Discrete-Option Multiple-Choice: Evaluating the Psychometric Properties of a New Method of Knowledge Assessment

Inaugural-Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Sonja Willing

aus Witten

Düsseldorf, August 2013

aus dem Institut für Experimentelle Psychologie

der Heinrich-Heine Universität Düsseldorf

Gedruckt mit der Genehmigung der

Mathematisch-Naturwissenschaftlichen Fakultät der

Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Jochen Musch

Korreferent: Prof. Ute J. Bayen, Ph.D.

Tag der mündlichen Prüfung: 11.10.2013

Die Neugier steht immer an erster Stelle eines Problems, das gelöst werden will.

(Galileo Galilei)

Danke

Bei der Verwirklichung meiner Promotion haben mich eine Reihe von Personen und Institutionen unterstützt, denen ich an dieser Stelle herzlich danken möchte.

Mein größter Dank gilt Herrn Prof. Jochen Musch für die hervorragende Betreuung. Als inhaltlicher Ratgeber unterstützte er mich in allen Phasen meiner Promotion. Prof. Ute Bayen danke ich für die Übernahme der Zweitbegutachtung.

Des Weiteren möchte ich mich bei allen 6153 Probanden bedanken, die bis zuletzt engagiert an den hier vorliegenden Studien teilgenommen haben.

Meinen Eltern, Lisa Budniak, Conny Bickmann, Sabine Breuing und Marie Uhlig danke ich für ihre uneingeschränkte Unterstützung in allen Lebenslagen der vergangenen Jahre. Von Herzen möchte ich mich bei Thomas Labryga bedanken, auf dessen emotionale Unterstützung ich immer zählen konnte.

Für die inhaltlichen Diskussionen und ihre Hilfsbereitschaft danke ich meinen Arbeitskollegen Adrian Hoffmannn, Berenike Waubert de Puiseau, Birk Diedenhofen, Jana Sommer, Martin Ostapczuk, Martin Papenberg, Meik Michalke und Sebastian Ullrich, die mich in den letzten Jahren begleitet haben.

Meine Promotion wurde von der Friedrich-Naumann-Stiftung für die Freiheit mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) gefördert.

Eigenständigkeitserklärung

Ich versichere an Eides statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist. Alle wörtlich oder dem Sinn nach aus anderen Texten entnommenen Stellen sind als solche kenntlich gemacht.
Die Dissertation wurde in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, 28. August 2013

(Sonja Willing)

Table of Contents

Abstract

Multiple-choice (MC) tests are currently the most popular test format for the assessment of knowledge. Foster and Miller (2009) recently proposed the discrete-option multiple-choice (DOMC) format as an alternative computerized test format. DOMC tests are based on a sequential rather than simultaneous presentation of answer options and therefore resemble the conduct of sequential police lineups in eyewitness research.

In the domain of knowledge assessment, Foster and Miller (2009) discussed several potential advantages to DOMC testing. DOMC might enhance test economy and test security, since due to item stopping criteria only half of the answer options have to be presented for each item on average. Furthermore, DOMC testing presumably allows for a better control of testwiseness. Testwiseness – that is, the ability to find and to make use of subtle cues to the solution by scrutinizing all available answer options (Gibb, 1964) – threatens the validity of MC tests (Haladyna & Downing, 2004). Unintended cues to the solution are nevertheless present and useful on many MC tests (Brozo, Schmelzer & Spires, 1984). As answer options are presented one at a time in DOMC testing and because the majority of testwiseness strategies require the simultaneous comparison of all answer options (Millman, Bishop & Ebel, 1965), the use of DOMC presumably prevents the use of testwiseness cues. If construct-irrelevant variance due to testwiseness can be reduced, an increase in test reliability and validity may be expected. Surprisingly, the effects of the use of DOMC on the psychometric properties of a test have not yet been tested. However, DOMC testing also has some potential drawbacks that might interfere with reliable and valid measurement. One important potential drawback of DOMC testing is the possibility of serial position effects. The order in which answer options are presented and, in particular, the position of the solution, might be critical and might influence the psychometric quality of a DOMC test.

In a series of experiments with a total of 4911 participants, the psychometric properties of DOMC items were therefore systematically evaluated. The main purpose was to examine whether the psychometric properties of test items are improved by a sequential presentation of answer options, and whether the hypothesized benefits of the new sequential answering format outweigh its potential drawbacks.

Results show that the DOMC procedure was capable of preventing the use of testwiseness cues better than the traditional MC format (Experiment 1). In addition, DOMC testing also reduced the use of testwiseness cues on a continuing medical education test, which was aimed at developing and maintaining the knowledge of professionals in the medical field and that had been criticized for being susceptible to testwiseness strategies (Rotthoff, Fahron, Baehring & Scherbaum, 2008; Experiment 2a). This result also held when participants were not informed of the presence of testwiseness cues (Experiment 2b), and could be replicated with medical professionals (Experiment 2c) as well as with participants from outside the medical field (Experiments 2a and 2b). DOMC was thus shown to allow for a better control of testwiseness than MC testing. Owing to item stopping criteria, DOMC testing reduced the number of answer options that were presented per item and that were available for comparison when trying to arrive at the correct solution (Experiments 1 and 2). Testing time was thus reduced (Experiments 1 and 2). These advantages, however, are accompanied by serial position effects. This was shown in Experiment 3. For this experiment, a test for the assessment of political knowledge was developed and validated in an additional series of 8 validation studies. Although political knowledge is a necessary requirement for political participation in public life (Popkin & Dimock, 1999), a validated German test for the assessment of political knowledge has been lacking. In Experiment 3, when answer options were presented sequentially for the political knowledge test, item difficulty increased as the serial position of the correct answer increased. Items were also more difficult when the most attractive distractor was presented prior to the correct answer. This effect was small in the MC

format and of medium size in the DOMC format. The reliability and validity of a sequential DOMC test were nevertheless comparable to that of a parallel MC test (Experiment 3). Examinees identified MC items as easier and generally indicated a preference for the MC format over the DOMC format; however, they also viewed DOMC tests as superior with regard to the assessment of factual knowledge and as requiring a deeper understanding of the subject area (Experiment 3).

Taken together, the psychometric properties of DOMC testing did not surpass but were able to match those of the format hitherto considered to be the most valid for an objective assessment of knowledge. In view of some of its unique new features, the sequential answering format therefore seems to offer a promising alternative to the traditional MC format.

Zusammenfassung

Das Multiple-Choice (MC) Verfahren ist die am weitesten verbreitete Methode der Wissensdiagnostik. Kürzlich ist Discrete-Option Multiple-Choice (DOMC) als ein computerbasiertes alternatives Antwortformat vorgeschlagen worden (Foster & Miller, 2009). Es beruht auf einer sequentiellen anstelle der simultanen Präsentation der Antwortoptionen und weist deshalb Ähnlichkeit zur Durchführung von polizeilichen sequentiellen Gegenüberstellungen auf.

Im Bereich der Wissensdiagnostik diskutierten Foster und Miller (2009) mehrere mögliche Vorteile dieses neuen sequentiellen Antwortformats. Infolge der Implementation von Abbruchkriterien könnte DOMC durch den Verzicht auf die Präsentation eines Teils der Antwortoptionen die Testzeit reduzieren und somit die Testökonomie verbessern. Dies käme potentiell auch dem Testschutz zugute. Darüber hinaus könnte DOMC besser als MC geeignet sein, die Anwendung von *Testwiseness* (dt. „Testschläue“) zu kontrollieren. *Testwiseness* ist definiert als die Verwendung metakognitiver Antwortstrategien, mit deren Hilfe die Lösung auch ohne inhaltliches Wissen alleine durch die Identifikation eines durch den Vergleich aller Antwortoptionen erschließbaren Hinweises auf die richtige Lösung zu identifizieren ist (Gibb, 1964). *Testwiseness* beeinträchtigt die Konstruktvalidität von MC-Tests (Haladyna & Downing, 2004); dennoch sind in vielen MC-Tests unbeabsichtigte Hinweise auf die richtige Lösung enthalten (Brozo, Schmelzer & Spires, 1984). Die Mehrzahl der *Testwiseness*-Strategien erfordert den simultanen Vergleich aller Antwortoptionen (Millman, Bishop & Ebel, 1965). Unter Verwendung von DOMC werden die Antwortoptionen jedoch einzeln nacheinander präsentiert, weshalb DOMC möglicherweise die Identifikation und Anwendung von Hinweisen auf die richtige Lösung zu verhindern vermag. Wenn konstruktirrelevante Varianz in Form von *Testwiseness* zur Identifikation von Lösungshinweisen reduziert wird,

kann DOMC die Reliabilität und Validität möglicherweise verbessern. Deshalb überrascht es, dass die psychometrischen Eigenschaften von DOMC Items bislang nicht systematisch untersucht wurden. Den möglichen Vorteilen des neuen Verfahrens stehen nämlich auch eine Reihe potentieller Nachteile gegenüber, die möglicherweise mit einer reliablen und konstruktvaliden Messung nicht vereinbar sind. Zu den möglichen Nachteilen eines sequentiellen Antwortformats gehört, dass potentiell konstruktirrelevante Varianz in Folge der seriellen Position der Lösung und attraktiver Distraktoren entsteht. Bislang wurden solche möglichen seriellen Positionseffekte jedoch nicht untersucht.

In einer Experimentreihe mit insgesamt 4911 Probanden wurden daher die psychometrischen Eigenschaften von DOMC Items systematisch geprüft. Das Ziel war eine experimentell gestützte Antwort auf die Frage, ob die psychometrischen Eigenschaften von Items durch eine sequentielle Präsentation der Antwortoptionen verbessert werden können, und ob etwaige Vorteile des DOMC-Verfahrens seine möglichen Nachteile überwiegen.

Die Ergebnisse zeigen, dass ein DOMC-Test die testwisenessbasierte Nutzung von Lösungshinweisen besser zu kontrollieren vermochte als ein MC-Test (Experiment 1). In einem für die Fortbildung von Medizinern entwickelten „Continuing Medical Education"-Test, der für die Anwendbarkeit von *Testwiseness* aufgrund der enthaltenen Lösungshinweise kritisiert wurde (Rotthoff, Fahron, Baehring & Scherbaum, 2008), vermochte das DOMC-Verfahren die Anwendung von *Testwiseness* besser zu kontrollieren als ein MC-Test (Experiment 2a). Dies war auch der Fall, wenn die Probanden nicht über das Vorhandensein von Lösungshinweisen vorab informiert wurden (Experiment 2b) und war nicht auf Teilnehmer ohne medizinischen Hintergrund beschränkt (Experimente 2a und 2b), sondern konnte auch für Mediziner als unmittelbare Zielgruppe nachgewiesen werden (Experiment 2c). Insgesamt konnte somit gezeigt werden, dass ein DOMC-Test eine bessere Kontrolle von *Testwiseness* ermöglicht als ein MC-Test. Infolge der Abbruchkriterien reduziert DOMC die Anzahl der Antwortoptionen, die je Item präsentiert werden und die für einen Vergleich der

Antwortoptionen und zur Identifikation möglicher Lösungshinweise zur Verfügung stehen (Experimente 1 und 2). Dies führte zu einer Reduktion der Testzeit (Experimente 1 und 2). Diesen Vorteilen standen jedoch serielle Positionseffekte gegenüber, wie sich in einem weiteren Experiment zeigte. Für dieses wurde zunächst in insgesamt acht Validierungsstudien ein Test zur Erfassung von Politikwissen entwickelt; dieses bildet eine unverzichtbare Voraussetzung für die kompetente Teilhabe am öffentlichen Leben (Popkin & Dimock, 1999), es gab bislang hierfür jedoch noch kein validiertes Testverfahren. Die sequentielle Präsentation der Antwortoptionen erhöhte bei einem Politikwissenstest die Itemschwierigkeit, je später die Lösung präsentiert wurde und wenn vor der Lösung der attraktivste Distraktor präsentiert wurde (Experiment 3). Unter Verwendung von MC war dieser Effekt vernachlässigbar; unter Verwendung von DOMC wies er hingegen eine mittle Effektstärke auf. Dennoch erlaubte die Anwendung des DOMC-Verfahrens konstruktvalide Messungen; das sequentielle Antwortformat wies im Vergleich zum MC-Format eine vergleichbare Reliabilität und Validität auf (Experiment 3). In einer Befragung hielten Probanden eine sequentielle Antwortpräsentation für besser geeignet zur Erfassung von Faktenwissen und zur Förderung eines tieferen Lernverständnisses (Experiment 3), präferierten jedoch das MC-Format und bewerteten DOMC-Items als schwerer im Vergleich zu MC-Items. Insgesamt vermochte das sequentielle DOMC-Verfahren die psychometrischen Eigenschaften von Tests nach dem Antwortwahlverfahren zwar nicht zu verbessern, es wies jedoch eine vergleichbare Reliabilität und Validität wie das MC-Verfahren auf und stellt deshalb aufgrund seiner spezifischen Vorteile eine interessante Alternative zu MC-Tests dar.

1 Introduction and Theoretical Background

In the following, the multiple-choice (MC) format is described with its benefits and disadvantages (Chapter 1.1). One important drawback of MC testing is that the construct validity of MC tests can be compromised by testwiseness. In Chapter 1.2, the construct of testwiseness and its importance for MC tests are therefore explained. Chapter 1 concludes with a detailed description of a computer-based alternative answer format for MC tests, called discrete-option multiple-choice (DOMC, Chapter 1.3), which is the main subject of the present dissertation. Results obtained with this new test format are then presented in Chapters 2 and 3.

1.1 Multiple-Choice Testing

The MC format is one of the most valid and hence, popular testing formats for the assessment of knowledge. It was first introduced during World War I as the general basis for the Army Alpha test, which allowed the U.S. Army to classify 1.5 million soldiers for military purposes (Downing, 2006a). Today, MC is widely used in diverse settings including school tests, university exams, vocational aptitude tests, and even TV quiz shows. In its standard form, a MC item consists of a stem and a set of three to five answer options (Foster & Miller, 2009). The stem is the beginning part of an item and presents the question that has to be answered. Next to the stem, all possible answer options are presented. One of the answer options is keyed as the correct answer; the remaining answer options - called distractors - are scored as incorrect (Haladyna, 2004). Usually, all answer options are presented simultaneously to the examinees who are asked to choose the correct option.

1.1.1 Advantages of Multiple-Choice Testing

Owing to its many advantages, the MC format is the most common selected-response item format (Downing, 2006a). MC tests can be used for classroom testing as well as in large-scale testing programs for purposes of graduation, certification, licensure, evaluation, placement, and admission (Haladyna, 2004). MC items can be suitably applied to test all levels of learning and understanding. Unlike other test formats such as open questions or essays, MC tests can be scored easily, objectively, and even in an automated manner (Clegg & Cashin, 1986; Downing, 2006a). Most importantly, the psychometric properties of MC tests comply with high standards. Using MC, highly reliable and valid tests can be constructed (Downing, 2006a).

1.1.2 Drawbacks of Multiple-Choice Testing

However, MC items have several limitations. First, the person who creates the test is required to develop incorrect yet plausible options that can be difficult to devise (Downing, 2006a). When used repeatedly, MC items are also threatened by student memorization, copying, and sharing (Foster & Miller, 2009; Kingston, Foster, Miller & Tiemann, 2010). Most importantly, critics have remarked that the selection of an answer option for an MC item does not directly reveal the actual knowledge of a respondent, but rather indicates the alternative that the respondent considers to be most likely to be true (Holmes, 2002). This choice is based on a comparison that is performed by taking all available options into account simultaneously. Therefore, a drawback of the MC test format is that cues that indicate which solution is correct may be derived or identified by comparing the various answer options. Hence, the results of MC tests can be influenced by testwiseness strategies (Foster & Miller, 2009; Martinez, 1999) as described in Chapter 1.2.

1.2 Testwiseness

Testwiseness is defined as the construct-unrelated ability to find and to make use of subtle cues to the solution by scrutinizing all available answer options (Gibb, 1964). Regardless of the knowledge domain to be assessed, test takers therefore are often able to identify the correct answer option or to eliminate one or more of the distractors of an MC item solely on the basis of surface characteristics or by using content-independent reasoning processes (Millman, Bishop & Ebel, 1965; Woodley, 1975). Millman et al. (1965) presented the first comprehensive classification of testwiseness strategies. The majority of testwiseness strategies require the simultaneous comparison of all available answer options. By comparing the various answer options, the test takers try to identify a cue that facilitates the identification of the correct answer. Any use of such cues is likely to improve test scores (Allan, 1992). Rost and Sparfeldt (2007) surprisingly found that by comparing all available answer options, pupils could often identify the correct solution without even knowing the question. Representing construct-irrelevant variance, however, the use of cues threatens the construct validity of MC tests (Haladyna & Downing, 2004; Rost & Sparfeldt, 2007). Another problem is that individual differences in testwiseness skills may selectively reward highly testwise examinees and penalize individuals who lack such skills (Edwards, 2003; Hammond, McIndoe, Sansome & Spargo, 1998; Millman et al., 1965; Sarnacki, 1979; Taylor & Gardner, 1999).

1.2.1 Testwiseness Cues in Multiple-Choice Tests

In principle, items on carefully constructed and evaluated tests should not be solvable solely by the use of unintended cues if the guidelines of effective item writing are followed (Haladyna, 2004). However, Farley (1989) estimated that even an experienced item writer needs one hour to create a valid MC item and in practice, many MC items are created under

time pressure and by authors who have little experience in test development (Downing, 2006b). It is therefore hardly surprising that a comprehensive analysis by Brozo, Schmelzer, and Spires (1984) demonstrated that answer strategies are applicable and useful on many tests. Brozo et al. (1984) examined 1,220 MC items from 43 exams that had been administered at U.S. colleges. They found that about 44% of these items contained a cue that could be used by examinees with high testwiseness skills to exclude one or several distractors. On average, for these flawed items, using the available cues almost tripled the probability of a correct solution as compared to a baseline of random guessing. In a more recent study, Tarrant and Ware (2008) analyzed 10 tests that had been used for high-stakes assessments in a nursing program. They also found that between 28 - 75% of the MC test items contained flaws, most of which favored testwise students.

1.2.2 Testwiseness Cues in Continuing Medical Education Tests

The fact that there are many poorly constructed MC tests in classroom practice is disturbing (Brozo et al., 1984). It must be considered even more disturbing that some of the administered tests, such as continuing medical education (CME) tests, are used to make decisions that have serious consequences both for the examinees and their patients. Since 2004, medical professionals in Germany are committed by social law (German social security code V § 95d) to participate in CME trainings to maintain and develop their professional knowledge and skills after receiving their medical license. All medical professionals are thus required to acquire at least 250 CME points every five years. In cases of non-compliance, medical professionals are faced with penalties ranging from a reduction in income to the withdrawal of their medical license. Private studies based on printed lectures followed by an achievement test containing 10 MC items constitute one prominent and approved method of acquiring CME points (German Medical Association, 2004). If test takers answer at least 7

out of the 10 items correctly, they receive 3 CME points, which is considered equivalent to three weeks of training and thereby complies with their medical training requirements.

Whereas the requirement to maintain the professional skills of medical professionals is generally embraced, there has been a controversial debate about whether MC items in CME programs are sufficiently challenging. Critics have argued that due to the testwiseness cues they often contain, many of these items can be solved even without training and without sufficient knowledge in the domain of the test (Kühne-Eversmann, Nussbaum, Reincke & Fischer, 2007; Rotthoff, Fahron, Baehring & Scherbaum, 2008; Stagnaro-Green & Downing, 2006). Stagnaro-Green and Downing (2006) studied 40 CME items that were published in the New England Journal of Medicine in 2003. They found that 50% of all MC items contained the cue “Longest Alternative” (Brozo et al., 1984; Edwards, 2003; Gibb, 1964; Millman et al., 1965; Sarnacki, 1979). For these items, an elaborate solution consisted of a greater number of words than all of the distractors. When comparing all answer options simultaneously, this provided the examinees with a helpful cue to the solution. Kühne-Eversmann et al. (2007) systematically evaluated the quality of MC items in three established German medical journals. The most frequent flaws were testwiseness cues and again, up to 30% of all items contained the cue “Longest Alternative”. Rotthoff et al. (2008) investigated 200 items in twenty training units of four established German medical journals in 2006. They found that without exception, all training units included items with concealed cues that indicated the correct answer. Depending on the journal, the proportion of these items varied between 30% and 40%. One example of the frequent occurrence of testwiseness cues was the CME unit for "The diagnosis and management of upper abdominal pain" from the German Medical Journal (“Deutsches Ärzteblatt”) 06/2006: In this unit, 8 out of 10 items could be solved by simply applying testwiseness cues. It might therefore not be too surprising that 83.3% of the nearly 24,000 physicians taking this test answered all items in this unit correctly and thus complied with their medical training requirement.

