N=16$) müsste strenggenommen mehr Studierende umfassen. Wünschenswert wäre es deswegen, die gesamten Ergebnisse anhand einer größeren Stichprobe weiter abzusichern.

Darüber hinaus wird deutlich, dass geschulte Personen, die über kein thematisches Wissen verfügen, gegenüber Personen, die zwar thematisches Wissen besitzen, aber keine Schulung erhalten haben, signifikant bessere Ergebnisse erzielen. Der Effekt der Schulung ist stärker als der des thematischen Wissens. Dies deutet wiederum darauf hin, dass sich die Items von Personen mit thematischem Wissen nicht besonders gut lösen lassen, was durch den Test intendiert wurde und für den Test spricht. Die Testdurchführung erfolgte nach standardisierten Richtlinien und alle Personen in der Gruppe „Schulung“ wurde identisches Material ausgehändigt. Für die Durchführung einer Schulung sollte in Zukunft stärker die unterschiedliche Schwierigkeit in der Anwendung der TW-Strategien berücksichtigt werden (vgl. Morse 1998).

Durch die vorliegende Studie sind einige Fragen hinsichtlich eines deutschsprachigen Tests zur Erfassung von TW geklärt worden. Es ergeben sich jedoch weitere Forschungsansätze. Zunächst erscheint es wichtig, TW und dessen Dimensionalität noch vertiefender zu analysieren. Es weisen theoretische Erkenntnisse und unterschiedliche Studien darauf hin, dass TW sehr komplex ist und sich aus verschiedenen Strategien zusammensetzt. Die Befunde zur Dimensionalität in dieser Studie jedoch weisen auf eine Eindimensionalität hin. Offen bleibt jedoch, welche Fähigkeiten die Anwendung der einzelnen Strategien erfordern bzw. ob es eine Fähigkeit gibt, die allen Strategien zu Grunde liegt. Hinsichtlich der Fähigkeitsdimensionen bzw. des Konstrukts selbst besteht somit weiterer Forschungs- und Optimierungsbedarf (vgl. Yang 2000). Vertieftes Wissen über das Konstrukt könnte zudem zu noch besseren Ergebnissen hinsichtlich der psychometrischen Güte des Tests führen.

Darüber hinaus erscheint es aufschlussreich, in einer weiterführenden Studie, den Expertisegrad im Hinblick auf das Thema der TW-Items zu variieren, um noch stärker den Einfluss des thematischen Wissens auf TW untersuchen zu können. Wünschenswert wäre es auch, wenn zwei Studien mit einem Prä-Post-Design getrennt für das thematische Wissen und Schulung umgesetzt werden könnte. Dies könnte helfen, die Befunde dieser Arbeit weiter abzusichern.

Weiterhin stellt sich die Frage, welche weiteren Faktoren für TW eine Rolle spielen und wie vertiefender untersucht werden kann, ob eine Person geraten, wirklich thematisches Wissen oder eine TW-Strategie angewendet hat. Besonders der Aspekt des Ratens bei der Beantwortung eines TW-Tests wäre von großer Bedeutung, da sich die Ratewahrscheinlichkeit, die bei MCA mit vier Antwortmöglichkeiten 25 % beträgt, durch die Anwendung von TW-Strategien noch erhöht werden kann. Dies wäre beispielsweise der Fall, wenn eine Person in der Lage ist, einige der Antwortmöglichkeiten auszuschließen und sich dann nur noch zwischen zwei Antworten entscheiden muss. Bezüglich TW sind also noch einige Fragen offen, die in Zukunft geklärt werden müssen, da TW ein nicht zu unterschätzendes Konstrukt in der Testentwicklung und -durchführung darstellt. Der technische Fortschritt zeigt hier ein großes Potenzial im Hinblick auf Methoden, die uns helfen, dem Einfluss

von TW ein Stück näher zu kommen. Beispielsweise Eyetrackingstudien, bei denen die Blickbewegungen einer Person bei der Testbearbeitung aufgezeichnet werden, haben – gerade auch in Verbindung mit weiteren Methoden wie das retrospektive Laute Denken – das Potenzial, Antworten auf offene Fragen zu liefern.