To sum up, MC tests often contain cues to the solution and are therefore vulnerable to testwiseness strategies. To remedy the validity concerns associated with this problem, it is certainly worthwhile to make an effort to minimize the impact of such cues on MC test scores.

In the last several decades, many variants of the MC format have been proposed (Downing, 2006b; Rodriguez, 2005). These variants differ with regard to the number of answer options, the number of correct answer options, the test administration procedure, and the scoring system (Haladyna, 2004). However, all of these variants offer the simultaneous presentation of answer options to the examinee.

1.3 Discrete-Option Multiple-Choice Testing

By contrast, Foster and Miller (2009) recently introduced a computer-based alternative to MC testing called discrete-option multiple-choice (DOMC) testing. Like any MC item, a DOMC item consists of a stem, an answer option that is the correct solution, and several distractors (Foster & Miller, 2009). However, there are two important differences between MC and DOMC testing. First, a DOMC item is defined by the sequential rather than simultaneous presentation of the answer options. Answer options are presented sequentially one at a time in a random order, and examinees have to decide on the correctness of each separately presented option. This procedure is implemented in a forward-only direction, and examinees are not given the opportunity to review items or to change previous answers (Foster & Miller, 2009; Kingston, Tiemann, Miller & Foster, 2012). Second, in DOMC testing, three stopping criteria are implemented. The processing of an item ends and no further answer options are presented when one of the following conditions is met: (a) the solution has been correctly identified as such (in this case, no more answer options need to be presented), (b) the solution has incorrectly been rejected, or (c) a distractor has incorrectly been accepted.

In the last two cases, there is no need to present additional answer options because the item has already been answered incorrectly. In other words, the presentation of a DOMC item ends as soon as it has been answered correctly or incorrectly (Foster & Miller, 2009). Unlike MC items, DOMC items are therefore usually answered before all answer options have been presented. If one of these three stopping criteria is met, the item is scored. No further answer option is presented after the correct answer option because one of the stopping criteria is necessarily met after the correct answer is presented (i.e., it is either correctly accepted or incorrectly rejected; Foster & Miller, 2009).

It is interesting to note that there is one area of research that bears some striking resemblance to sequential MC testing, namely, the use of sequential police lineups in eyewitness research. In the usual simultaneous police lineups, which largely prevail in current practice, the witness is presented with a number of individuals and is asked to indicate the person who committed the crime, if present. However, this kind of police lineup, which parallels the usual MC procedure, has been criticized because witnesses may simply decide to choose the lineup member who most resembles the perpetrator without making an absolute comparison between their memory of the image of the perpetrator and each lineup member. Similarly, MC testing in knowledge assessment has also been criticized as allowing respondents to simply choose the most plausible alternative rather than to identify the solution with some certainty. In a sequential lineup procedure, as in DOMC testing, the eyewitness is therefore presented with one lineup member at a time and has to decide whether or not that person is the perpetrator before being allowed to view the next member. The idea behind this one-at-a-time procedure is to discourage the eyewitness from relying on a relative judgment and to simply decide whether each suspect looks like the perpetrator. In a sequential lineup, an eyewitness may decide that the current lineup member looks more like the perpetrator than the one before, but he or she cannot be sure whether the next person will not look even more

like the perpetrator. A sequential lineup can therefore be used to force eyewitnesses to make a real decision about whether or not they have identified the perpetrator.

However, research on the sequential versus simultaneous presentation of suspects in a police lineup has provided somewhat equivocal results and has not yet arrived at a definite conclusion (Mickes, Flowe & Wixted, 2012; Steblay, Dysart, Fulero & Lindsay, 2001). It is also important that given the visual nature of the lineup task, results obtained in this paradigm cannot be readily transferred to the case of a sequential versus simultaneous presentation of answer options in MC testing. In contrast to typical eyewitness experiments that apply the lineup procedure, there is always a solution - equivalent to a guilty suspect - for MC items.

1.3.1 Potential Benefits of Discrete-Option Multiple-Choice Testing

In the domain of knowledge assessment, there are several potential advantages to a sequential testing procedure that have hardly been evaluated. Owing to the stopping criteria, only half of the answer options have to be presented for each item on average. This potentially helps to reduce testing time in spite of the sequential presentation procedure, and Foster and Miller (2009) indeed observed that, compared to MC, DOMC reduced testing time by about 10%. Foster and Miller (2009) also identified the more limited exposure of the various answer options as another advantage of the new testing format. If an answer option is never presented to a participant, he or she cannot recall it or give it away to future participants. This makes it easier to reuse DOMC items on future exams and enhances test security.

Additionally, in a first test of their new format, Foster and Miller (2009) found that DOMC items were more difficult than standard MC items. This finding was replicated in a subsequent study using a larger sample (Kingston et al., 2012). A likely explanation for this higher difficulty is that in the DOMC format, it is no longer possible to compare the plausibility of all available answer options; rather, the examinee repeatedly has to make

decisions on the basis of the limited information that is provided by each single option. To make correct decisions in sequential DOMC testing, the examinee therefore has to be able to assess the correctness of each answer option separately, whereas in MC testing, all answer options can be considered simultaneously to identify the correct solution. Foster and Miller (2009) surmised that DOMC testing might therefore motivate deeper learning because the solution has to be identified by the learner without the help of accompanying distractors.

Most important, however, is that not being able to compare sequentially presented answer options may help to prevent the use of testwiseness cues. As the majority of testwiseness strategies – as classified by Millman et al. (1965) – require the simultaneous comparison of all answer options, DOMC potentially prevents the comparison of all options before answering an item, and reduces the probability that a cue to the solution can be identified. In their studies, both Foster and Miller (2009) and Kingston et al. (2012) have therefore argued that the increased item difficulty was probably the result of the reduced impact of testwiseness, although they did not ensure that the mathematical problems they presented actually did contain testwiseness cues. It is thus presently unknown whether DOMC testing is indeed capable of reducing the effect of testwiseness.

If construct-irrelevant variance due to testwiseness can be controlled using a sequential presentation of answer options, an increase in test reliability and validity may be expected. Surprisingly, the effects of the use of DOMC on the psychometric properties of a test have not yet been tested. However, such a test is necessary because DOMC testing also has some potential drawbacks that might interfere with reliable and valid measurement.

1.3.2 Potential Drawbacks of Discrete-Option Multiple-Choice Testing

One important potential drawback of DOMC testing is the possibility of serial position effects. The order in which answer options are presented and in particular, the position of the solution, might be critical and might influence the psychometric quality of a DOMC test. No

previous work has investigated potential order effects for DOMC items. However, for traditional MC items with a simultaneous presentation of the answer options, some research has been conducted on positional effects. Under conditions of time pressure, Clark (1956) observed that some participants had a tendency to neglect the later positions, which he interpreted as a failure to read all answer options before giving a response. Several other studies have also reported effects of the position of the correct answer on the difficulty of MC items (Gustav, 1963; Rapaport & Berg, 1955; Wevrick, 1962). However, such biases were typically weak (Attali & Bar-Hillel, 2003) and were difficult to interpret because of methodological and conceptual problems. In particular, previous studies on serial position effects in MC testing failed to randomize the position of the correct answer option, making it difficult to interpret the observed effects (Fagley, 1987).

Several researchers have surmised that rather than the position of the correct answer option, the relative position of the correct answer and of the most attractive distractor may influence the difficulty of MC items (Friel & Johnstone, 1979; Marcus, 1963). Friel and Johnstone (1979) found that presenting the most attractive distractor immediately prior to the correct answer reduced the difficulty of MC items. Thus, somewhat surprisingly, placing the most plausible distractor immediately before the solution did not attract responses away from the solution; rather, a close proximity of the solution and the most plausible distractor seemed to help test takers to better discriminate between the distractor and the solution. However, the positions of the solution and the most attractive distractor were not systematically varied by Friel and Johnstone (1979), and an opposite effect was found by Clark and Davey (2005) in two police lineup studies. In sequential lineups, presenting foils who were very similar to the target prior to the presentation of the target led to an increased number of incorrect identifications of the foil. Clark and Davey (2005) argued that when a very similar foil is presented after the target in a sequential procedure, the identification of the target is no longer complicated by the necessity of rejecting the similar foil first. If, however, a very similar foil

is presented prior to the target, this will most likely result in some erroneous identifications of the foil. In DOMC testing, a similar effect is easily conceivable because unlike in MC testing, a decision sometimes has to be made about the most plausible distractor before the solution is presented to the test taker. If it is more difficult for an examinee to dismiss the most attractive distractor in DOMC testing, presenting the most attractive distractor prior to the solution may increase the number of false alarms and may thus harm the psychometric properties of the test.

Surprisingly, no previous work has investigated potential order effects in DOMC items in what as yet is still a small body of research on the properties of the DOMC answer format.

2 Research Questions

To arrive at a fair and balanced assessment, both the potential benefits of the new sequential answering format and the accompanied potential drawbacks were investigated. In a series of experiments with a total of 4911 participants, the psychometric properties of DOMC items were systematically evaluated. The main purpose of the dissertation was to examine whether the psychometric properties of test items are improved by a sequential presentation of answer options, and whether any benefits of the new sequential answering format outweigh its potential drawbacks. In the following, the research questions of all studies are outlined and the rationale of the research of this dissertation is explained. The studies are described later in detail in Chapters 3.1 to 3.4.

Experiment 1 tested a major promise of the new format and investigated whether DOMC testing can indeed prevent the use of testwiseness cues more effectively than the usual MC format. To this end, a test consisting of items that included cues to their solutions was constructed. As answer options are presented one at a time in DOMC testing, and because the majority of testwiseness strategies require the simultaneous comparison of all answer options (Millman et al., 1965), we expected that the use of DOMC would prevent the comparison of all options before participants answered an item and would reduce the probability that a cue to the solution would be identified. We therefore expected that DOMC testing would be less affected by the impact of testwiseness strategies (see Article 3 in the appendix).

To extend the findings of Experiment 1 to a domain with higher stakes, we investigated whether a sequential presentation of answer options also allows for a better control of testwiseness in Continuing Medical Education (CME) tests - as described in Chapter 1.2.2 - in a series of three additional experiments (Experiment 2a-2c; see Article 2 in the appendix). To

this end, we presented examinees with a published CME test that has been criticized for being susceptible to testwiseness strategies (Rotthoff et al., 2008).

A secondary purpose of the Experiments 1 and 2 was to investigate the efficiency of the proposed new answer format. This was done by calculating the reduction in the number of answer options that needed to be presented to the examinee by using the DOMC format and by determining the associated decrease in testing time. Even though less than half of the answer options had to be presented per item on average, Foster and Miller (2009) observed a reduction in testing time of only about 10%. As DOMC testing requires test takers to make more decisions per item than MC testing, a significant reduction in the total testing time of somewhat less than 50% was expected. However, the time needed for the instructions was expected to be much higher in the DOMC condition due to the necessity to explain this new procedure to all participants (see Article 2 and 3 in the appendix).

Experiment 3 examined the psychometric properties of DOMC tests and tested for potential serial position effects (see Article 1 in the appendix). In particular, we investigated the reliability and convergent validity of the DOMC format. We expected that the DOMC scores would be more strongly associated with relevant external criteria than the MC scores due to the reduction of construct-irrelevant variance that would presumably be achieved by employing a sequential presentation procedure. Moreover, potential moderators of item difficulty in DOMC testing were investigated. DOMC testing likely increases mean item difficulty as compared with MC testing because the decisions of the test taker have to be made on the basis of the more limited information that is provided by the single options that are presented on DOMC tests. We therefore expected that DOMC items would be more difficult than MC items. We also wanted to examine effects of the serial position of the solution and the serial position of the most attractive distractor relative to the solution. Using a computer-based presentation, we therefore varied the positions of the correct answer option and the most attractive distractor. We expected that item difficulty would increase in the

DOMC format as the serial position of the correct answer increased because over the course of a sequential decision procedure, it may be difficult for the test taker to dismiss up to three distractors that are presented prior to the solution. However, additional false alarms may also potentially result in an increased difficulty of DOMC items. Therefore, we also expected that the serial position of the most attractive distractor would influence the difficulty of DOMC items as has previously been reported in police lineup research. In particular, given that the erroneous acceptance of a distractor is most likely to occur when the distractor is highly attractive, we expected that item difficulty would increase in the DOMC format when the most attractive distractor was presented prior to the correct answer option. Such an effect would challenge the application of the DOMC format or at least require a careful monitoring of serial position effects. However, we also wanted to determine whether such serial position effects, if present, would interfere with the valid measurement of the subject matter under investigation. This is not necessarily the case because, if less able participants fall prey to serial position effects more easily and therefore achieve a lower score, this is exactly what is expected from a valid measurement procedure.

As new answer formats are less familiar to the respondents, these foreign answer formats are frequently met with both curiosity and skepticism. When answer options are presented sequentially, less information is available to the test taker who also has to make more decisions per item than for items with a simultaneous presentation of answer options. The increased difficulty of DOMC items that was reported by Foster and Miller (2009) may well result in more favorable evaluations of the MC format. We therefore evaluated the perceived difficulty, face validity, stimulation of deeper learning, and general attractiveness of the two answer formats.

To summarize, the present dissertation reports five experiments that investigated the psychometric properties and potential benefits of the DOMC test format, and determined whether DOMC testing has drawbacks which interfere with reliable and valid measurement.

The development and validation of a political knowledge test in another series of 8 validation studies is also reported. The development of this test was a secondary objective of the present dissertation. It served not only as a vehicle for conducting the DOMC studies, but also closed a gap because a carefully validated test to measure individual differences in political knowledge in Germany has been lacking (see Article 4 in the appendix).

3 Overview of Studies

In the following, the main results of the experiments are presented. Additionally, the main results of the development and validation of the PWT are described. The original research articles are listed in the appendix.

3.1 Experiment 1:

DOMC prevents the Use of Testwiseness Cues in MC Testing

In Experiment 1, we investigated whether DOMC testing allows for a better control of testwiseness than the traditional MC format. To this end, we presented examinees with a test that contained cues pointing towards the correct solution in each item and checked whether these cues could be used less easily in DOMC testing.

Several tests have been constructed to measure the ability of individuals to take advantage of the existence of item cues (e.g., Gibb, 1964). A test of testwiseness needs to fulfill the following criteria: First, the test items must be rather difficult; participants should normally not have much knowledge that would allow them to answer the questions correctly. Second, each item must contain an item cue, which, if used cleverly, will allow the test taker to identify the correct solution or at least to increase the person's probability of identifying the correct solution. If these criteria are met, an item on a test of testwiseness can be solved if the item cue is recognized and applied by the test taker. The number of items that can be solved correctly can then be used as an index of the examinee's testwiseness. Unfortunately, to the best of our knowledge, no test of testwiseness has ever been published in the German language. Because the content of existing instruments is often rather culture-specific, a new German testwiseness test was constructed for Experiment 1.

The 24 test items contained one of four testwiseness cues where the correct answer option was more wordy than the distractors, contained a lower degree of generalization, or was directly opposite in meaning to a second answer option, or - if the answer options were numbers - was located in the middle between the two extremes.

After constructing this test of testwiseness, we also created a parallel control test by removing all cues from the testwiseness test items. In Experiment 1, we were thus able to create a condition in which participants were asked to solve items that did not contain any cues (no cue condition) or in which they were asked to solve items containing such cues (cue condition). To establish an additional group that would take a test that was even more susceptible to the use of item cues, we asked a third group of participants to work on a test that also contained item cues, and we additionally informed the participants in this group about the presence and the nature of these cues (informed cue condition). We created this third condition to examine whether DOMC can reduce the use of testwiseness even when examinees are explicitly informed about the presence of cues.

The experiment used a 2 x 3 between-subjects design with the first factor *testing format* (MC, DOMC) and the second factor *availability of testwiseness cues* (no cue, cue, cue & informed). For each participant, all responses were recorded, and a total test score for the 24 items was computed. Additionally, we recorded the time needed to read the instructions and to complete all items. 181 psychology students (155 female, 85.64%) participated and were randomly assigned to each of the six conditions.

Experiment 1 showed that the DOMC answer format was capable of preventing the use of item cues better than the traditional MC format. Although test items were generally more difficult in the DOMC than in the MC format, the availability of item cues led to an increase in test scores that was considerably larger in the MC condition. DOMC was thus shown to allow for a better control of testwiseness than MC. However, it is also true that the control of testwiseness afforded by DOMC was less than perfect, considering that participants profited

from the availability of item cues even in the DOMC condition. The most probable explanation for this is that some item cues can be used even under sequential presentation; for example, when all answer options are presented before one of the stopping criteria is met. Nevertheless, the DOMC format allowed for an improved control of testwiseness that was greatly superior to that under MC testing conditions. Moreover, DOMC testing reduced the number of answer options that were presented to the examinee and that were available for comparison when trying to arrive at the correct solution. This enhances both test difficulty and test security. In addition, participants in the DOMC condition completed the test significantly faster than participants in the MC condition. Thus, due to the smaller number of answer options that had to be presented in the DOMC condition, the time needed to answer all items was reduced by 21% when answer options were presented sequentially. However, participants needed longer to read the extended instructions in the DOMC condition. DOMC therefore seems to have the potential to reduce testing time, at least once the test takers get accustomed to the new format and no longer need lengthy instructions.

To sum up, Experiment 1 shows that DOMC testing offers a promising alternative to the traditional MC format in consideration of the control of testwiseness cues and the improvement in testing time (see Article 3 in the appendix).

3.2 Experiment 2:

DOMC prevents the Use of Testwiseness Cues on CME Tests

To extend the findings of Experiment 1 to a domain with higher stakes, we investigated whether a sequential presentation of answer options also allows for a better control of testwiseness in Continuing Medical Education (CME) tests (Experiment 2). We expected that this effect would be limited to the less carefully constructed items for which the solution was more likely to be identified by making use of the cues they contained. To this end, we used a

published test intended and previously employed for high-stakes purposes, namely, the test accompanying the CME unit for "The diagnosis and management of upper abdominal pain" from the German Medical Journal (“Deutsches Ärzteblatt”, 06/2006; Niedergethmann & Post, 2006). In a series of three experiments, we randomly assigned participants taking this test to either the MC or the DOMC condition. In each condition, every test taker was provided with the 10 items accompanying this CME unit, of which 2 items did not contain cues, whereas the remaining 8 items could be solved on the sole basis of construct-unrelated cues to the solution (Rotthoff et al., 2008). For these 8 items, the correct answer option was more wordy than the distractors, contained a lower degree of generalization, or was grammatically better aligned with the stem.

The three serial experiments used a 2 x 2 mixed design with the independent between-subjects variable *answer format* (MC vs. DOMC) and the within-subjects variable *availability of cues* (items with vs. items without cues to their solution). As dependent variables, we recorded the proportion of correctly solved items and the time needed to read the instructions and to complete all items.

3.2.1 Experiment 2a

The aim of Experiment 2a was to examine whether DOMC testing would allow for a better control of testwiseness than MC testing. Furthermore, the efficiency of DOMC testing was explored by determining the decrease in testing time associated with using DOMC rather than MC testing.

The sample consisted of 48 (27 female) native German speakers. At the beginning of the questionnaire, participants were told that MC items are frequently criticized for the presence of cues to their solution, and that finding such cues facilitates the solution of an item, even if a test taker is not familiar with the topic at hand. The above three testwiseness cues were explained, and an example item was given for each.

In summary, the sequential presentation of answer options in the DOMC condition was more successful at preventing the use of cues to the solution than the simultaneous presentation of answer options that is customary in MC testing. However, opposite to our expectations and contrary to the results of Foster and Miller (2009), a main effect of the answer format was not found. Although DOMC items were also more difficult than MC items in our experiment, the difference between the two answer formats was not statistically significant. However, this lack of a significant main effect of test format may have been the result of the lack of power associated with the small sample size. More importantly, with regard to the usefulness of a sequential presentation of answer options, we observed a substantial reduction in testing time (30%) when using DOMC rather than MC testing. This effect was the result of the smaller number of answer options that had to be presented in the DOMC condition. As a result, a perfect test taker could be expected to be presented with an average of 3 out of the 5 possible answer options in the DOMC condition. For the test takers in the Experiment 2a, owing to their frequent erroneous acceptance of a distractor early in the course of the presentation of an item, this presentation ended even earlier: on average, after 1.68 (SD = 0.44) out of the 5 possible answer options.

However, an obvious and legitimate criticism of Experiment 2a is that the cues to the solution were explained explicitly to all participants prior to the presentation of the CME items. It may be argued that participants were able to make use of the cues to the solution only because the cues were explicitly revealed to them before they began working on the test. Therefore, to ensure the reliability and replicability of our finding, we conducted a second experiment in which participants were not informed about the presence of testwiseness cues.

3.2.2 Experiment 2b

In Experiment 2b, we expected that participants in the MC condition would be more successful than participants in the DOMC condition at making use of cues to the solution,

even when both groups were not informed about the presence of such cues. In order to increase the statistical power to detect a main effect of answer mode, we used a larger sample size than was used in Experiment 2a.

In Experiment 2b, 86 (45 female) native German speakers participated. The procedure was the same as in Experiment 2a with the exception that no participants were informed about the nature of the cues that could be used to detect the solutions of some of the items.

Summarizing, Experiment 2b replicated the findings of Experiment 2a and supported the notion that DOMC testing allows for a better control of testwiseness than MC testing. This effect was found even when we did not inform participants about the cues contained in some of the items. Experiment 2b also replicated the finding that DOMC allows for a reduction of the average number of answer options that had to be presented to the test takers (M = 1.88, SD = 0.49) and in the total testing time (23%). It is important to note, however, that participants in Experiments 2a and 2b did not have any medical background, as indicated by their chance performance on items that did not contain cues to their solution. We therefore wanted to replicate our findings using a sample of medical students and trained physicians in Experiment 2c.

3.2.3 Experiment 2c

For obvious reasons, we were particularly interested in whether DOMC would allow for the control of testwiseness and would decrease testing time not only in a convenience sample of participants with heterogeneous educational backgrounds, but also in the target group of individuals who were in continuing medical education. Therefore, medical professionals were used as the sample in Experiment 2c to examine whether the findings of the two previous experiments could be replicated.

In Experiment 2c, the sample consisted of 106 (66 female) medical students and medical doctors. The participants were all German native speakers and were recruited via

announcements in a learning portal for medical professionals (MEDI-LEARN) as well as in the bulletin board of a student council for medical students. The procedure was the same as in Experiment 2b.

Summarizing, the results of Experiment 2c confirmed the hypothesis that the sequential presentation of answer options would allow for the control of testwiseness for poorly constructed CME items. This was shown not only for participants with a heterogeneous educational background, but also for medical professionals who achieved higher test scores than participants in the two previous studies; these medical professionals were less dependent on relying on such cues in the first place due to their medical knowledge. In addition, the sequential presentation of answer options reduced total testing time by 25%.

In conclusion, the three experiments reveal three straightforward benefits of the new DOMC answer format. First, DOMC allows for a better control of testwiseness than MC testing. Second, DOMC testing reduces the number of answer options that are presented per item and that are available for comparison when trying to arrive at the correct solution, thereby enhancing test security. Last, the use of DOMC reduces testing time, in spite of the additional time that is needed for participants to read the extended instructions to understand the new format, and even though none of our participants was acquainted with the new format (see Article 2 in the appendix).

3.3 Experiment 3:

Evaluating the Psychometric Properties of a DOMC Test

The main goal of Experiment 3 was to investigate whether the psychometric properties of a test would be improved by employing this alternative test format for MC testing. To address this question, we compared item difficulty and discrimination as well as reliability and validity for a test that was randomly administered in either the MC or the DOMC format.

Additionally, we evaluated the acceptability of the DOMC format. To arrive at a balanced assessment of the DOMC format, we also tested for serial position effects. Accordingly, a set of items with known properties was needed. Therefore, we developed a pretest to administer eight political knowledge items to a sample of 258 panelists who did not participate in the main study (see Study 4). Across all items, the most frequently chosen alternative was the solution, and there was one distractor that was clearly more attractive than the remaining distractors. Using these items, we were able to systematically investigate the impact of the positions of the correct answer and the most attractive distractor relative to the solution. Using a computer-based presentation, the positions of the correct answer option and the most attractive distractor were therefore varied.

The experiment used an incomplete 2 x 4 x 2 mixed design with the independent between-subjects variable *answer format* (MC, DOMC), the within-subjects variable *position of the correct answer option* (1, 2, 3, 4), and the within-subjects variable *position of the most attractive distractor* (before the correct answer, after the correct answer). Some cells in this design were empty because the most attractive distractor could not be presented prior to the correct answer if the correct answer was presented as the first option, or after the correct answer if the correct answer was presented as the last option. 4,490 (2,653 female, 59.09%) native German speakers were recruited via the SoSci Panel, an open panel for the recruitment of participants for scientific investigations (Leiner, 2012).

Results showed that item difficulty increased as the serial position of the correct answer increased when answer options were presented sequentially. Items were also more difficult when the most attractive distractor was presented prior to the correct answer. This effect was small in the MC format and of medium size in the DOMC format. The reliability and validity of a sequential DOMC test were nevertheless comparable to that of a parallel MC test. Examinees identified MC items as easier and generally indicated a preference for the MC format over the DOMC format; however, they also viewed DOMC tests as superior with

regard to the assessment of factual knowledge and as requiring a deeper understanding of the subject area (see Article 1 in the appendix). The benefits and drawbacks of DOMC testing are therefore discussed in Chapter 4.

3.4 Study 4:

Development and Validation of the Political Knowledge Test (PWT)

Political knowledge is a necessary precondition for successful participation in public life and for the development of political culture in democracies (Popkin & Dimock, 1999). Surprisingly, no carefully validated test for measuring an individual's level of political knowledge has ever been published in the German language with the exception of three short subscales from German general knowledge tests. These subscales are included in (a) the Differential Knowledge Test (DWT; Jäger & Fürntratt, 1968), (b) the Bochum Knowledge Test (BOWIT; Hossiep & Schulte, 2008), and (c) the SPIEGEL students PISA test (Trepte & Verbeet, 2010). However, items from the political knowledge subscales of these general knowledge tests (DWT, BOWIT, and SPIEGEL students PISA test) are not sufficient for a reliable assessment of the construct of political knowledge because they are too few in number, partly unreliable or obsolete and not sufficiently validated, and sometimes refer to domains other than politics. Therefore, Study 4 presents the development and validation of the German political knowledge test (Politikwissenstest; PWT). In a pretest, 31 MC items were selected to construct the final form of the PWT. In two cross-validation studies, the PWT showed good psychometric properties and a clear one-factor loading structure. Evidence for the convergent validity of the PWT was obtained in six additional validation studies by relating the test to (a) self-ratings of political knowledge, interests, and media use, (b) political knowledge items taken from general knowledge tests, and (c) various measures of intelligence. Discriminant validity was established with regard to the scientific knowledge

subscales of a general knowledge test. Taken together, we found evidence that the PWT has high construct validity; thus, it is able to provide a reliable and valid assessment of an individuals' level of political knowledge. In developing the PWT - in contrast to the procedures followed for the DWT and the SPIEGEL students PISA test - we selected only items that are not subject to rapid change over time. The PWT therefore offers a reliable tool for future inquiries into the field of political knowledge (see Article 4 in the appendix).

4 General Discussion

Discrete-option multiple-choice testing is based on a sequential rather than a simultaneous presentation of answer options. The main goal of the series of experiments was to investigate whether the psychometric properties of a test are improved by employing this alternative test format for MC testing.

Across all five experiments, MC items were easier to answer correctly than DOMC items. This is probably because for the latter, a simultaneous comparison of all answer options is not possible and the test taker therefore has to base his decision on the much more limited amount of information that is provided by single options (Foster & Miller, 2009; Kingston et al., 2012). Over the course of this sequential decision making, it seems to be difficult for test takers to dismiss up to three distractors and to wait for the correct answer, resulting in an increased number of false alarms in the DOMC format. Accordingly, item difficulty depended on the position of the correct answer when tests were conducted sequentially, and item difficulty increased as the serial position of the correct answer increased (Experiment 3). An effect of the position of the correct answer on item difficulty was also observed in the MC condition, in which answers were presented simultaneously. A likely explanation for this finding is that some participants failed to read all options thoroughly, resulting in a small response bias toward the early options. Such a response bias necessarily results in a somewhat lower proportion of correct solutions when the correct answer is presented as one of the later options (Fagley, 1987). However, in the MC condition, this effect was of relatively small magnitude and statistically significant only because of the large sample size used in Experiment 3. It is also important to note that in spite of these format-specific effects, the reliability and validity of the sequential DOMC test did not differ from that of the parallel MC test.

As in sequential police lineups, Experiment 3 showed that the position of the most attractive distractor also influenced the difficulty of the task. If an item contained a very attractive distractor that was considered correct by many examinees, it was difficult for the test takers to dismiss this distractor if it was presented prior to the correct answer. Whereas the magnitude of this effect was rather small in the MC format, it was of medium size in the DOMC format, presumably because under a sequential presentation procedure, the potentially even more attractive solution was not yet available to the test taker when he had to make a decision about the correctness of the attractive distractor. Thus, in contrast to the findings by Friel and Johnstone (1979), who incidentally did not randomize the position of the correct answer, the item difficulty of MC items did not decrease in Experiment 3 if the most attractive distractor was presented immediately before the solution. Rather, we observed the exact opposite effect: Item difficulty was higher for the MC format if the most attractive distractor was presented before the correct answer option. However, this effect was of very small magnitude in the MC format and therefore seems rather negligible. A likely explanation for the small effect of the serial position of the most attractive distractor relative to the solution is a failure of some of the participants to read all answer options before providing their response. If a highly attractive distractor is presented prior to the solution, these participants are likely to wrongly accept it and never even reach the correct answer.

In addition, Experiments 1 and 2 showed that DOMC also increased item difficulty due to a better control of testwiseness. DOMC outperformed MC in preventing the comparison of all available options before answering an item, and thus in identification of testwiseness cues to the solution. However, it is also true that the control of testwiseness afforded by DOMC was less than perfect, considering that participants profited from the availability of item cues even in the DOMC condition. This was most likely because some item cues can be used even under sequential presentation, for example when all answer options are presented before one of the stopping criteria is met. Nevertheless, the DOMC format allows for an improved

control of testwiseness that is greatly superior to that under MC testing conditions (Experiments 1 and 2). This was also true for medical professionals on a CME test that was most likely created under time pressure and by authors who had little experience in the development of tests. According to Brozo et al. (1984), it is especially this kind of test that often contains cues to the solution. The results also suggest that the benefits of DOMC testing will probably generalize to professional participants in high-stakes testing situations.

Moreover, assessing the number of answer options that had to be presented for each DOMC item allowed us to understand why this format is more efficient at controlling for testwiseness than MC testing. In Experiments 1 and 2, less than half of the answer options were presented per item. This great reduction in the number of answer options that were available for comparison made it difficult to take full advantage of the available cues in the DOMC condition. Owing to the decreased number of answer options that have to be presented per item, a considerable reduction in testing time seems to be an additional advantage of DOMC testing. Unlike Foster and Miller (2009), who observed a decreased testing time of only about 10%, we found a reduction in testing time up to 30%. Only in the first experiment, however, this reduction was no longer significant when the time needed for the extended instructions was taken into account. Although on average, less than half of the answer options had to be presented per item, testing time was reduced by less than 50%. This was probably due to the need to make decisions after the presentation of each answer option, which prevented the savings in time from amounting to the same magnitude as the savings in the number of items. However, the savings in testing time may well increase up to 50% once test takers become more familiar with the new testing format.

Since the DOMC format allows for a better control of testwiseness than the MC format, it therefore can potentially improve the fairness of knowledge assessment (Foster & Miller, 2009). The present results also show, however, that the difficulty of DOMC items is influenced by the serial position of the correct answer and by the position of the most

attractive distractor relative to the correct answer. A careful consideration of such effects therefore seems necessary. If the order in which answer options are presented is randomized, the attractiveness of the distractors shown to the examinees will vary. It is therefore possible that in spite of having the same ability, two examinees will receive different test scores. This poses a potential threat to the fairness and validity of test scores and may make it necessary to hold the order of answer options constant within items.

However, even though no such control of the order of answer options was used in Experiment 3, we found no difference between the reliability and validity of the sequential DOMC procedure and the simultaneous MC testing procedure. This finding suggests that the small to medium serial position effects we observed are not harmful to the psychometric quality of DOMC tests if the position of answer options is randomized for all items and all participants. On the other hand, there was also no superiority of the psychometric properties of DOMC tests in spite of the better control that DOMC testing affords for construct-irrelevant variance due to testwiseness. Counter to our expectations, no significant differences in convergent validity with regard to a number of external criteria were observed between MC and DOMC testing. Although DOMC items were generally more difficult, their discrimination and internal consistency did not differ from those of MC items.

In Experiment 3, the acceptability of the DOMC format was investigated. Examinees rated MC items as easier, and potentially for this reason, they generally indicated a preference for the MC format over the DOMC format. The notion that a perceived higher difficulty of DOMC items is responsible for this preference is supported by the finding that low-scoring participants showed an even stronger preference for the MC test format. However, participants also rated DOMC tests as superior with regard to the assessment of factual knowledge and as requiring a deeper understanding of the subject area. Arguably, participants recognized that searching for cues and relying on testwiseness to identify the solution is less helpful in DOMC than in MC testing. If participants of Experiment 3 have a point in

believing that the DOMC format requires a deeper understanding of the subject area at hand, it seems worthwhile to investigate in an instructional context whether the expectation of having to take a test with a sequential answer format is capable of encouraging deeper learning and more careful preparation among test takers.

Taken together, with the availability of the new DOMC test format, examiners are faced with the necessity of making a decision. Is the implementation of DOMC testing worth the extra effort? Potential benefits of DOMC testing come at a price. First, DOMC testing increases item difficulty as a function of the positions of the correct answer option and the most attractive distractor. Even though in Experiment 3, such serial position effects did not harm the psychometric properties of the test, they potentially challenge the application of the DOMC format. A careful monitoring of serial position effects and further investigations into whether they moderate the reliability and validity of DOMC tests are therefore desirable. It is also a drawback of the DOMC format that test administration software is needed, posing an additional burden to the examiner even though the costs of implementation are one-time expenses during the initial stages of implementation (Exam Innovations, 2010). However, many tests are now being conducted online anyway and can easily be adapted to a sequential presentation procedure. Therefore, it may be more important that straightforward advantages associated with the use of DOMC testing have been found. First, the number of answer options that are presented per item and that are available for comparison when trying to arrive at the correct solution is reduced, thereby enhancing test security (Foster & Miller, 2009). Second, DOMC reduces testing time, in spite of the additional time that is needed to read the extended instructions for the new format. Third, the DOMC format provides the opportunity to increase item difficulty without changing item discrimination. DOMC testing thereby allows test creators to construct more challenging and demanding tests that are perceived as more difficult by the examinees, who may thereby be motivated to process the information on a deeper level and prepare for the exam more carefully. Fourth and perhaps most importantly,

the sequential presentation procedure allows for a better control of testwiseness than traditional MC tests. DOMC outperforms MC in preventing the identification of cues on tests that had been criticized for being susceptible to testwiseness strategies (Rotthoff et al., 2008). This unique advantage of DOMC over MC testing deserves attention because it addresses a major concern about the use of MC testing in general.

In spite of the different drawbacks and advantages of the two answer formats, the reliability and validity of a sequential DOMC test were found to be equivalent to those of a parallel MC test in Experiment 3. Thus, the psychometric properties of DOMC testing did not surpass but were able to match those of the format hitherto considered to be the most valid for an objective assessment of knowledge. In view of some of its unique new features, the sequential answer format therefore seems to offer a promising alternative to the traditional MC format and it therefore certainly seems worthwhile to further evaluate the usefulness of this new answer format.

References

Allan, A. (1992). Development and validation of a scale to measure testwiseness in EFL/ESL reading test takers. *Language Testing*, *9*, 101-119.

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement, 40,* 109-128.

Brozo, W. G., Schmelzer, R. V., & Spires, H. A. (1984). A study of testwiseness clues in college and university teacher-made tests with implications for academic assistance centers *(Technical Report 84-0*1). Georgia State University: College Reading and Learning Assistance. ERIC database (ED240928), http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED240928

Clark, E. L. (1956). General response pattern to five-choice items. *Journal of Educational Psychology, 47*, 110-117.

Clark, S. E., & Davey, S. L. (2005). The target-to-foils shift in simultaneous and sequential lineups. *Law and Human Behavior, 29,* 151-172.

Clegg, V. L., & Cashin, W. E. (1986). *Improving multiple-choice tests.* IDEA Paper No. 16. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Downing, S. M. (2006a). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287-301). Mahwah, NJ: Lawrence Erlbaum.

Downing, S. M. (2006b). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum.

Edwards, B. D. (2003). *An examination of factors contributing to a reduction in race-based subgroup differences on a constructed response paper-and-pencil test of achievement* (Unpublished doctoral dissertation). Texas A&M University.

Exam Innovations (2010). *Return on investment for discrete option multiple choice.* Exam Innovations, Inc. http://dl.dropbox.com/u/528723/ROI%20for%20DOMC.pdf

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.

Farley, J. K. (1989). The multiple choice test: Writing the questions. *Nurse Educator, 14,* 10-12, 39.

Foster, D., & Miller, H. L. (2009). A new format for multiple-choice testing: Discrete-option multiple-choice. Results from early studies. *Psychology Science Quarterly, 51*, 355-369.

Friel, S., & Johnstone, A. H. (1979). Does the position matter? *Education in Chemistry, 16,* 175.

German Medical Association (2004). *Regulations for Continuing Education and Continuing Education Certificate.* http://www.bundesaerztekammer.de/downloads/ADFBSatzungEn.pdf

German social security code (SGB). *§ 95d SGB V Pflicht zur fachlichen Fortbildung* [Obligation to professional training]. http://www.sozialgesetzbuch-sgb.de/sgbv/95d.html

Gibb, B. G. (1964). *Testwiseness as secondary cue response* (Doctoral dissertation). Stanford University, Ann Arbor, Michigan: University Microfilms, 1964. No. 64-7643.

Gustav, A. (1963). Response set in objective achievement tests. *Journal of Psychology, 56,* 421-427.

Haladyna, T. M. (2004). *Developing and validating multiple choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*, 17-27.

Hammond, E. J., McIndoe, A. K., Sansome, A. J., & Spargo, P. M. (1998). Multiple-choice examinations: Adopting an evidence-based approach to exam technique. *Anaesthesia, 53,* 1105-1108.

Holmes, P. (2002). *Multiple evaluation versus multiple choice as testing paradigm* (Unpublished doctoral dissertation). Twente University, Enschede.

Hossiep, R., & Schulte, M. (2008). *BOWIT - Bochumer Wissenstest* [*BOWIT - Bochum knowledge test*]. Göttingen: Hogrefe.

Jäger, A. O., & Fürntratt, E. (1968). *Differentieller-Wissens-Test - DWT* [*Differential Knowledge Test - DWT*]. Göttingen: Hogrefe.

Kingston, N. M., Foster, D., Miller, H. L., & Tiemann, G. C. (2010). *Building a better mousetrap: Using discrete options to improve the multiple-choice question.* Paper presented to the ATP Innovations in Testing Conference, February 7-10th, Orlando, Florida.

Kingston, N. M., Tiemann, G. C., Miller, H. L., & Foster, D. (2012). An analysis of the discrete-option multiple-choice item type. *Psychological Test and Assessment Modeling, 54,* 3-19.