Anhang

Tab. 5 Ladungen aus einer varimax-rotierten 6-Faktorenlösung für die Erfassung von TW

TW-Items	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5	Faktor 6
<i>Stem-option cues</i>						
MCA 12	–	–	–	–0,45	–	–
MCA 14	–	–	–	–0,33	–	–0,37
MCA 15	–	–	0,43	–	–	–0,31
MCA 23	–	–	0,62	–	–	–
<i>Absurd options</i>						
MCA 7	–	–0,44	–	–	–	–
MCA 10	0,38	–	–	–	–	–
MCA 13	–	–	–	–	0,48	0,52
MCA 22	–	–	–	–	0,76	–
<i>Specific determiners</i>						
MCA 11	–	–0,52	–	–	–0,32	–
MCA 16	–	–	0,66	–	–	–
MCA 17	–	–	–	0,41	–	–0,36
MCA 19	–	–	–	–	0,32	–
<i>Precise option</i>						
MCA 4	0,48	–	–	0,30	–0,33	–
MCA 5	0,53	–	–	–	–	–
MCA 9	0,49	–	0,43	–	–	–
MCA 24	–	–	–	0,53	–	–
<i>Longer correct option</i>						
MCA 1	–	0,33	–	0,43	–	–
MCA 3	–	–	–	–	–	0,56
MCA 18	–	0,53	–	–	–	–
MCA 21	–	0,57	–	–	–	–
<i>Grammatical cues</i>						
MCA 2	0,55	–	–	–	–	–
MCA 6	0,40	–	–	–	–	–
MCA 8	0,60	–	–	–	–	–
MCA 20	0,47	–	–	–	–	–

Hauptkomponentenanalyse; Es sind nur Faktorladungen >0,30 dargestellt

Tab. 6 Ladungen aus einer varimax-rotierten 2-Faktorenlösung für die Erfassung von TW

TW-Items	Faktor 1	Faktor 2
<i>Stem-option cues</i>		
MCA 12	–	–
MCA 14	–	–0,34
MCA 15	0,31	–
MCA 23	–	–
<i>Absurd options</i>		
MCA 7	–	0,52
MCA 10	–	–
MCA 13	–	0,31
MCA 22	–	0,38
<i>Specific determiners</i>		
MCA 11	–	–
MCA 16	–	0,47
MCA 17	–	–
MCA 19	–	–
<i>Precise option</i>		
MCA 4	0,35	–
MCA 5	0,50	–
MCA 9	0,60	–
MCA 24	–	–
<i>Longer correct option</i>		
MCA 1	–	–
MCA 3	–	–
MCA 18	–	–0,54
MCA 21	–	–0,40
<i>Grammatical cues</i>		
MCA 2	0,54	–
MCA 6	0,35	–
MCA 8	0,59	–
MCA 20	0,48	–

Hauptkomponentenanalyse; Es sind nur Faktorladungen >0,30 dargestellt

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz (<http://creativecommons.org/licenses/by/4.0/deed.de>) veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Literatur

- Adams, R., & Khoo, S. (1996). *Quest: the interactive test analysis system (version 2.1)*. Melbourne: Australian Council for Educational Research.
- Bajtelsmit, J. W. (1975). *Development and validation of an adult measure of secondars cue-using strategies on objective examination: the test of obscure knowledge (TOOK)*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C..

- Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. L. C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research*, 53, 571–585.
- Ben-Simon, A., Budesu, D. V., & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, 21(1), 65–88.
- Bruch, M. A. (1981). Relationship of test-taking strategies to test anxiety and performance: toward a task analysis of examination behavior. *Cognitive Therapy and Research*, 1(5), 41–56.
- Brunner, M., Artelt, C., Krauss, S., & Baumert, J. (2007). Coaching for the PISA test. *Learning and Instruction*, 17, 111–122.
- Callenbach, C. (1973). The effects of instruction and practice in content-dependent test-taking techniques upon the standardized reading test scores of selected second grade students. *Journal of Educational Measurement*, 10, 25–30.
- Chang, S.-H., Lin, P.-C., & Lin, Z. C. (2007). Measures of partial knowledge and unexpected responses in multiple-choice tests. *Educational Technology & Society*, 10(4), 95–109.
- Cohen, A. C. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, 3(4), 307–331.
- Diamond, J. J., & Evans, W. J. (1972). An investigation of the cognitive correlates of test-wiseness. *Journal of Educational Measurement*, 9, 145–150.
- Dodeen, H. (2008). Assessing test-taking strategies of university students: developing a scale and estimating its psychometric indices. *Assessment & Evaluation in Higher Education*, 33(4), 409–419.
- Dolly, J. P., & Vick, D. S. (1986). An attempt to identify predictors of test-wiseness. *Psychological Reports*, 58, 663–672.
- Dolly, J. P., & Williams, K. S. (1986). Using test-taking strategies to maximize multiple-choice test scores. *Educational and Psychological Measurement*, 46, 619–625.
- Edwards, B. D. (2003). An examination of factors contributing to a reduction in race-based subgroup differences on a constructed response paper-pencil test of achievement. Unpublished Doctoral dissertation, Texas A&M University. <http://repository.tamu.edu/bitstream/handle/1969.1/128/etd-tamu-2003B-2003062513-Edwa-1.pdf?sequence=1>. Zugegriffen: 1. Apr. 2011.
- Fagley, N. S. (1987). Positional response bias in multiple-choice tests of learning: Its relation to testwiseness and guessing strategy. *Journal of Educational Psychology*, 79, 95–97.
- Gibb, B. G. (1964). Test-wiseness as a secondary cue response. University Microfilms Document No. 64-7643. Unpublished doctoral dissertation, Stanford University.
- Glug, I. (2009). Entwicklung und Validierung eines Multiple-Choice-Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung. http://eldiss.uni-kiel.de/macau/servlets/MCRFileNodeServlet/dissertation_derivate_00002755/dissertation_inga_glug.pdf. Zugegriffen: 15. Jan. 2017.
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York: Routledge.
- Harmon, M. G., Morse, D. T., & Morse, L. W. (1996). Confirmatory factor analysis of the Gibb Experimental Test of Test-wiseness. *Educational and Psychological Measurement*, 56, 276–286.
- Hayati, A. M., & Ghogh, A. N. (2008). Investigating the influence of proficiency and gender on the use of selected test-wiseness strategies in higher education. *English Language Teaching*, 1(2), 169–181.
- Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision*. Göttingen: Beltz.
- Houston, S. E. (2005). *Test-wiseness Training: An Investigation of the Impact of Test-wiseness in an Employment Setting*. Unpublished PH.D Dissertation. The Graduate Faculty of the University of Akron.
- Kelava, A., & Moosbrugger, H. (2012). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 75–102). Heidelberg: Springer.
- Miller, P. M., Fagley, N. S., & Lane, D. S. (1988). Stability of the Gibb (1964) experimental test of testwiseness. *Educational and Psychological Measurement*, 48, 1123–1127.
- Miller, P. M., Fuqua, D. R., & Fagley, N. S. (1990). Factor structure of the Gibb experimental test of test-wiseness. *Educational and Psychological Measurement*, 50, 203–208.
- Millman, J. (1966). *Test-wiseness in taking objective achievement and aptitude examinations. Final Report*. New York: College Entrance Examination Board.
- Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, 25, 707–726.
- Morse, D. T. (1998). The relative difficulty of selected test-wiseness skills among college students. *Educational and Psychological Measurement*, 58(3), 399–408.
- Oakland, T. (1972). The effects of test-wiseness materials on standardized test performance of preschool disadvantaged children. *Journal of School Psychology*, 4(10), 355–360.

- Omvig, C. P. (1971). Effects of guidance on the results of standardized achievement testing. *Measurement and Evaluation in Guidance*, 4, 47–52.
- Pryczak, F. (1973). *Use of similarities between stems and keyed choices in multiple-choice items*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.
- Rogers, W. T., & Bateson, D. J. (1991). The influence of test-wisness on performance of high school seniors on school leaving examinations. *Applied Measurement in Education*, 4(2), 159–183.
- Rogers, W. T., & Yang, P. (1996). Testwisness: its nature and application. *European Journal of Psychological Assessment*, 12(3), 247–259.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Huber.
- Rowley, G. L. (1974). Which examinees are most favored by the use of multiple choice tests? *Journal of Educational Measurement*, 11, 15–23.
- Samson, G. E. (1985). Effects of training in test-taking skills on achievement test performance: a quantitative synthesis. *The Journal of Educational Research*, 78(5), 261–266.
- Sarnacki, R. E. (1979). An examination of test-wisness in the cognitive test domain. *Review of Educational Research*, 2(49), 252–279.
- Slakter, M. J., Koehler, R. A., & Hampton, S. H. (1970). Grade level, sex, and selected aspects of test-wisness. *Journal of Educational Measurement*, 7(2), 119–122.
- Smith, R. M. (1995). *Using item mean squares to evaluate fit to the rasch model*. Paper presented at the annual Meeting of the American Educational Research Association, San Francisco.
- Tomkiewicz, J., & Rogers, W. T. (2005). The use of one-, two-, and three-parameter and nominal item response scoring in place of number-right-scoring in the presence of test-wisness. *The Alberta Journal of Educational Research*, 51(3), 200–215.
- Woodley, K. K. (1973). *Test-wisness program development and evaluation*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). *ACER ConQuest 2.0—Generalised item response modelling software*. Camberwell: ACER Press.
- Yang, P. (2000). *Effects of Test-wisness upon Performance on the Test of English as a Foreign Language*. Dissertation. Alberta, Canada: University of Alberta.