Kühne-Eversmann, L., Nussbaum, C., Reincke, M., & Fischer, M. R. (2007). CME-Fortbildungsangebote in medizinischen Fachzeitschriften: Strukturqualität der MC-Fragen als Erfolgskontrollen [CME activities of medical journals: Quality of multiple-choice questions as evaluation tool]. *Medizinische Klinik, 102,* 993-1001.

Leiner, D. J. (2012). *SoSci Panel: The noncommercial online access panel.* Poster presented at the GOR 2012, March 6th, Mannheim. https://www.soscisurvey.de/panel/download/SoSciPanel.GOR2012.pdf.

Marcus, A. (1963). The effect of correct response location on the difficulty level of multiple-choice questions. *Journal of Applied Psychology, 47,* 48-51.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34, 207-218.

Mickes, L., Flowe, H., & Wixted, J. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied, 18,* 361-376.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of testwiseness. *Educational and Psychological Measurement*, *25*, 707-726.

Niedergethmann, M., & Post, S. (2006). Differentialdiagnose des Oberbauchschmerzes [The diagnosis and management of upper abdominal pain]. *Deutsches Ärzteblatt, 13,* A862-A871.

Popkin, S. L., & Dimock, M. A. (1999). Political knowledge and citizen competence. In S. L. Elkin & K. E. Soltan (Eds.), *Citizen competence and democratic institutions* (pp. 117-146). University Park: Pennsylvania State University Press.

Rapaport, G. M., & Berg, I. A. (1955). Response sets in a multiple-choice test. *Educational and Psychological Measurement, 15*, 58-62.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A metaanalysis of 80 years of research. *Educational Measurement: Issues and Practice, 24,* 3-13.

Rost, D. H., & Sparfeldt, J. R. (2007). Leseverständnis ohne Lesen? Zur Konstruktvalidität von multiple-choice-Leseverständnistestaufgaben [Reading comprehension without reading? On the construct validity of multiple-choice reading comprehension test items]. *Zeitschrift für Pädagogische Psychologie, 21,* 305-314.

Rotthoff, T., Fahron, U., Baehring, T., & Scherbaum, W. A. (2008). Die Qualität von CME-Fragen in der ärztlichen Fortbildung - eine empirische Studie [The quality of CME

questions as a component part of continuing medical education - an empirical study]. *Zeitschrift für ärztliche Fortbildung und Qualität im Gesundheitswesen, 101,* 667-674.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. *Review of Educational Research, 49,* 252-279.

Stagnaro-Green, A. S., & Downing, S. M. (2006). Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. *Medical Teacher, 28,* 566-568.

Steblay, N. K., Dysart, J. E., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 25*, 459-473.

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, *42*, 198-206.

Taylor, C., & Gardner, P. L. (1999). An alternative method of answering and scoring multiple choice tests. *Research in Science Education, 29,* 353-363.

Trepte, S., & Verbeet, M. (Eds.). (2010). *Allgemeinbildung in Deutschland. Erkenntnisse aus dem SPIEGEL-Studentenpisa-Test* [*General knowledge in Germany. Findings from the SPIEGEL students PISA test*]. Wiesbaden: VS Verlag.

Wevrick, L. (1962). Response set in multiple-choice test. E*ducational and Psychological Measurement, 22*, 533-538.

Woodley, K. K. (1975). *Test-wiseness: A cognitive function?* Paper presented at the annual meeting of the National Council on Measurement in Education, March 31-April 2, Washington, D.C.

Appendix: Original Research Articles

Article 1:

Willing, S., & Musch, J. (2013). *Evaluating the psychometric properties of a new method of knowledge assessment: Discrete-option multiple-choice testing.* Manuscript submitted for publication.

Article 2:

Willing, S., Ostapczuk, M., & Musch, J. (2013). *Do sequentially presented answer options prevent the use of testwiseness cues on continuing medical education tests?* Manuscript submitted for publication.

Article 3:

Papenberg, M., Willing, S., & Musch, J. (2013). *Sequentially presented answer options prevent the use of testwiseness cues in multiple-choice testing.* Manuscript submitted for publication.

Article 4:

Willing, S., & Musch, J. (2013). *The political knowledge test (PWT): Development and validation of a new instrument for assessing political knowledge.* Manuscript submitted for publication.

Evaluating the Psychometric Properties of a New Method of Knowledge Assessment: Discrete-Option Multiple-Choice Testing

Sonja Willing* and Jochen Musch

Institute of Experimental Psychology

University of Duesseldorf

* Correspondence concerning this article should be addressed to:

Sonja Willing

University of Duesseldorf

Institute of Experimental Psychology

Building 23.03

Universitaetsstr. 1

D-40225 Düsseldorf

Germany

Email: sonja.willing@hhu.de

Abstract

Multiple-choice (MC) tests are currently the most popular test format for the assessment of knowledge. Foster and Miller (2009) recently proposed the discrete-option multiple-choice (DOMC) format as an alternative computerized test format. DOMC tests are based on a sequential rather than simultaneous presentation of answer options. The present study examined the psychometric properties of DOMC tests and tested for potential serial position effects. When answer options were presented sequentially, item difficulty increased as the serial position of the correct answer increased. Items were also more difficult when the most attractive distractor was presented prior to the correct answer. This effect was small in the MC format and of medium size in the DOMC format. The reliability and validity of a sequential DOMC test was nevertheless comparable to that of a parallel MC test. Examinees identified MC items as easier and generally indicated a preference for the MC format over the DOMC format; however, they also viewed DOMC tests as superior with regard to the assessment of factual knowledge and as requiring a deeper understanding of the subject area. The benefits and drawbacks of DOMC testing are discussed.

Keywords:

discrete-option multiple-choice, multiple-choice, sequential item presentation, serial position effects, reliability, validity.

Evaluating the Psychometric Properties of a New Method of Knowledge Assessment: Discrete-Option Multiple-Choice Testing

The multiple-choice (MC) test format is one of the most valid and hence, popular testing formats for the assessment of knowledge. The MC format was first introduced during World War I as the general basis for the Army Alpha test, which allowed the U.S. Army to classify 1.5 million soldiers for military purposes (Downing, 2006). An MC item consists of an item stem followed by a list of several answer options. One of these answer options is keyed as the correct answer; the remaining answer options are distractors and are scored as incorrect. Test takers are requested to choose the correct answer option for each item.

Owing to its many advantages, the MC format is the most common selected-response item format (Downing, 2006). MC tests can be used for classroom testing as well as in large-scale testing programs for purposes of graduation, certification, licensure, evaluation, placement, and admission (Haladyna, 2004). MC items can be suitably applied to test all levels of learning and understanding, and they allow examiners to efficiently achieve a high degree of objectivity because the scoring procedures are automated (Clegg & Cashin, 1986; Downing, 2006). Most importantly, highly reliable and valid tests can be constructed using MC items (Downing, 2006).

However, MC testing also has several limitations. First, item writers have to develop incorrect yet plausible answer options, which can be difficult to create. Poorly written items containing testwiseness cues compromise the validity of MC items (Foster & Miller, 2009; Haladyna & Downing, 2004). When used repeatedly, MC items are also threatened by student memorization, copying, and sharing (Foster & Miller, 2009; Kingston, Foster, Miller, & Tiemann, 2010).

In the last several decades, many variants of the MC format have been proposed. These variants differ with regard to the number of answer options, the number of correct answer

options, the test administration procedure, and the scoring system (Haladyna, 2004). However, all of these variants offer the simultaneous presentation of answer options to the examinee.

By contrast, Foster and Miller (2009) recently introduced a computer-based alternative to MC testing called discrete-option multiple-choice (DOMC) testing. Analogous to an MC item, a DOMC item consists of a stem and several answer options, one of which has to be identified as the solution. However, there are two important differences between MC and DOMC testing. First, a DOMC item is defined by the sequential rather than simultaneous presentation of the answer options. Answer options are presented sequentially one at a time in a random order, and examinees have to decide on the correctness of each separately presented option. This procedure is implemented in a forward-only direction, and examinees are not given the opportunity to review items or to change previous answers (Foster & Miller, 2009; Kingston, Tiemann, Miller & Foster, 2012). Second, in DOMC testing, three stopping criteria are implemented. The processing of an item ends and no further answer options are presented after the examinee has answered an item either correctly or incorrectly. Whereas an item can be solved correctly only by identifying the correct answer as the solution, there are two ways for an examinee to provide an incorrect answer: first, to erroneously reject the solution as a distractor and second, to mistakenly select a distractor as the solution (Foster & Miller, 2009). If one of these three stopping criteria is met, the item is scored. No further answer option is presented after the correct answer option because one of the stopping criteria is necessarily met after the correct answer is presented (i.e., it is either correctly accepted or incorrectly rejected; Foster & Miller, 2009).

In the domain of knowledge assessment, there are several potential advantages to a sequential testing procedure. Owing to the stopping criteria, only half of the answer options have to be presented for each item on average. DOMC testing thereby improves test

efficiency and reduces the danger that test items will be copied or shared because many answer options do not have to be presented at all (Foster & Miller, 2009). Willing, Ostapczuk, and Musch (2013) also found that, owing to the sequential presentation of answer options, DOMC was able to reduce the use of testwiseness cues. If construct-irrelevant variance due to testwiseness can be controlled using a sequential presentation of answer options, an increase in test reliability and validity may be expected. This is why we decided to conduct an investigation into the psychometric properties of DOMC tests.

It is interesting to note that there is one area of research that bears some striking resemblance to sequential MC testing, namely, the use of sequential police lineups in eyewitness research. In the usual simultaneous police lineups, which largely prevail in current practice, the witness is presented with a number of individuals and is asked to indicate the person who committed the crime, if present. However, this kind of police lineup, which parallels the usual MC procedure, has been criticized because witnesses may simply decide to choose the lineup member who most resembles the perpetrator without making an absolute comparison between their memory of the image of the perpetrator and each lineup member. Similarly, MC testing in knowledge assessment has also been criticized as allowing respondents to simply choose the most plausible alternative rather than to identify the solution with some certainty. In a sequential lineup procedure, as in DOMC testing, the eyewitness is therefore presented with one lineup member at a time and has to decide whether or not that person is the perpetrator before being allowed to view the next member. The idea behind this one-at-a-time procedure is to discourage the eyewitness from relying on a relative judgment and to simply decide whether each suspect looks like the perpetrator. In a sequential lineup, an eyewitness may decide that the current lineup member looks more like the perpetrator than the one before, but he or she cannot be sure whether the next person will not look even more like the perpetrator. A sequential lineup can therefore be used to force eyewitnesses to make a

real decision about whether or not they have identified the perpetrator. In a similar vein, Foster and Miller (2009) surmised that in the domain of knowledge assessment, DOMC testing might motivate deeper learning because the correct answer option has to be identified without the help of accompanying distractors. In a meta-analysis comparing the two competing suspect-identification procedures across 30 studies, Steblay, Dysart, Fulero, and Lindsay (2001) found that sequential lineups were superior to simultaneous lineups for overall correct decisions (.56 vs. .48).

More recent research has questioned this judgment, however. This research was conducted on the basis of signal detection analyses of the lineup procedure, in which some participants viewed a lineup in which the suspect was, in fact, the perpetrator, whereas other participants viewed a lineup in which the suspect was an innocent person who resembled the perpetrator. In such experiments, a hit rate can be computed as the proportion of target-present lineups from which the guilty suspect is correctly identified, and a false alarm rate can be computed as the proportion of target-absent lineups from which the innocent suspect is incorrectly identified. In their later review of the diagnostic performance of simultaneous and sequential lineups, Steblay, Dysart, and Wells (2011) reported an average hit rate and false alarm rate for the simultaneous lineup procedure of .52 and .28, respectively. For the sequential lineup procedure, they reported values of .44 and .15, respectively. Thus, on average, the sequential procedure yielded both a lower hit rate and a lower false alarm rate. In terms of signal detection analysis, this pattern does not allow one of the procedures to be identified as superior. Mickes, Flowe, and Wixted (2012) therefore constructed receiver operating characteristics to allow for a more direct comparison of the diagnostic performances of the two procedures by computing the area under the respective ROC curve. Contrary to virtually all previous research, they found that the sequential procedure was inferior to the simultaneous procedure in discriminating between the presence versus absence of a guilty

suspect in a police lineup. Mickes et al. (2012) therefore argued that employing a sequential procedure may simply have induced a more conservative response criterion, which, however, would have no direct bearing on the superiority of one of the two procedures. Consequently, they asked police departments not to prematurely switch to a sequential presentation procedure.

To summarize, research on the sequential versus simultaneous presentation of suspects in a police lineup has provided somewhat equivocal results and has not yet arrived at a definite conclusion. It is also important that given the visual nature of the lineup task, results obtained in this paradigm cannot be readily transferred to the case of a sequential versus simultaneous presentation of answer options in MC testing. Therefore, our main goal was to do some fundamental research on the effects that a sequential testing procedure would have on item difficulty and to determine the psychometric properties of DOMC testing. In doing so, however, it is important to note that unlike in typical eyewitness experiments that apply the lineup procedure, there is always a solution - equivalent to a guilty suspect - for MC items. For this reason, the signal detection approach proposed by Mickes et al. (2012) could not be applied to the present research question.

To arrive at a balanced assessment, the second goal of our study was to determine whether DOMC testing also has some potential drawbacks that might interfere with reliable and valid measurement. One important potential drawback of DOMC testing is the possibility of serial position effects. The order in which answer options are presented and in particular, the position of the solution, might be critical and might influence the psychometric quality of a DOMC test. No previous work has investigated potential order effects for DOMC items. However, for traditional MC items with a simultaneous presentation of the answer options, some research has been conducted on positional effects. Under conditions of time pressure, Clark (1956) observed that some participants had a tendency to neglect the later positions,

which he interpreted as a failure to read all answer options before giving a response. Several other studies have also reported effects of the position of the correct answer on the difficulty of MC items (Gustav, 1963; Rapaport & Berg, 1955). However, such biases were typically weak (Attali & Bar-Hillel, 2003) and were difficult to interpret because of methodological and conceptual problems. In particular, previous studies on serial position effects in MC testing failed to randomize the position of the correct answer option, making it difficult to interpret the observed effects (Fagley, 1987).

Several researchers have surmised that rather than the position of the correct answer option, the relative position of the correct answer and of the most attractive distractor may influence the difficulty of MC items (Friel & Johnstone, 1979; Marcus, 1963). Friel and Johnstone (1979) found that presenting the most attractive distractor immediately prior to the correct answer reduced the difficulty of MC items. Thus, somewhat surprisingly, placing the most plausible distractor immediately before the solution did not attract responses away from the solution; rather, a close proximity of the solution and the most plausible distractor seemed to help test takers to better discriminate between the distractor and the solution. However, the positions of the solution and the most attractive distractor were not systematically varied by Friel and Johnstone (1979), and an opposite effect was found by Clark and Davey (2005) in two police lineup studies. In sequential lineups, presenting foils who were very similar to the target prior to the presentation of the target led to an increased number of incorrect identifications of the foil. Clark and Davey (2005) argued that when a very similar foil is presented after the target in a sequential procedure, the identification of the target is no longer complicated by the necessity of rejecting the similar foil first. If, however, a very similar foil is presented prior to the target, this will most likely result in some erroneous identifications of the foil. In DOMC testing, a similar effect is easily conceivable because unlike in MC testing, a decision sometimes has to be made about the most plausible distractor before the solution is

presented to the test taker. If it is more difficult for an examinee to dismiss the most attractive distractor in DOMC testing, presenting the most attractive distractor prior to the solution may increase the number of false alarms and may thus harm the psychometric properties of the test.

To summarize, the primary purpose of the present study was to examine whether the psychometric properties of test items would be improved by a sequential presentation of answer options due to a better control of the use of unwanted cues to the solution because these cues can only be used when answer options are presented simultaneously (Willing et al., 2013). To this end, we investigated the reliability and convergent validity of the DOMC format. Cronbach's α was computed as an estimate of internal consistency. Convergent validity was examined by correlating the MC and DOMC scores with external criteria that were relevant to the construct of political knowledge that was tested using either the MC or the DOMC format. We expected that the DOMC scores would be more strongly associated with relevant external criteria than the MC scores due to the reduction of construct-irrelevant variance that would presumably be achieved by employing a sequential presentation procedure.

The second purpose of the present study was to investigate potential moderators of item difficulty in DOMC testing. DOMC testing likely increases mean item difficulty as compared with MC testing because the decisions of the test taker have to be made on the basis of the more limited information that is provided by the single options that are presented on DOMC tests. We therefore expected that DOMC items would be more difficult than MC items. We also wanted to examine effects of the serial position of the solution and the serial position of the most attractive distractor relative to the solution. We expected that item difficulty would increase in the DOMC format as the serial position of the correct answer increased because over the course of a sequential decision procedure, it may be difficult for the test taker to

dismiss up to three distractors that are presented prior to the solution. However, additional false alarms may also potentially result in an increased difficulty of DOMC items. Therefore, we also expected that the serial position of the most attractive distractor would influence the difficulty of DOMC items as has previously been reported in police lineup research. In particular, given that the erroneous acceptance of a distractor is most likely to occur when the distractor is highly attractive, we expected that item difficulty would increase in the DOMC format when the most attractive distractor was presented prior to the correct answer option. Such an effect would challenge the application of the DOMC format or at least require a careful monitoring of serial position effects. However, we also wanted to determine whether such serial position effects, if present, would interfere with the valid measurement of the subject matter under investigation. This is not necessarily the case because, if less able participants fall prey to serial position effects more easily and therefore achieve a lower score, this is exactly what is expected from a valid measurement procedure.

As new answer formats are less familiar to the respondents, these foreign answer formats are frequently met with both curiosity and skepticism. When answer options are presented sequentially, less information is available to the test taker who also has to make more decisions per item than for items with a simultaneous presentation of answer options. The increased difficulty of DOMC items that was reported by Foster and Miller (2009) may well result in more favorable evaluations of the MC format. However, in an exploratory investigation, Kingston et al. (2010) observed a generally positive response of student test takers to the DOMC test format. These participants also felt that DOMC testing made cheating and test theft more difficult. We wanted to expand on this evaluation of the acceptability of DOMC testing; therefore, we evaluated the perceived difficulty, face validity, stimulation of deeper learning, and general attractiveness of the two answer formats.

To address our research questions, we randomly assigned participants to either the MC or the DOMC condition. In each condition, every test taker was asked to answer political knowledge items. Using a computer-based presentation, the positions of the correct answer option and the most attractive distractor were varied.

Methods

Participants. 4,490 (2,653 female, 59.09%) native German speakers were recruited via the SoSci Panel, an open panel for the recruitment of participants for scientific investigations (Leiner, 2012). Their ages ranged from 15 to 92 years (M = 31.96, SD = 12.83). When asked about their highest academic degree, participants indicated a Ph.D. (2.7%), a master's degree (28.4%), a bachelor's degree (17.1%), or a final secondary-school examination (the German "Abitur"; 40.3%). Another 10.7% of the participants held a junior high school diploma, and 0.8% participants had not graduated from junior high school. The number of years they spent in school and university education ranged from 7 to 38 years (M = 16.07, SD = 3.33). Participants were randomly assigned to either the MC condition (N = 2,283) or the DOMC condition (N = 2,207). Across conditions, participants did not differ regarding (i) their average age ($t(4488) = 1.18$, $p = .24$, $d = 0.04$), (ii) the length of their education ($t(4488) = 0.11$, $p = .92$, $d = 0.01$), or (iii) their self-assessed intelligence ($t(4488) = 0.81$, $p = .42$, $d = 0.02$).

Design. The study used an incomplete 2 x 4 x 2 mixed design with the independent between-subjects variable *answer format* (MC, DOMC), the within-subjects variable *position of the correct answer option* (1, 2, 3, 4), and the within-subjects variable *position of the most attractive distractor* (before the correct answer, after the correct answer). Some cells in this design were empty because the most attractive distractor could not be presented prior to the correct answer if the correct answer was presented as the first option or after the correct answer if the correct answer was presented as the last option. Mean item difficulty, measured

by the proportion of correct solutions, and mean item discrimination, measured by the part-whole corrected correlation between item score and total test score, were computed as dependent variables. Additionally, test score reliability, validity, and acceptability were determined for both answer formats.

Material. For the present study, a set of items with known properties was needed. Therefore, we developed a pretest to administer eight political knowledge items to a sample of 258 panelists who did not participate in the main study. One of the four answer options in each item was keyed as the correct answer; the remaining three answer options were scored as incorrect. Averaged across all items, the proportions of respondents who chose the solution, the most attractive distractor, the second most attractive distractor, and the least attractive distractor were .67, .24, .05, and .04, respectively. Thus, across all items, the most frequently chosen alternative was the solution, and there was one distractor that was clearly more attractive than the remaining distractors. Using these items, we were able to systematically investigate the impact of the positions of the correct answer and the most attractive distractor relative to the solution. The mean item discrimination of the political knowledge items ranged from .32 to .53 (M = .41, SD = .06). An example item is: "Where does the International Court of Justice reside?" (The Hague* / Strasbourg / Luxembourg / Brussels).

Instruments. Political knowledge items from general knowledge tests and various measures of self-assessed political knowledge were used as external criteria to determine the convergent validity of the two answer formats.

Tests of political knowledge. As a first external criterion, political knowledge was assessed using adaptions of nine of the 14 items of the subscale "politics and society" of Bochum's general knowledge test (henceforth: BOWIT; Hossiep & Schulte, 2008). An example item is: "Which country is not a member of the European Union?" (Norway* / Finland / Austria / Denmark). The remaining five items on the subscale dealt with society

rather than politics and were therefore excluded. The BOWIT political knowledge items were presented in an MC format and in a random order. Nine additional political knowledge items were taken from the Spiegel PISA student test (Trepte & Verbeet, 2010). They were also presented in an MC format and in a random order to obtain a second independent external criterion of political knowledge. One example item is: "Which country is not a permanent member of the United Nations Security Council?" (Germany* / France / Great Britain / China).

Self-ratings of intelligence and knowledge. To obtain three measures of cognitive ability as additional external validity criteria, participants were asked to compare their self-assessed intelligence, their self-assessed general knowledge, and their self-assessed political knowledge with that of the adult population at large. To this end, they were asked to estimate the percentage of the German adult population they considered likely to be more intelligent, to have better general knowledge, and to have better political knowledge than themselves.

Self-reported measures of interest in politics and participation in political discussions. Participants' interest in politics was assessed using a 7-point Likert scale ranging from very low (1) to very high (7). Additionally, participants were asked to indicate how often they participated in political discussions in their circle of friends and family, using a 7-point Likert scale ranging from very rarely (1) to very often (7).

Political media consumption. To obtain an index of the participants' consumption of political media, we asked them for a self-report of the number of political books they had read over the course of the last three years and the number of hours they had spent reading the political sections of a newspaper, magazine, or online service over the course of the past week.

Education. As an additional external criterion, we asked for the number of years participants had spent in school and university education and for the last grade they obtained in "politics and the social sciences."

Evaluation of DOMC's acceptability. The acceptability of the DOMC format was assessed by asking for the perceived item difficulty ("Which answer format makes it easier to solve an item correctly?"), face validity ("Which answer format do you think is superior for assessing factual knowledge?"), encouragement of deeper learning ("In your opinion, which answer format requires a deeper understanding of the subject?"), and attractiveness to the respondent ("Which answer format do you prefer?"; "Which answer format would you prefer for a high-stakes test?"; "In your opinion, which answer format is more fun?"). Participants were asked to indicate their evaluations on a 7-point Likert scale ranging from MC (1) to DOMC (7).

Procedure. The questionnaire was delivered online using the software Unipark (Version 8; Globalpark AG, 2012). At the beginning of the questionnaire, participants completed self-ratings of their intelligence and their general and political knowledge. Afterwards, they provided a self-assessment of their interest in politics, their participation in political discussions, and their political media consumption. Subsequently, participants worked on the political BOWIT and PISA items that were both presented in an MC format to obtain external criteria for the purpose of validating the two answer formats. Next, participants were randomly assigned to either the DOMC or the MC condition and completed the political knowledge items in which the position of the correct solution and the position of the most attractive distractor relative to the solution were varied. In the MC condition, one item was presented per page along with all possible answer options. In the DOMC condition, answer options were presented sequentially. As is usual also in police lineup procedures, in either condition, respondents were not given the opportunity to review previous options or items.

The DOMC procedure was explained using a sample item. Additionally, in the DOMC condition, participants were told about all item stopping criteria that were employed for the sequential presentation of answer options. In both the MC and DOMC conditions, items were presented in a random order. Within items, answer options were also presented in a random order. After completing the political knowledge items, participants in the DOMC condition worked on the six items that asked for their ratings of the acceptability of the DOMC format relative to the MC format. Participants in the MC condition did not answer these questions because they had not been introduced to the DOMC format and were thus unable to provide ratings for this procedure. Next, all participants were asked to indicate their age, sex, the number of years they spent in school and university education, as well as their last grade in "politics and the social sciences." To thank them for their participation, all individuals received feedback on their performance after they completed the test.

Results

In all statistical tests, an alpha level of .05 was used. ANOVA effect sizes were computed using eta-squared (η^2), which can be interpreted as the proportion of the variance explained by each factor or interaction. $\eta^2 \geq 0.01$ implies a small effect, $\eta^2 \geq 0.06$ a moderate effect, and $\eta^2 \geq 0.14$ a large effect (Cohen, 1988). Effect sizes for the difference between two means were calculated using Cohen's *d*. According to Cohen (1988), an effect of $d \geq 0.20$ is considered small, an effect of $d \geq 0.50$ is considered medium, and an effect of $d \geq 0.80$ is considered large. Effect sizes in 2 x 2 contingency tables were computed using the phi coefficient (φ) as a measure of the correlation between the two variables. According to Cohen (1988), $\varphi \geq 0.10$ implies a small effect, $\varphi \geq 0.30$ a medium effect, and $\varphi \geq 0.50$ a large effect. Effect sizes for the relation between two or more variables measured on a nominal scale were

computed as Cramer's V, for which the same effect size conventions according to Cohen (1988) apply.

Order of answer options. The proportions of choices between the four answer options of the eight political knowledge items were .63, .14, .08, and .07 for the correct solution, the most attractive distractor, the second most attractive distractor, and the least attractive distractor, respectively. Thus, for each item, there was a distractor that was more attractive than the remaining distractors, confirming the results of our pretest. Consequently, using these items, the impact of the position of the correct answer and the most attractive distractor on item difficulty and item discrimination could be investigated.

Item difficulty. A t-test was computed to compare the test scores between the DOMC and MC conditions. Participants in the MC condition correctly solved 5.27 of the 8 items (*SD* = 1.54). Participants in the DOMC condition correctly solved only 4.77 of the 8 items (*SD* = 1.72). This difference was statistically significant, $t(4488) = 10.27$, $p < .001$, $d = 0.31$. Thus, test difficulty was higher for the DOMC test format presumably because decisions have to be made on the basis of a smaller amount of information on DOMC tests.

Item difficulty as a function of the position of the correct answer. The average proportion of correctly solved items as a function of the position of the correct answer option across conditions is presented in figure 1.

--- Insert figure 1 here ---

We computed χ^2 tests to explore whether the proportion of correct solutions varied as a function of the position of the solution. Although this was true for both the MC condition, $\chi^2(3) = 9.09$, $p = .03$, and the DOMC condition, $\chi^2(3) = 383.30$, $p < .001$, the influence of the position of the solution varied as a function of the test format. The magnitude of the effect

was small to medium in the DOMC condition (Cramer's V = 0.15) and very small in the MC condition (Cramer's V = 0.02). In the MC condition, the proportion of correct solutions differed significantly only between positions 1 and 4, $\chi^2(1) = 7.47$, $p < .01$, $\varphi = 0.03$, and between positions 2 and 4, $\chi^2(1) = 5.34$, $p = .02$, $\varphi = 0.02$. In the DOMC condition, all of the six possible pairwise comparisons between the four serial positions were statistically significant. An item was solved more frequently when the solution was presented in the first rather than in the second position, $\chi^2(1) = 21.28$, $p < .001$, $\varphi = 0.05$, when the solution was presented in the first rather than in the third position, $\chi^2(1) = 136.80$, $p < .001$, $\varphi = 0.13$, and when the solution was presented in the first rather than in the fourth position, $\chi^2(1) = 328.45$, $p < .001$, $\varphi = 0.19$. An item was also solved more frequently when the solution was presented in the second rather than in the third position, $\chi^2(1) = 51.18$, $p < .001$, $\varphi = 0.08$, and when the solution was presented in the second rather than in the fourth position, $\chi^2(1) = 185.89$, $p < .001$, $\varphi = 0.15$. Last, a DOMC item was also correctly solved more often if the solution was presented in the third rather than in the fourth position, $\chi^2(1) = 41.72$, $p < .001$, $\varphi = 0.07$. Thus, taken together, solution frequency decreased as the serial position of the solution increased in both answer formats. However, this effect was significantly larger in the DOMC condition as was confirmed by a planned contrast that showed a significant difference between the linear trends in a 2 (MC vs. DOMC) x 4 (serial position) repeated-measures ANOVA, $F(1, 8978) = 653.80$, $p < .001$.

Item difficulty as a function of the position of the most attractive distractor. To investigate whether item difficulty varied as a function of the position of the most attractive distractor, a mixed between-within subjects ANOVA with answer format (MC vs. DOMC) as the between-subjects factor and the position of the most attractive distractor (before vs. after the correct answer) as the within-subjects factor was conducted. The position of the solution could not be added to this ANOVA as an additional factor. This was because, as explained

above, an incomplete design had to be used to accommodate the fact that the manipulation of the position of the most attractive distractor relative to the solution imposes a constraint on the serial position in which the solution can be presented. As can be seen in figure 2, there was a substantial main effect of the answer format, indicating that participants in the MC condition solved a greater proportion of items correctly (M = .66, SD = .19) than participants in the DOMC condition (M = .60, SD = .21). This difference was statistically significant, $F(1, 4469) = 98.71$, $p < .001$, $\eta^2 = 0.02$. The proportion of correctly solved items also increased as a function of the position of the most attractive distractor. Items for which the most attractive distractor was presented only after the solution were correctly solved more often (M = .67, SD = .26) than items for which the most attractive distractor was presented before the solution (M = .58, SD = .28). This effect of the position of the most attractive distractor was significant, $F(1, 4469) = 407.71$, $p < .001$, $\eta^2 = 0.08$. Most importantly, there also was a significant interaction between the answer format and the position of the most attractive distractor, $F(1, 4469) = 246.15$, $p < .001$, $\eta^2 = 0.05$. To investigate this interaction more closely, additional paired-samples t-tests were computed. In the DOMC condition, we found a significantly higher proportion of correctly solved items if the most attractive distractor was presented after the solution (M = .68, SD = .27) than if the most attractive distractor was presented before the solution (M = .51, SD = .30), $t(2187) = 22.77$, $p < .001$, d = 0.60. Even in the MC condition, however, we found a significantly higher proportion of correctly solved items if the most attractive distractor was presented after the solution (M = .67, SD = .24) than if the most attractive distractor was presented before the solution (M = .65, SD = .25), $t(2282) = 3.56$, $p < .001$, d = 0.08. Although presenting the most attractive distractor before the solution led to a decrease in the proportion of correctly solved items in both conditions, this decrease was considerable larger in the DOMC condition.

--- Insert figure 2 here ---

Item discrimination and reliability. A t-test was computed to compare item discrimination across conditions. The part-whole corrected correlations between the items and the total test performance in the MC condition ranged from .18 to .33 (M = .25, SD = .06). In the DOMC condition, the respective correlations ranged from .20 to .31 (M = .27, SD = .04). The difference in mean discrimination was not statistically significant, $t(14) = 0.51$, $p = .62$, $d = 0.25$. Thus, although item difficulty was higher on the DOMC test, there was no reduction in the mean item discrimination.

As an estimate of the reliability of the tests, Cronbach's coefficient α (Cronbach, 1951) was calculated. The program AlphaTest by Lautenschlager and Meade (2008) was applied to test for differences in coefficient α across conditions. We found that the items in the MC condition (α = .54) did not differ from the items in the DOMC condition (α = .55) with regard to their internal consistency, $\chi^2(1) = 0.76$, $p = .38$. Thus, presenting item answers sequentially did not increase or decrease test reliability.

Convergent validity. Table 1 displays the correlations of the MC and DOMC test scores with external validation criteria. Tests for comparing independent correlations (Cohen, Cohen, West & Aiken, 2003) showed that none of these correlations differed significantly between the MC and DOMC conditions (see table 1). Thus, DOMC scores were not associated more strongly with external criteria than MC scores. For all external criteria, the convergent validity of DOMC test scores was comparable to that of MC test scores.

--- Insert table 1 here ---

Evaluation of DOMC's acceptability. A one-sample t-test was computed to test whether the test takers' evaluations of the test format differed from a mean value of 4 toward either the MC end of the scale (1) or the DOMC end of the scale (7). Participants indicated that they viewed DOMC as superior to MC with regard to the assessment of factual knowledge (M = 4.81, SD = 2.11, $t(2206) = 18.05$, $p < .001$, $d = 0.54$). They also considered DOMC testing to require a deeper understanding of the subject matter (M = 5.80, SD = 1.63, $t(2206) = 52.00$, $p < .001$, $d = 1.56$). However, participants strongly believed that it was easier to solve an item when it was presented in the MC format (M = 2.14, SD = 1.56, $t(2206) = 55.93$, $p < .001$, d = 1.69). For the three items regarding the relative attractiveness of the two answer formats, participants indicated that they strongly preferred the MC format (M = 2.69, SD = 1.77; $t(2206) = 34.76$, $p < .001$, $d = 1.05$). This preference for MC also held for high-stakes tests (M = 2.21, SD = 1.69; $t(2206) = 49.96$, $p < .001$, $d = 0.22$). Finally, participants rated the MC format as more fun (M = 3.62, SD = 2.02), $t(2206) = 8.74$, $p < .001$, $d = 0.27$.

To investigate whether ability moderated these effects, the sample was split into two groups of low versus high scorers at the median test score of 5. The 995 low scorers with a score of up to 4 points achieved a mean score of 3.21 (SD = 0.94), which was considerably below the mean score of the 1,212 high scorers who had scores of 5 or more points (M = 6.05, SD = 1.02, $t(2205) = 67.42$, $p < .001$, $d = 2.90$). Low scorers rated the DOMC format as worse for the assessment of factual knowledge than high scorers (M = 4.61, SD = 2.20 vs. M = 4.97, SD = 2.02, $t(2205) = 4.08$, $p < .001$, $d = 0.17$). A difference between low and high scorers was also found for the three items that were used to assess the attractiveness of the test. First, participants with low scores on the DOMC test showed a greater preference for the MC format than high scorers (M = 2.47, SD = 1.71 vs. M = 2.87, SD = 1.80, $t(2205) = 5.26$, $p < .001$, $d = 0.23$). For high-stakes testing, low scorers also indicated a stronger preference for the MC format than high scorers (M = 2.08, SD = 1.60 vs. M = 2.31, SD = 1.75), $t(2205)$ =

3.07, $p < .01$, $d = 0.14$. And finally, low scorers also rated the MC format as more fun than the DOMC format ($M = 3.37$, $SD = 2.03$ vs. $M = 3.83$, $SD = 1.97$, $t(2205) = 5.40$, $p < .001$, $d = 0.23$). Other differences in the evaluations of low and high scorers did not achieve statistical significance.

Discussion

Discrete-option multiple-choice testing is based on a sequential rather than a simultaneous presentation of answer options. The main goal of the present study was to investigate whether the psychometric properties of a test would be improved by employing this alternative test format for MC testing. To address this question, we compared item difficulty and discrimination as well as reliability and validity for a test that was randomly administered in either the MC or the DOMC format. Additionally, we evaluated the acceptability of the DOMC format. To test for serial position effects, we varied the positions of the correct answer and the most attractive distractor.

One finding was that MC items were easier to answer correctly than DOMC items. This is probably because for the latter, a simultaneous comparison of all answer options is not possible and the test taker therefore has to base his decision on the much more limited amount of information that is provided by single options (Foster & Miller, 2009; Kingston et al., 2012). Over the course of this sequential decision making, it seems to be difficult for test takers to dismiss up to three distractors and to wait for the correct answer, resulting in an increased number of false alarms in the DOMC format. Accordingly, item difficulty depended on the position of the correct answer when tests were conducted sequentially, and item difficulty increased as the serial position of the correct answer increased. The magnitude of this effect was small to medium according to Cohen (1988). However, an effect of the position of the correct answer on item difficulty was also observed in the MC condition, in

which answers were presented simultaneously. A likely explanation for this finding is that some participants failed to read all options thoroughly, resulting in a small response bias toward the early options. Such a response bias necessarily results in a somewhat lower proportion of correct solutions when the correct answer is presented as one of the later options (Fagley, 1987). However, in the MC condition, this effect was of relatively small magnitude and statistically significant only because of the large sample size used in the present study. It is also important to note that in spite of these format-specific effects, the reliability and validity of the sequential DOMC test did not differ from that of the parallel MC test.

As in sequential police lineups, we found that the position of the most attractive distractor also influenced the difficulty of the task. If an item contained a very attractive distractor that was considered correct by many examinees, it was difficult for the test takers to dismiss this distractor if it was presented prior to the correct answer. Whereas the magnitude of this effect was rather small in the MC format, it was of medium size in the DOMC format, presumably because under a sequential presentation procedure, the potentially even more attractive solution was not yet available to the test taker when he had to make a decision about the correctness of the attractive distractor. Thus, in contrast to the findings by Friel and Johnstone (1979), who incidentally did not randomize the position of the correct answer, the item difficulty of MC items did not decrease in our study if the most attractive distractor was presented immediately before the solution. Rather, we observed the exact opposite effect: Item difficulty was higher for the MC format if the most attractive distractor was presented before the correct answer option. However, this effect was of very small magnitude in the MC format and therefore seems rather negligible. A likely explanation for the small effect of the serial position of the most attractive distractor relative to the solution is a failure of some of the participants to read all answer options before providing their response. If a highly

attractive distractor is presented prior to the solution, these participants are likely to wrongly accept it and never even reach the correct answer.

Willing et al. (2013) found that the DOMC test format allowed for a better control of testwiseness than the MC format; therefore, the DOMC format can potentially improve the fairness of knowledge assessment (Foster & Miller, 2009). The present results also show, however, that the difficulty of DOMC items is influenced by the serial position of the correct answer and by the position of the most attractive distractor relative to the correct answer. A careful consideration of such effects therefore seems necessary. If the order in which answer options are presented is randomized, the attractiveness of the distractors shown to the examinees will vary. It is therefore possible that in spite of having the same ability, two examinees will receive different test scores. This poses a potential threat to the fairness and validity of test scores and may make it necessary to hold the order of answer options constant within items. However, even though no such control of the order of answer options was used in the present study, we found no difference between the reliability and validity of the sequential DOMC procedure and the simultaneous MC testing procedure. This finding suggests that the small to medium serial position effects we observed are not harmful to the psychometric quality of DOMC tests if the position of answer options is randomized for all items and all participants. On the other hand, there was also no superiority of the psychometric properties of DOMC tests in spite of the better control that DOMC testing affords for construct-irrelevant variance due to testwiseness (Willing et al., 2013). Counter to our expectations, no significant differences in convergent validity with regard to a number of external criteria were observed between MC and DOMC testing. Although DOMC items were generally more difficult, their discrimination and internal consistency did not differ from those of MC items.

The second goal of the present study was to take a closer look at the acceptability of the DOMC format. Examinees rated MC items as easier, and potentially for this reason, they generally indicated a preference for the MC format over the DOMC format. The notion that a perceived higher difficulty of DOMC items is responsible for this preference is supported by the finding that low-scoring participants showed an even stronger preference for the MC test format. However, participants also rated DOMC tests as superior with regard to the assessment of factual knowledge and as requiring a deeper understanding of the subject area. Arguably, participants recognized that searching for cues and relying on testwiseness to identify the solution is less helpful in DOMC than in MC testing. If our participants have a point in believing that the DOMC format requires a deeper understanding of the subject area at hand, it seems worthwhile to investigate in an instructional context whether the expectation of having to take a test with a sequential answer format is capable of encouraging deeper learning and more careful preparation among test takers.

Taken together, with the availability of the new DOMC test format, examiners are faced with the necessity of making a decision. Is the implementation of DOMC testing worth the extra effort? Potential benefits of DOMC testing come at a price. First, DOMC testing increases item difficulty as a function of the positions of the correct answer option and the most attractive distractor. Even though in the present study, such serial position effects did not harm the psychometric properties of the test, they potentially challenge the application of the DOMC format. A careful monitoring of serial position effects and further investigations into whether they moderate the reliability and validity of DOMC tests are therefore desirable. It is also a drawback of the DOMC format that test administration software is needed, posing an additional burden to the examiner even though the costs of implementation are one-time expenses during the initial stages of implementation (Exam Innovations, 2010). However, many tests are now being conducted online anyway and can easily be adapted to a sequential

presentation procedure. Therefore, it may be more important that several advantages associated with the use of DOMC testing have been found. First, the number of answer options that are presented per item is reduced, thereby enhancing test security (Foster & Miller, 2009; Willing, 2013). Second, the DOMC format provides the opportunity to increase item difficulty without changing item discrimination. DOMC testing thereby allows test creators to construct more challenging and demanding tests that are perceived as more difficult by the examinees, who may thereby be motivated to process the information on a deeper level and prepare for the exam more carefully. Third and perhaps most importantly, Willing et al. (2013) observed that a sequential presentation procedure allows for a better control of testwiseness than traditional MC tests in continuing medical education (CME) programs aimed at developing and maintaining the knowledge of professionals in the medical field. Willing et al. (2013) found that DOMC outperformed MC in preventing the identification of cues on tests that had been criticized for being susceptible to testwiseness strategies (Rotthoff, Fahron, Baehring & Scherbaum, 2008). This unique advantage of DOMC over MC testing deserves attention because it addresses a major concern about the use of MC testing in general.

In spite of the different drawbacks and advantages of the two answer formats, the reliability and validity of a sequential DOMC test were found to be equivalent to those of a parallel MC test in the present study. Thus, the psychometric properties of DOMC testing did not surpass but were able to match those of the format hitherto considered to be the most valid for an objective assessment of knowledge. In view of some of its unique new features, it therefore certainly seems worthwhile to further evaluate the usefulness of this new answer format.

References

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement, 40,* 109-128.

Clark, E. L. (1956). General response pattern to five-choice items. *Journal of Educational Psychology, 47*, 110-117.

Clark, S. E., & Davey, S. L. (2005). The target-to-foils shift in simultaneous and sequential lineups. *Law and Human Behavior, 29,* 151-172.

Clegg, V. L., & Cashin, W. E. (1986). *Improving multiple-choice tests.* IDEA Paper No. 16. Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/ correlation analysis for the behavioral sciences.* Mahwah, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287-301). Mahwah, NJ: Lawrence Erlbaum.

Exam Innovations (2010). *Return on investment for discrete option multiple choice.* Exam Innovations, Inc. http://dl.dropbox.com/u/528723/ROI%20for%20DOMC.pdf

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.

Foster, D., & Miller, H. L. (2009). A new format for multiple-choice testing: Discrete-option multiple-choice. Results from early studies. *Psychology Science Quarterly, 51,* 355-369.

Friel, S., & Johnstone, A. H. (1979). Does the position matter? *Education in Chemistry, 16,* 175.

Globalpark AG (2012). *Enterprise Feedback Suite. EFS Survey.* http://www.unipark.info

Gustav, A. (1963). Response set in objective achievement tests. *Journal of Psychology, 56,* 421-427.

Haladyna, T. M. (2004). *Developing and validating multiple choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*, 17-27.

Hossiep, R., & Schulte, M. (2008). *BOWIT - Bochumer Wissenstest* [*Bochum's knowledge test*]. Göttingen: Hogrefe.

Kingston, N. M., Foster, D., Miller, H. L., & Tiemann, G. C. (2010). *Building a better mousetrap: Using discrete options to improve the multiple-choice question.* Paper presented to the ATP Innovations in Testing Conference, February 7-10th, Orlando, Florida.

Kingston, N. M., Tiemann, G. C., Miller, H. L., & Foster, D. (2012). An analysis of the discrete-option multiple-choice item type. *Psychological Test and Assessment Modeling, 54,* 3-19.

Lautenschlager, G. J., & Meade, A. W. (2008). AlphaTest: A windows program for tests of hypotheses about coefficient alpha. *Applied Psychological Measurement, 32,* 502-503.

Leiner, D. J. (2012). *SoSci Panel: The noncommercial online access panel.* Poster presented at the GOR 2012, March 6th, Mannheim. https://www.soscisurvey.de/panel/download/SoSciPanel.GOR2012.pdf.

Marcus, A. (1963). The effect of correct response location on the difficulty level of multiple-choice questions. *Journal of Applied Psychology, 47,* 48-51.

Mickes, L., Flowe, H., & Wixted, J. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied, 18,* 361-376.

Rapaport, G. M., & Berg, I. A. (1955). Response sets in a multiple-choice test. *Educational and Psychological Measurement, 15*, 58-62.

Rotthoff, T., Fahron, U., Baehring, T., & Scherbaum, W. A. (2008). Die Qualität von CME-Fragen in der ärztlichen Fortbildung - eine empirische Studie [The quality of CME questions as a component part of continuing medical education - an empirical study]. *Zeitschrift für ärztliche Fortbildung und Qualität im Gesundheitswesen, 101,* 667-674.

Steblay, N. K., Dysart, J. E., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 25*, 459-473.

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law, 17,* 99-139.

Trepte, S., & Verbeet, M. (Eds.). (2010). *Allgemeinbildung in Deutschland. Erkenntnisse aus dem SPIEGEL-Studentenpisa-Test* [*General knowledge in Germany. Findings from the SPIEGEL PISA student test*]. Wiesbaden: VS Verlag.

Willing, S., Ostapczuk, M., & Musch, J. (2013). *Do sequentially presented answer options prevent the use of testwiseness cues in continuing medical education tests?* Manuscript submitted for publication.

Table 1

Validation criteria and their convergent correlations with an MC test and a DOMC test of political knowledge

	MC (N = 2,283)	DOMC (N = 2,207)	Comparison of correlation coefficients	
	r	*r*	*Z*	*p*
BOWIT political knowledge items	.57**	.56**	0.10	.92
Student PISA political knowledge items	.54**	.56**	-0.82	.41
Self-assessed intelligence	.14**	.17**	-1.17	.24
Self-assessed general knowledge	.29**	.30**	0.40	.69
Self-assessed political knowledge	.36**	.38**	-0.62	.54
Self-assessed interest in politics	.38**	.40**	-0.67	.50
Self-reported frequency of participation in political discussions	.33**	.36**	-1.10	.27
Self-reported number of political books read in the last three years	.17**	.17**	-0.21	.84
Self-reported intensity of reading political information	.39**	.39**	-0.06	.97
Number of years spent in school and university education	.14**	.12**	0.54	.59
Last school grade in politics	.20**	.24**	-1.12	.27

Note: **$p < .01$.

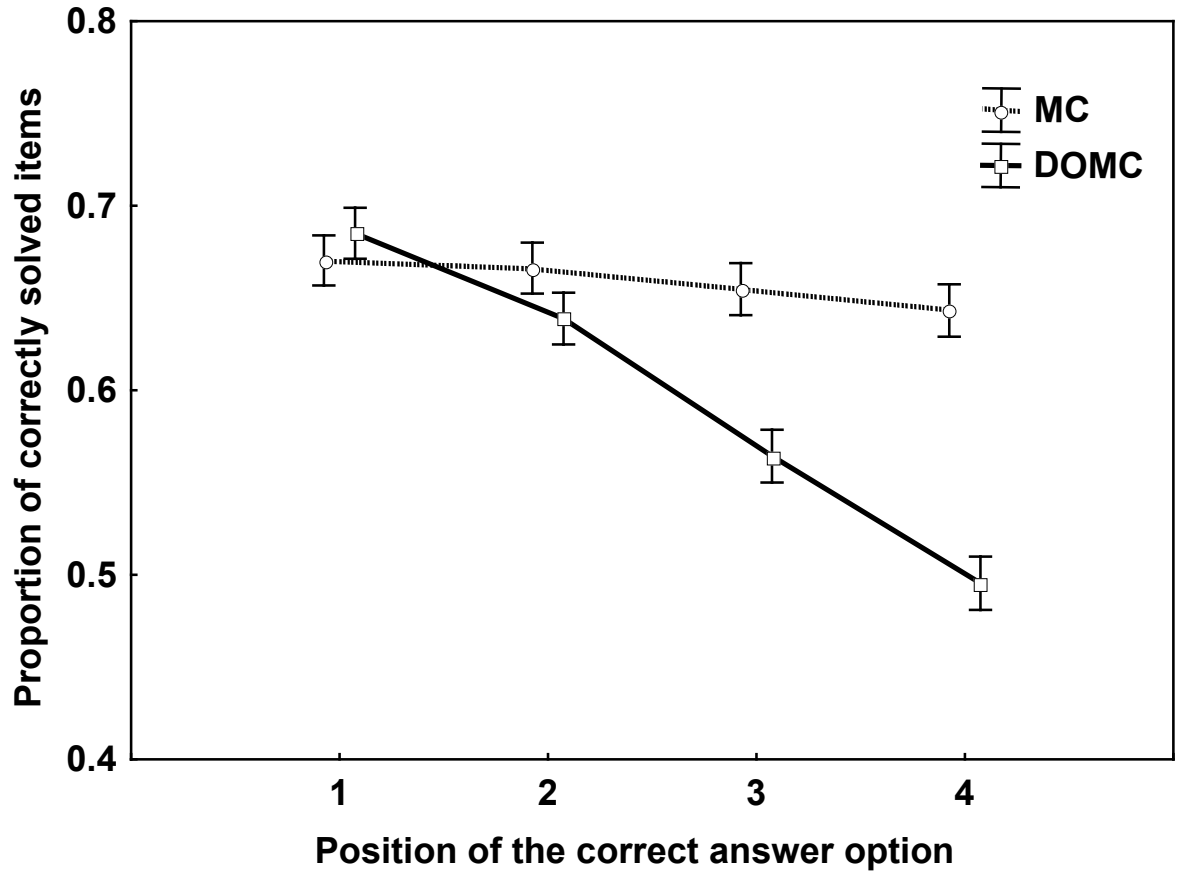

Figure 1. Proportions of correctly solved items and their 95% confidence intervals as a function of answer format and the position of the correct answer option.

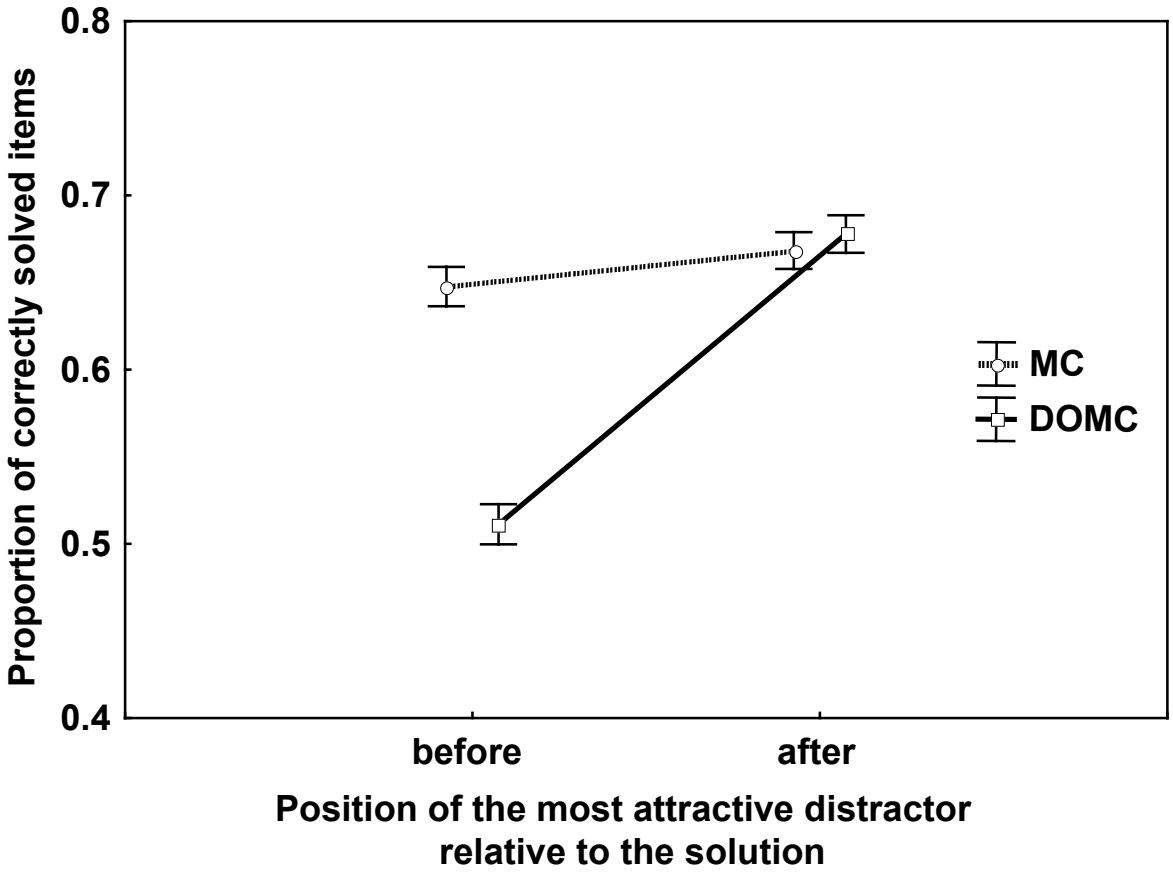

Figure 2. Proportions of correctly solved items and their 95% confidence intervals as a function of answer format and the position of the most attractive distractor relative to the solution.

Do Sequentially Presented Answer Options Prevent the Use of Testwiseness Cues on Continuing Medical Education Tests?

Sonja Willing*, Martin Ostapczuk, and Jochen Musch

Institute of Experimental Psychology

University of Duesseldorf

* Correspondence concerning this article should be addressed to:

Sonja Willing

University of Duesseldorf

Institute of Experimental Psychology

Building 23.03

Universitaetsstr. 1

D-40225 Düsseldorf

Germany

Email: sonja.willing@hhu.de

Fon: +49-211-81-11524

Fax: +49-211-81-11753

Abstract

The multiple-choice (MC) format is one of the most valid and hence, popular testing formats for the assessment of knowledge. However, testwiseness – that is, the ability to find subtle cues that point toward the solution – threatens the validity of MC tests. The majority of testwiseness cues require the simultaneous comparison of all available answer options. A computerized alternative testing format for MC tests has recently been proposed by Foster and Miller (2009); it is based on a sequential rather than a simultaneous presentation of the answer options. Discrete-option multiple-choice (DOMC) testing presumably allows for a better control of testwiseness. Test items that have been criticized for being susceptible to testwiseness strategies (Rotthoff et al., 2008) are used in continuing medical education (CME) programs aimed at developing and maintaining the knowledge of professionals in the medical field. We found that presenting answer options sequentially reduced the use of testwiseness cues on a CME test (Experiment 1). This result also held when participants were not informed of the presence of testwiseness cues (Experiment 2), and could be replicated with medical professionals (Experiment 3) in addition to working for participants from outside the medical field (Experiments 1 and 2). The sequential DOMC answer format was thus shown to allow for a better control of testwiseness than traditional MC testing.

Keywords:

multiple-choice test, discrete-option multiple-choice, sequential item presentation, testwiseness, continuing medical education test.

Introduction

The multiple-choice format (henceforth: MC) has been the format most widely used to assess knowledge for almost 100 years (Moreno et al., 2006). All MC items consist of an item stem and several answer options, including one correct option accompanied by several plausible distractors (Clegg and Cashin, 1986). Usually, all answer options are presented simultaneously to the examinees who are asked to choose the correct option.

There are many advantages to MC testing. MC items can be adapted to a wide range of different content and difficulty levels, allow standardized tests to be administered to large groups, and can be scored quickly and at low cost (Clegg and Cashin, 1986; Downing, 2006a; Taylor and Gardner, 1999). Most importantly, the psychometric properties of MC tests comply with high standards. Using MC, highly reliable and valid tests can be constructed (Downing, 2006a).

A disadvantage of MC testing, however, is that the person who creates the test is required to develop incorrect yet plausible options that can be difficult to devise (Downing, 2006a). MC tests are also easily compromised by student memorization, copying, and sharing (Foster and Miller, 2009). More importantly, the results of MC tests can be influenced by *testwiseness* (Foster and Miller, 2009; Martinez, 1999). Testwiseness is defined as the use of metacognitive answer strategies by test takers during examinations. Regardless of the knowledge domain to be assessed, test takers are often able to identify the correct answer option or to eliminate one or more of the distractors of an MC item solely on the basis of surface characteristics or by using content-independent reasoning processes (Millman et al., 1965, p. 707). Millman et al. (1965) presented the first comprehensive classification of testwiseness strategies. The majority of all testwiseness strategies require the simultaneous comparison of all available answer options. By comparing the various answer options, the test

takers try to identify a cue that facilitates the identification of the correct answer. Any use of such cues is likely to improve test scores. Representing construct-irrelevant variance, however, the use of cues can threaten the construct validity of MC tests (Haladyna and Downing, 2004). Another problem is that individual differences in testwiseness skills may selectively reward highly testwise examinees and penalize individuals who lack such skills (Edwards, 2003; Hammond et al., 1998; Millman et al., 1965; Sarnacki, 1979; Taylor and Gardner, 1999). In principle, items on carefully constructed and evaluated tests should not be solvable solely by the use of unintended cues if the guidelines of effective item writing are followed (Haladyna, 2004). However, Farley (1989) estimated that even an experienced item writer needs one hour to create a valid MC item and in practice, many MC items are created under time pressure and by authors who have little experience in test development (Downing, 2006b). It is therefore hardly surprising that a comprehensive analysis by Brozo et al. (1984) demonstrated that answer strategies are applicable and useful on many tests. Brozo et al. (1984) examined 1,220 MC items from 43 exams that had been administered at U.S. colleges. They found that about 44% of these items contained a cue that could be used by examinees with high testwiseness skills to exclude one or several distractors. No less than 70% of the faulty items could even be answered correctly without any knowledge of the content domain because an unintended cue allowed test takers to exclude all distractors.

The fact that there are many poorly constructed MC tests in classroom practice is disturbing (Brozo et al., 1984). It must be considered even more disturbing that some of them, such as continuing medical education (CME) tests, are used to make decisions that have serious consequences both for the examinees and their patients. Since 2004, medical professionals in Germany are committed by social law (German social security code V § 95d) to participate in CME trainings to maintain and develop their professional knowledge and skills after receiving their medical license. All medical professionals are thus required to

acquire at least 250 CME points every five years. In cases of non-compliance, medical professionals are faced with penalties ranging from a reduction in income to the withdrawal of their medical license. Private studies based on printed lectures followed by an achievement test containing 10 MC items constitute one prominent and approved method of acquiring CME points (German Medical Association, 2004). If test takers answer at least 7 out of the 10 items correctly, they receive 3 CME points, which is considered equivalent to three weeks of training and thereby complies with their medical training requirements.

Whereas the requirement to maintain the professional skills of medical professionals is generally embraced, there has been a controversial debate about whether MC items in CME programs are sufficiently challenging. Critics have argued that due to the testwiseness cues they often contain, many of these items can be solved even without training and without sufficient knowledge in the domain of the test (Rotthoff et al., 2008; Stagnaro-Green and Downing, 2006). Stagnaro-Green and Downing (2006) studied 40 CME items that were published in the New England Journal of Medicine in 2003. They found that 50% of all MC items contained the cue "Longest Alternative" (Brozo et al., 1984; Gibb, 1964; Millman et al., 1965; Sarnacki, 1979). For these items, an elaborate solution consisted of a greater number of words than all of the distractors. When comparing all answer options simultaneously, this provided the examinees with a helpful cue to the solution. Kühne-Eversmann et al. (2007) systematically evaluated the quality of MC items in three established German medical journals. The most frequent flaws were testwiseness cues and again, up to 30% of all items contained the cue "Longest Alternative". Rotthoff et al. (2008) investigated 200 items in twenty training units of four established German medical journals in 2006. They found that without exception, all training units included items with concealed cues that indicated the correct answer. Depending on the journal, the proportion of these items varied between 30% and 40%. One example of the frequent occurrence of testwiseness cues was the CME unit for

"The diagnosis and management of upper abdominal pain" from the German Medical Journal ("Deutsches Ärzteblatt") 06/2006: In this unit, 8 out of 10 items could be solved by simply applying testwiseness cues. It might therefore not be too surprising that 83.3% of the nearly 24,000 physicians taking this test answered all items in this unit correctly and thus complied with their medical training requirement.

To sum up, MC tests often contain cues to the solution and are therefore vulnerable to testwiseness strategies. To remedy the validity concerns associated with this problem, it is certainly worthwhile to make an effort to minimize the impact of such cues on MC test scores. In view of the many variants of the MC test format that have been proposed over the last 90 years (Downing, 2006b; Rodriguez, 2005), it is therefore surprising that up to now, no method allowing for an effective control of testwiseness on MC tests has been developed.

Foster and Miller (2009), however, recently proposed *discrete-option multiple-choice* (henceforth: DOMC) testing as a computer-based alternative answer format for MC tests. Like any MC item, a DOMC item consists of a stem, an answer option that is the correct solution, and several distractors (Foster and Miller, 2009). The difference from standard MC testing is that in DOMC testing, answer options are not presented simultaneously. Instead, they are presented one at a time in a random order. For each option, the test taker therefore has to make a decision about whether it is the correct solution or not. This presumably allows for a better control of cue utilization because the test taker can no longer compare all available options before providing his or her answer.

DOMC items are usually completed before all answer options have been presented because the presentation of each item ends when one of the following conditions is met: (a) the solution has been correctly identified as such (in this case, no more answer options need to be presented), (b) the solution has incorrectly been rejected, or (c) a distractor has incorrectly been accepted. In the last two cases, there is no need to present additional answer options

because the item has already been answered incorrectly. In other words, the presentation of a DOMC item ends as soon as it has been answered correctly or incorrectly. If this is the case, none of the remaining answer options is shown; instead, the next question is presented. This characteristic of DOMC testing potentially helps to reduce testing time in spite of the sequential presentation procedure, and Foster and Miller (2009) indeed observed that, compared to MC, DOMC reduced testing time by about 10%. Foster and Miller (2009) also identified the more limited exposure of the various answer options as another advantage of the new testing format. If an answer option is never presented to a participant, he or she cannot recall it or give it away to future participants. This makes it easier to reuse DOMC items on future exams and enhances test security.

In several studies, it was observed that DOMC items were usually more difficult than MC items (Foster and Miller, 2009; Kingston et al., 2012). Foster and Miller (2009) argued that the increased item difficulty was probably the result of the reduced impact of testwiseness, although they did not ensure that the mathematical problems they presented actually did contain testwiseness cues. It is therefore possible that in their study, item difficulty increased because in DOMC testing, test takers have to base their decisions on a smaller amount of information than in MC testing where all options are presented simultaneously, allowing test takers to compare them before having to decide on the solution. It is thus presently unknown whether DOMC testing is indeed capable of reducing the effect of testwiseness. We therefore decided to investigate this issue further and wanted to test whether a sequential presentation of the answer options was indeed able to prevent the use of testwiseness cues. As answer options are presented one at a time in DOMC testing, and because the majority of testwiseness strategies – as classified by Millman et al. (1965) – require the simultaneous comparison of all answer options, we expected that the use of DOMC would prevent the comparison of all options before participants answered an item and

would reduce the probability that a cue to the solution would be identified. We therefore expected that DOMC testing would be less affected by the impact of testwiseness strategies.

To investigate whether DOMC testing could prevent the use of testwiseness cues more effectively than the usual MC test format, we used a published test intended and previously employed for high-stakes purposes, namely, the test accompanying the CME unit for "The diagnosis and management of upper abdominal pain" from the German Medical Journal (“Deutsches Ärzteblatt”, 06/2006). In a series of three experiments, we randomly assigned participants taking this test to either the MC or the DOMC condition. In each condition, every test taker was provided with the 10 items accompanying this unit, of which 8 could be solved solely on the basis of some cues to the solution, whereas the remaining 2 items did not contain such cues (Rotthoff et al., 2008, p. 671).

We expected a higher proportion of correctly solved items in the MC condition than in the DOMC condition because for such poorly constructed items, we expected that a sequential presentation of the answer options would allow for a better control of testwiseness than a simultaneous presentation of the answer options. Another reason to expect this main effect was the observation of both Foster and Miller (2009) and Kingston et al. (2012) that MC items are typically easier to answer than sequentially presented DOMC items. However, we expected that this effect would be limited to the less carefully constructed items for which the solution was more likely to be identified by making use of the cues they contained. Thus, we expected an interaction between the answer format (MC, DOMC) and the availability of testwiseness cues (present in 8 items, not present in 2 items). We compared the proportion of correct solutions for the two carefully constructed CME items that did not contain any cues with the proportion of correct solutions for the 8 CME items that could be solved by identifying the cue to their solution. For items with a testwiseness cue, we expected a higher proportion of correctly solved items in the MC condition than in the DOMC condition

because we expected only participants in the MC condition to be able to apply these cues. By contrast, we did not expect participants in the DOMC condition to be able to identify cues to the solution. For items without a testwiseness cue, we expected no difference between the proportion of correctly solved items in the MC condition and in the DOMC condition.

A secondary aim of the three studies was to investigate the efficiency of the proposed new answer format by determining the associated decrease in testing time. Even though less than half of the answer options had to be presented per item on average, Foster and Miller (2009) observed a reduction in testing time of only about 10%. As DOMC testing requires test takers to make more decisions per item than MC testing, a significant reduction in the total testing time of somewhat less than 50% was expected. However, the time needed for the instructions was expected to be much higher in the DOMC condition due to the necessity to explain this new procedure to all participants.

Experiment 1

The aim of Experiment 1 was to examine whether DOMC testing would allow for a better control of testwiseness than MC testing. Furthermore, the efficiency of DOMC testing was explored by determining the decrease in testing time associated with using DOMC rather than MC testing.

Method

Participants. Forty-eight (27 female) native German speakers who had previously participated in online studies conducted by our department were recruited via an online panel. Data from an additional 11 participants who did not complete the questionnaire had to be discarded. Participants' ages ranged from 17 to 68 years (M = 41.46, SD = 16.52). Four participants indicated a Ph.D. as their highest academic degree. Another 18 participants held a

master's degree, 3 held a bachelor's degree, 15 had completed their final secondary-school examination (the German “Abitur”), 7 held a junior high-school diploma, and one participant had not graduated from high school. The number of years spent in school, high school, and university education ranged from 10 to 24 years (M = 16.00, SD = 3.53). For the 40 participants who had successfully completed their secondary-school examinations, the final school exam grade (the German “Abiturnote”) averaged 2.30 (SD = 0.62) on a scale ranging from 1 (best) to 4 (worst).

Participants were randomly assigned to either the DOMC condition (N = 25) or the MC condition (N = 23). Participants in the two conditions did not differ with regard to the length of their education ($t(46) = 1.07$, $p = .29$, $d = 0.31$) or their final school exam grade ($t(38) = 0.40$, $p = .69$, $d = 0.13$).

Design. The experiment used a 2 x 2 mixed design with the independent between-subjects variable *answer format* (MC vs. DOMC) and the within-subjects variable *availability of cues* (items with vs. items without cues to their solution). As dependent variables, we recorded the proportion of correctly solved items and the time needed to read the instructions and to complete all items.

Material. We used items from the CME unit "The diagnosis and management of upper abdominal pain" published in the German Medical Journal (Niedergethmann and Post, 2006). This unit has been certified by the North Rhine Academy for medical training and education. It includes an instructional unit followed by 10 MC items. Each item consists of a stem and five answer options, one of which is the solution. Only 2 of the 10 items do not contain any testwiseness cues. By contrast, 8 of the 10 items contain cues that allow the test taker to determine the correct solution or at least to increase the probability of a correct guess (Rotthoff et al., 2008). Specifically, these items comprise the following three testwiseness cues:

1. *Longest Alternative* (Brozo et al., 1984, p. 5; Gibb, 1964, p. 20; Millman et al., 1965, p. 712; Sarnacki, 1979, p. 255). The correct answer option is often formulated in more detail and is therefore more wordy than the distractors. In a study by Brozo et al. (1984), 54 (4.4%) of 1,220 items contained this cue. On the present CME test, we identified 6 items that contained the cue "Longest Alternative" because the solution was at least 1/3 line of print longer than the remaining options.

2. *Categorical Exclusives* (Gibb, 1964, p. 20). By including overgeneralizations based on quantifiers such as “never,” “always,” or “absolutely,” answer options are often overqualified and thus identified as incorrect. For such items, the solution is the only answer option that does not include any overgeneralizing qualifiers. In the present CME unit, 2 of 10 items contained the cue of categorical exclusives.

3. *Grammatical Error* (Brozo et al., 1984, p. 6; Gibb, 1964, p. 20). Often distractors are formulated less carefully than the solution and contain grammatical errors. If only one answer option is grammatically aligned with the item stem, it can easily be identified as correct. On the present CME test, one item contained this cue in addition to the cue “Longest Alternative”.

Procedure. The questionnaire was delivered online using the software Unipark (Version 8; Globalpark AG, 2011). At the beginning of the questionnaire, participants were told that MC items are frequently criticized for the presence of cues to their solution and that finding such cues facilitates the solution of an item even if a test taker is not familiar with the topic at hand. The above three testwiseness cues were explained, and an example item was given for each. Next, the participants were randomly assigned to either the DOMC or the MC condition. Participants were first introduced to the answer format that was used for the test. As the DOMC format was expected to be less familiar to the test takers, it had to be described in more detail. The DOMC procedure was explained using a sample item. Additionally,

participants were told about all stopping criteria employed during the sequential presentation of the answer options in DOMC testing. Subsequently, participants completed all 10 CME items – 8 items with a testwiseness cue and 2 items without a testwiseness cue. All items were presented in a random order in both the MC and DOMC conditions. Within all items, all answer options were also presented in a random order. In the MC condition, one item was presented per page along with all possible answer options. In the DOMC condition, answer options were presented sequentially. After completing the 10 items, participants were asked to indicate their age, sex, and education. At the end of the test, participants received feedback on their individual performance and were told whether they would have passed the medical examination. Participants were thanked and debriefed and were provided with some further information on continuing medical education in Germany.

Results

For each participant, all responses were recorded, and the proportion of correct solutions was computed separately for the 8 items with a testwiseness cue and the 2 items without such a cue. Additionally, the time needed to read the instructions and the time needed to complete all items were recorded. In all statistical tests, an alpha level of .05 was used. Effect sizes for the difference between two means were calculated using Cohen's *d*. According to Cohen (1988), an effect may be considered small when $d \geq 0.20$, medium when $d \geq 0.50$, and large when $d \geq 0.80$. ANOVA effect sizes were computed using eta-squared (η^2), the proportion of variance explained by each factor or interaction. According to a taxonomy proposed by Cohen (1988), $\eta^2 \geq 0.14$ implies a large effect, $\eta^2 \geq 0.06$ a moderate effect, and $\eta^2 \geq 0.01$ a small effect.

Control for testwiseness cues. To compare the proportion of correct solutions across conditions, a mixed between-within-subjects ANOVA was computed with *answer format*

(MC vs. DOMC) as the between-subjects factor and *availability of cues* (items with vs. without cues) as the within-subjects factor. As can be seen in figure 1, this resulted in a significant main effect of the availability of cues, indicating that the correct answer option could be identified more easily in items that contained a cue (M = .47, SD = .25) than in items without a cue (M = .23, SD = .31), $F(1, 46) = 21.53$, $p < .001$, $\eta^2 = 0.32$. Possibly owing to statistical power that was not sufficient for detecting this effect, the main effect of answer format was not statistically significant, $F(1, 46) = 1.34$, $p = .25$, $\eta^2 = 0.03$, suggesting no significant difference in the proportions of correctly solved items between the DOMC condition (M = .31, SD = .19) and the MC condition (M = .38, SD = .23). As predicted, however, the two-way interaction between these two variables was significant, indicating that participants in the MC condition were more successful at using the cues - if present - than participants in the DOMC condition, $F(1, 46) = 6.73$, $p = .01$, $\eta^2 = 0.13$. Additional t-tests were computed to explore the nature of this interaction. When cues were available, participants in the MC condition (M = .57, SD = .24) solved a higher proportion of items than participants in the DOMC condition (M = .36, SD = .21), $t(46) = 3.04$, $p < .01$, $d = 0.89$. By contrast, there was no significant difference in the proportion of correctly solved items between the MC condition (M = .20, SD = .32) and the DOMC condition (M = .26, SD = .30) for items without a cue to their solution, $t(46) = 0.68$, $p = .50$, $d = 0.19$.

The results show that participants in the MC condition were far more successful at making use of the cues contained in poorly constructed CME items than participants in the DOMC condition. If no cues were available, the proportion of correctly solved items did not differ from the chance baseline level of .20 both in the MC condition ($t(24) = 0$, $p = 1.0$, $d = 0$) and in the DOMC condition ($t(22) = 0.98$, $p = .34$, $d = 0.28$), indicating that participants who were not familiar with the content of the CME exam and who were not able to use cues to the solution were forced to rely on guessing. Thus, taken together, the sequential

presentation of answer options in DOMC testing prevented the use of cues to the solution more successfully than the simultaneous presentation of answer options in traditional MC testing.

--- Insert figure 1 here ---

Testing times. An independent-samples t-test was computed to compare the testing times of the DOMC and MC conditions. For DOMC testing, the time needed to answer the 10 CME items was reduced by 50.26% (M = 217.57 seconds, SD = 88.47) as compared with MC testing (M = 437.44 seconds, SD = 171.74), $t(46) = 5.50$, $p < .001$, $d = 1.61$. As the position of the solution was varied randomly, the stopping criteria considerably reduced the average number of answer options that had to be presented to the test takers in the DOMC condition. As a result, a perfect test taker could be expected to be presented with an average of 3 out of the 5 possible answer options in the DOMC condition. For the test takers in the present study, owing to their frequent erroneous acceptance of a distractor early in the course of the presentation of an item, this presentation ended even earlier: on average, after 1.68 (SD = 0.44) out of the 5 possible answer options.

Due to the need to read a more detailed explanation of the testing format, participants in the DOMC condition needed more time to read the instructions (M = 101.13 seconds, SD = 73.16) than participants in the MC condition (M = 16.24 seconds, SD = 6.11), $t(46) = 5.79$, $p < .001$, $d = 1.64$. However, when the time needed to read the instructions was added to the time needed to answer the items to compute the total testing time, the total time was reduced in the DOMC condition (M = 318.70 seconds, SD = 115.67) as compared with the MC condition (M = 453.68 seconds, SD = 172.96), even despite the longer amount of time spent reading the instructions, $t(46) = 3.15$, $p < .01$, $d = 0.92$. In summary, the results of Experiment

1 showed that under DOMC testing conditions, participants took about 29.75% less time to read the instructions and to complete all items, suggesting that having to deal with a smaller number of answer options in DOMC testing reduces the total testing time even in spite of having to read longer instructions and having to make more decisions.

Discussion

The sequential presentation of answer options in the DOMC condition was more successful at preventing the use of cues to the solution than the simultaneous presentation of answer options that is customary in MC testing. However, against our expectations and contrary to the results of Foster and Miller (2009), a main effect of the answer format was not found. Although DOMC items were also more difficult than MC items in our study, the difference between the two answer formats was not statistically significant. However, this lack of a significant main effect of test format may have been the result of the lack of power associated with the small sample size. More importantly, with regard to the usefulness of a sequential presentation of answer options, we observed a substantial reduction in testing time when using DOMC rather than MC testing. This effect was the result of the smaller number of answer options that had to be presented in the DOMC condition. In DOMC testing, the presentation of an item stopped whenever a distractor was mistakenly accepted as the solution. Moreover, the item presentation also stopped after the presentation of the solution because the solution could be either correctly accepted or wrongly rejected; in both of these cases, it was unnecessary to present additional answer options. Testing time was thus reduced considerably.

However, an obvious and legitimate criticism of Experiment 1 is that the cues to the solution were explained explicitly to all participants prior to the presentation of the CME items. It may be argued that participants were able to make use of the cues to the solution

only because the cues were explicitly revealed to them before they began working on the test. Therefore, to ensure the reliability and replicability of our finding, we conducted a second experiment in which participants were not informed about the presence of testwiseness cues.

Experiment 2

In Experiment 2, we expected that participants in the MC condition would be more successful than participants in the DOMC condition at making use of cues to the solution, even when both groups were not informed about the presence of such cues. In order to increase the statistical power to detect a main effect of answer mode, we used a larger sample size than was used in Experiment 1.

Method

Participants. Eighty-six (45 female) native German speakers who had previously participated in online studies in our department but who had not participated in Experiment 1 were recruited via an online panel. Their ages ranged from 16 to 80 years (M = 34.12, SD = 14.55). When asked to indicate their highest academic degree, 4 participants indicated a Ph.D., 22 indicated a master's degree, 16 indicated a bachelor's degree, 27 indicated that they had completed the final secondary-school examinations (the German “Abitur”), 15 indicated that they had acquired a junior high-school diploma, and two participants had not graduated from high school. The number of years spent in school, high school, and university education ranged from 9 to 23 years (M = 16.27, SD = 3.71). For the 69 participants who had successfully completed their secondary-school examinations, the final school exam grade (the German “Abiturnote”) averaged 2.17 (SD = 0.59). Participants were randomly assigned to either the DOMC condition (N = 46) or the MC condition (N = 40). Participants did not differ

in length of education ($t(84) = 0.22$, $p = .23$, $d = 0.05$) or in their average final school exam grade ($t(67) = 1.22$, $p = .13$, $d = 0.30$) across conditions. Data from an additional 16 participants who did not complete the questionnaire had to be discarded.

Design. The design was identical to Experiment 1.

Material. Examinees were provided with the same CME test that had already been used in Experiment 1.

Procedure. The procedure was the same as in Experiment 1 with the exception that no participants were informed about the nature of the cues that could be used to detect the solutions of some of the items.

Results

Control for testwiseness cues. As in Experiment 1, we computed a 2 x 2 (answer format [MC vs. DOMC] x availability of cues [cue vs. no cue] ANOVA to compare the proportion of correctly solved items across the experimental conditions. Participants in the MC condition solved a higher proportion of items ($M = .39$, $SD = .22$) than participants in the DOMC condition ($M = .30$, $SD = .17$). Probably owing to the larger sample size as compared with Experiment 1, this difference was statistically significant in Experiment 2, $F(1, 84) = 4.78$, $p = .03$, $\eta^2 = 0.05$. The proportion of correctly solved items decreased as a function of the availability of cues. Items with a testwiseness cue ($M = .46$, $SD = .24$) were correctly solved more often than items without such a cue ($M = .22$, $SD = .29$). This effect of the availability of cues was significant, $F(1, 84) = 44.87$, $p < .001$, $\eta^2 = 0.35$. Most importantly, there was a significant interaction between answer format and the availability of cues, $F(1, 84) = 13.18$, $p < .001$, $\eta^2 = 0.14$ (see figure 2). To investigate this interaction more closely, additional t-tests were computed. We found a significantly higher proportion of correctly solved items in the MC condition ($M = .58$, $SD = .24$) than in the DOMC condition ($M = .35$, $SD = .20$) for

items containing a cue to the solution, $t(84) = 4.81$, $p < .001$, $d = 1.04$. By contrast, we found no significant difference in the proportion of correctly solved items between the MC condition ($M = .20$, $SD = .30$) and the DOMC condition ($M = .24$, $SD = .29$) for items without such cues, $t(84) = 0.62$, $p = .54$, $d = 0.14$. If CME items did not include cues to the solution, the proportion of correctly solved items did not differ from the chance level of .20 both in the MC condition ($t(39) = 0$, $p = 1.00$, $d = 0$) and in the DOMC condition ($t(45) = 0.91$, $p = .37$, $d = 0.20$), indicating that when participants lacked knowledge in the domain of the test, they were forced to guess when no cues were available to help them identify the solution.

--- Insert figure 2 here ---

Testing times. Presenting answers sequentially in the DOMC condition considerably reduced the time needed to answer the 10 CME items ($M = 215.64$ seconds, $SD = 73.46$) as compared with the MC condition ($M = 342.95$ seconds, $SD = 140.49$), $t(84) = 5.36$, $p < .001$, $d = 1.14$. Due to the item stopping criteria, the average number of answer options that had to be presented to the test takers in the DOMC condition was only 1.88 ($SD = 0.49$). As a result, DOMC reduced testing time by 37.12%.

Because of its novelty, the DOMC format had to be explained to the participants in more detail than the MC format. Therefore, as in Experiment 1, participants in the MC condition ($M = 18.55$ seconds, $SD = 6.81$) needed significantly less time to read the instructions than participants in the DOMC condition ($M = 62.14$ seconds, $SD = 22.92$), $t(84) = 11.58$, $p < .001$, $d = 2.58$. To compute the total testing time, we added the time needed to read the instructions to the time to complete all items. Even though reading the instructions required more time in the DOMC condition, participants in the DOMC condition ($M = 277.78$ seconds, $SD = 87.32$) completed the test significantly more quickly than participants in the

MC condition (M = 361.50 seconds, SD = 144.13), $t(84) = 3.31$, $p < .02$, $d = 0.70$. As in Experiment 1, these results indicated that participants took about 23.16% less time to read the instructions and complete the 10 CME items, suggesting that DOMC testing leads to a considerable enhancement of test economy.

Discussion

Experiment 2 replicated the findings of Experiment 1and supported the notion that DOMC testing allows for a better control of testwiseness than MC testing. This effect was found even when we did not inform participants about the cues contained in some of the items. Experiment 2 also replicated the finding that DOMC allows for a reduction in the total testing time. It is important to note, however, that participants in Experiments 1 and 2 did not have any medical background, as indicated by their chance performance on items that did not contain cues to their solution. We therefore wanted to replicate our findings using a sample of medical students and trained physicians in Experiment 3.

Experiment 3

For obvious reasons, we were particularly interested in whether DOMC would allow for the control of testwiseness and would decrease testing time not only in a convenience sample of participants with heterogeneous educational backgrounds, but also in the target group of individuals who were in continuing medical education. Therefore, medical professionals were used as the sample in Experiment 3 to examine whether the findings of the two previous experiments could be replicated.

Method

Participants. The sample consisted of 106 (66 female) medical students and medical doctors between the ages of 19 and 56 years (M = 25.35, SD = 5.91). The participants were all German native speakers and were recruited via announcements in a learning portal for medical professionals (MEDI-LEARN) as well as in the bulletin board of a student council for medical students. 80 participants were medical students, 9 participants were final-year medical students, 14 participants were medical residents, and 3 participants were trained medical specialists. Their number of years of education ranged between 13 and 28 years (M = 16.88, SD = 2.71), and the grade on their final school examination averaged 1.79 (SD = 0.59). Participants were randomly assigned to either the DOMC condition (N = 53) or the MC condition (N = 53). Participants did not differ in length of education ($t(104) = 0.97$, $p = .34$, $d = 0.19$) or in average final school exam grade ($t(104) = 0.14$, $p = .89$, $d = 0.02$) between conditions. Data from an additional 34 participants who did not complete the questionnaire had to be discarded.

Design. The design was the same as in Experiments 1 and 2.

Material. The material was identical to the material used in Experiments 1 and 2.

Procedure. The procedure was the same as in Experiment 2.

Results

Control for testwiseness cues. We again computed a 2 x 2 (answer format [MC vs. DOMC] x availability of cues [cues vs. no cues] ANOVA to compare the proportion of correctly solved items across experimental conditions. Participants in the MC condition solved a higher proportion of items (M = .56, SD = .22) than participants in the DOMC condition (M = .42, SD = .20). This difference was statistically significant, $F(1, 104) = 12.61$, $p = .001$, $\eta^2 = 0.11$. Moreover, medical examinees correctly solved a higher proportion of

items that contained cues to their solution (M = .59, SD = .22) than items that did not contain such cues (M = .39, SD = .37). This effect of cue availability was also significant, $F(1, 104) = 25.32$, $p < .001$, $\eta^2 = 0.20$. Most importantly, even in this sample of medical professionals, a significant interaction between answer format and the availability of cues was observed, $F(1, 104) = 5.68$, $p = .02$, $\eta^2 = 0.05$ (see figure 3).

To investigate this interaction more closely, additional t-tests were computed. As in Experiments 1 and 2, we found a significantly higher proportion of correctly solved items in the MC condition (M = .71, SD = .20) than in the DOMC condition (M = .47, SD = .18) for items containing a cue to their solution, $t(104) = 6.41$, $p < .001$, $d = 1.31$. By contrast, we found no significant difference in the proportion of correctly solved items between the MC condition (M = .42, SD = .35) and the DOMC condition (M = .37, SD = .38) for items that did not contain such cues, $t(104) = 0.66$, $p = .51$, $d = 0.14$. In contrast to the two previous experiments, for the present sample of medical students and doctors, the proportion of items that were solved correctly even without a cue to their solution differed significantly from random guessing, both in the MC condition ($t(52) = 4.48$, $p < .001$, $d = 0.89$) and in the DOMC condition ($t(52) = 3.20$, $p = .002$, $d = 0.63$). This result indicates that participants were able to solve CME items better than chance without using cues only on the basis of some medical knowledge, suggesting that the CME test provides a valid assessment of medical knowledge.

--- Insert figure 3 here ---

Testing times. Due to the sequential presentation of answer options, testing time was significantly reduced in the DOMC condition (M = 265.83 seconds, SD = 113.18) as compared to the MC condition (M = 433.17 seconds, SD = 179.73), $t(104) = 5.74$, $p < .001$, d

= 1.11. DOMC reduced the testing time for the 10 CME items by 38.63%. Owing to the item stopping criteria, the presentation of the items stopped, on average, after 2.09 out of the five possible answer options (SD = 0.40). Again, participants in the MC condition (M = 18.52 seconds, SD = 7.87) needed significantly less time to read the instructions than participants in the DOMC condition (M = 74.11 seconds, SD = 43.77), $t(104) = 9.10$, $p < .001$, $d = 1.77$. Computing the total testing time, medical examinees in the DOMC condition (M = 339.94 seconds, SD = 141.91) completed the test significantly more quickly than participants in the MC condition, (M = 451.69 seconds, SD = 183. 73), $t(104) = 3.50$, $p = .001$, $d = 0.67$. Overall, using DOMC reduced total testing time by 24.74%.

Discussion

The results of the third experiment confirmed the hypothesis that the sequential presentation of answer options would allow for the control of testwiseness for poorly constructed CME items. This was shown not only for participants with a heterogeneous educational background, but also for medical professionals who achieved higher test scores than participants in the two previous studies; these medical professionals were less dependent on relying on such cues in the first place due to their medical knowledge. Moreover, we found that the sequential presentation of answer options reduced testing times in all three experiments even though none of our participants was acquainted with the new format.

General discussion

DOMC, an alternative testing format for MC items recently proposed by Foster and Miller (2009) is based on the sequential rather than simultaneous presentation of answer options. We evaluated whether this new procedure would offer better control of testwiseness

than traditional MC testing. To this end, we compared the proportions of correctly solved items on a published German medical education test for items without a cue to their solution and for items with such a cue using both a MC format and a DOMC format. On this test, there were items for which the correct answer option was more wordy than the distractors, contained a lower degree of generalization, or was grammatically better aligned with the stem. In a series of three studies, the sequential presentation of answer options (i) reduced the use of testwiseness cues and (ii) reduced total testing time.

Foster and Miller (2009) and Kingston et al. (2012) found that sequentially presented DOMC items seemed to be more difficult than MC items. In their studies, however, they could only surmise that this may have been due to the better control of testwiseness that is afforded by the DOMC answer format because they used only mathematical problems for their MC items and did not ensure that these items contained testwiseness cues that could be used to arrive at the solution. The present studies show that the DOMC answer format is indeed capable of preventing the use of testwiseness cues better than the traditional MC format. Although the availability of testwiseness cues led to an increase in the proportion of correctly solved items in both conditions, this increase was larger in the MC condition.

However, it is true that the control of testwiseness afforded by DOMC was less than perfect, considering that participants profited from the availability of item cues even in the DOMC condition. The most probable explanation for this is that some item cues can be used even under sequential presentation; for example, when all answer options are presented before one of the stopping criteria is met. Nevertheless, the DOMC format allowed for an improved control of testwiseness that was greatly superior to that under MC testing conditions. The findings of Experiment 1 support the notion that DOMC is capable of controlling testwiseness better than the traditional MC format, although participants in both conditions profited from the availability of cues. Participants in the MC condition, however, were more successful at

making use of the cues in poorly constructed CME items than participants in the DOMC condition if testwiseness cues were explained beforehand. The results of Experiment 2 demonstrated that this finding could be replicated even when no cues were explained prior to the administration of the test. The results of Experiment 3 were in full agreement with the first two experiments and demonstrated that sequentially presented items were able to reduce the use of cues among medical professionals on a CME test better than simultaneously presented items. This finding suggests that the benefits of DOMC testing will probably generalize to professional participants in high-stakes testing situations.

In continuing medical education, the review of professional knowledge by means of a test has important consequences for the test taker. On the one hand, passing the tests leads to important benefits because certifications are awarded. On the other hand, medical professionals are faced with different penalties, ranging from income reduction to withdrawal of their medical license if they fail to take or to pass the tests. For obvious reasons, it is essential that medical professionals educate themselves regularly. Hence, CME items should measure true knowledge and should not simply assess the participants' testwiseness skills. It therefore seems highly desirable to improve the quality of continuing medical education in a way that is as simple as substituting the DOMC format for the MC format.

To summarize, our studies show that DOMC items are superior to MC items in controlling testwiseness. This was valid for a test that was most likely created under time pressure and by authors who had little experience in the development of tests. According to Brozo et al. (1984), it is especially this kind of test that often contains cues to the solution.

Assessing the number of answer options that had to be presented for each DOMC item allowed us to understand why this format is more efficient at controlling for testwiseness than MC testing. In three studies, we found that the presentation of DOMC items was stopped, on average, after the presentation of 1.7 to 2.1 out of the five possible answer options. This great

reduction in the number of answer options that were available for comparison made it difficult to take full advantage of the available cues in the DOMC condition.

Owing to the decreased number of answer options that have to be presented per item, a considerable reduction in testing time seems to be an additional advantage of DOMC testing. Unlike Foster and Miller (2009), who observed a decreased testing time of only about 10%, we found a reduction in testing time ranging from 23% to 30% in our three studies. Thus, although on average, less than half of the answer options had to be presented per item, testing time was reduced by less than 50%. This was probably due to the need to make decisions after the presentation of each answer option, which prevented the savings in time from amounting to the same magnitude as the savings in the number of items. However, the savings in testing time may well increase up to 50% once test takers become more familiar with the new testing format. We observed that test takers needed consistently more time to read the extended instructions in the DOMC condition; however, less than two minutes were required to explain the new answer format the first time it was used, and this needs to be done only once. Hence, instructions can probably be shortened considerably once participants become better acquainted with the new DOMC format. DOMC testing may also considerably reduce the danger of unwanted copying and sharing of test items because many answer options do not have to be presented at all (Foster and Miller, 2009).

The social acceptance (Kersting, 2008) of DOMC testing remains to be investigated more closely, however. As a sequential presentation of answers options provides less information to the participants and decisions must be made after the presentation of each answer option, it seems quite possible that participants might prefer MC testing to the new format.

It is fair to mention that there are also some disadvantages to the new DOMC format. Test administration software is needed to present tests in the DOMC format. However, these

costs are one-time expenses, solely occurring during the initial stage of implementation (Exam Innovations, 2010).

In conclusion, the present studies reveal three straightforward benefits of the new DOMC answer format. First, our studies show that the DOMC answer format allows for a better control of testwiseness than MC testing. Second, DOMC testing reduces the number of answer options that are presented per item and that are available for comparison when trying to arrive at the correct solution, thereby enhancing test security. Last, the use of DOMC reduces testing time, in spite of the additional time that is needed for participants to read the extended instructions to understand the new format. For all of these reasons, it certainly seems worthwhile to further evaluate the usefulness of this new answer format.

References

Brozo, W. G., Schmelzer, R. V., & Spires, H. A. (1984). A study of testwiseness clues in college and university teacher-made tests with implications for academic assistance centers (*Technical Report 84-01*). Georgia State University: College Reading and Learning Assistance. Retrieved August 21, 2013, from ERIC database (ED240928).

Clegg, V. L., & Cashin, W. E. (1986). *Improving multiple-choice tests.* IDEA Paper No. 16. Manhattan: Kansas State University, Center for Faculty Evaluation and Development. Retrieved August 21, 2013 from http://www.theideacenter.org/sites/default/files/Idea_Paper_16.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Downing, S. M. (2006a). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287-301). Mahwah, NJ: Lawrence Erlbaum.

Downing, S. M. (2006b). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum.

Edwards, B. D. (2003). *An examination of factors contributing to a reduction in race-based subgroup differences on a constructed response paper-and-pencil test of achievement.* Unpublished doctoral dissertation, Texas A&M University.

Exam Innovations (2010). *Return on investment for discrete option multiple choice*. Exam Innovations, Inc. Retrieved August 21, 2013 from http://dl.dropbox.com/u/528723/ROI%20for%20DOMC.pdf

Farley, J. K. (1989). The multiple choice test: Writing the questions. *Nurse Educator, 14,* 10-12, 39.

Foster, D., & Miller, H. L. (2009). A new format for multiple-choice testing: Discrete-option multiple-choice. Results from early studies. *Psychology Science Quarterly, 51*, 355-369.

German Medical Association (2004). *Regulations for Continuing Education and Continuing Education Certificate.* Retrieved August 21, 2013 from http://www.bundesaerztekammer.de/downloads/ADFBSatzungEn.pdf

German social security code (SGB). *§ 95d SGB V Pflicht zur fachlichen Fortbildung* [Obligation to professional training]. Retrieved August 21, 2013 from http://www.sozialgesetzbuch-sgb.de/sgbv/95d.html

Globalpark AG (2011). *Enterprise Feedback Suite. EFS Survey.* Retrieved August 21, 2013 from http://www.unipark.info

Gibb, B. G. (1964). *Testwiseness as secondary cue response.* Doctoral dissertation, Stanford University, Ann Arbor, Michigan: University Microfilms, 1964. No. 64-7643.

Haladyna, T. M. (2004). *Developing and validating multiple choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*, 17-27.

Hammond, E. J., McIndoe, A. K., Sansome, A. J., & Spargo, P. M. (1998). Multiple-choice examinations: Adopting an evidence-based approach to exam technique. *Anaesthesia, 53,* 1105-1108.

Kersting, M. (2008). Zur Akzeptanz von Intelligenz- und Leistungstests [On the acceptance of intelligence and achievement tests]. *Report Psychologie, 33,* 420-433.

Kingston, N. M., Tiemann, G. C., Miller, H. L., & Foster, D. (2012). An analysis of the discrete-option multiple-choice item type. *Psychological Test and Assessment Modeling, 54,* 3-19.

Kühne-Eversmann, L., Nussbaum, C., Reincke, M., & Fischer, M. R. (2007). CME-Fortbildungsangebote in medizinischen Fachzeitschriften: Strukturqualität der MC-Fragen als Erfolgskontrollen [CME activities of medical journals: Quality of multiple-choice questions as evaluation tool]. *Medizinische Klinik, 102,* 993-1001.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34, 207-218.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of testwiseness. *Educational and Psychological Measurement*, *25*, 707-726.

Moreno, R., Martinez, R. J., & Muniz, J. (2006). New guidelines for developing multiple-choice items. *Methodology, 2,* 65-72.

Niedergethmann, M., & Post, S. (2006). Differentialdiagnose des Oberbauchschmerzes [The diagnosis and management of upper abdominal pain]. *Deutsches Ärzteblatt, 13,* A862-A871.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A metaanalysis of 80 years of research. *Educational Measurement: Issues and Practice, 24,* 3-13.

Rotthoff, T., Fahron, U., Baehring, T., & Scherbaum, W. A. (2008). Die Qualität von CME-Fragen in der ärztlichen Fortbildung - eine empirische Studie [The quality of CME questions as a component part of continuing medical education - an empirical study]. *Zeitschrift für ärztliche Fortbildung und Qualität im Gesundheitswesen, 101,* 667-674.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. *Review of Educational Research, 49,* 252-279.

Stagnaro-Green, A. S., & Downing, S. M. (2006). Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. *Medical Teacher, 28,* 566-568.

Taylor, C., & Gardner, P. L. (1999). An alternative method of answering and scoring multiple choice tests. *Research in Science Education, 29,* 353-363.

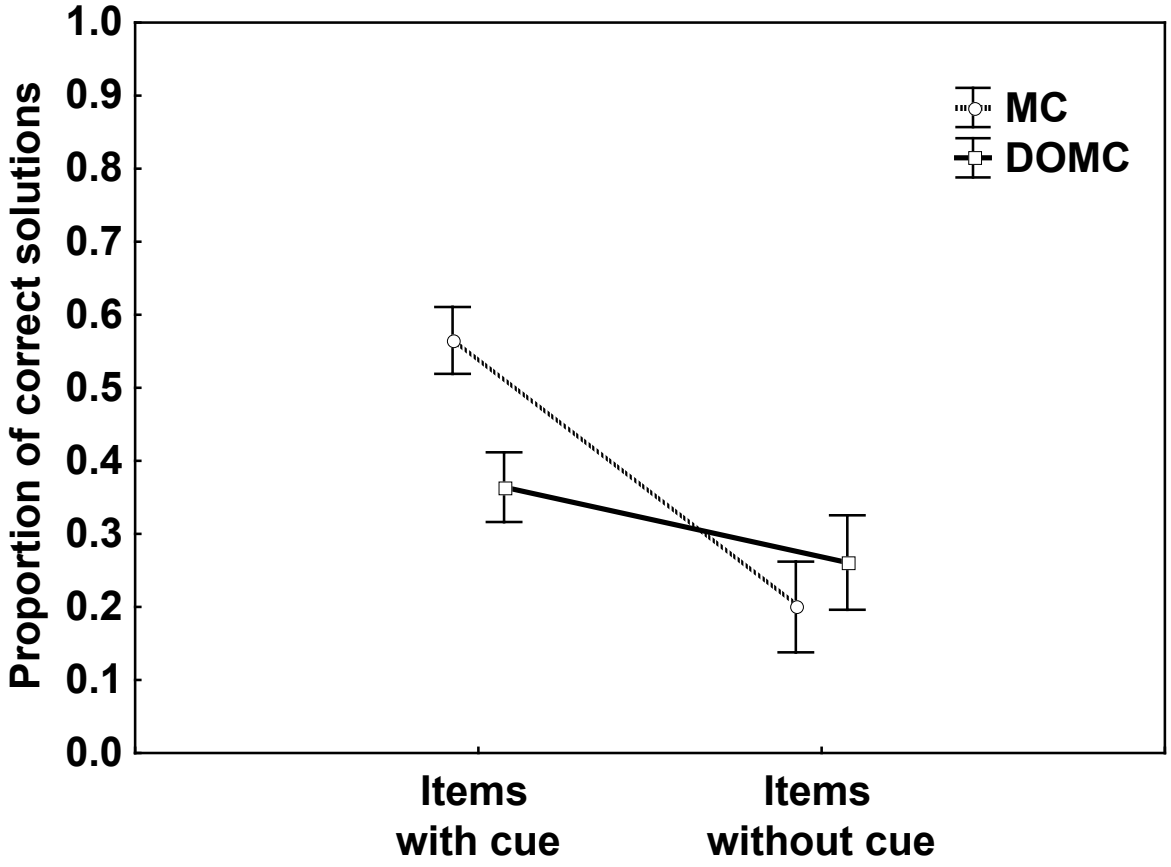

Figure 1. Proportions of correct solutions and their standard errors as a function of answer format and cue availability in Experiment 1.

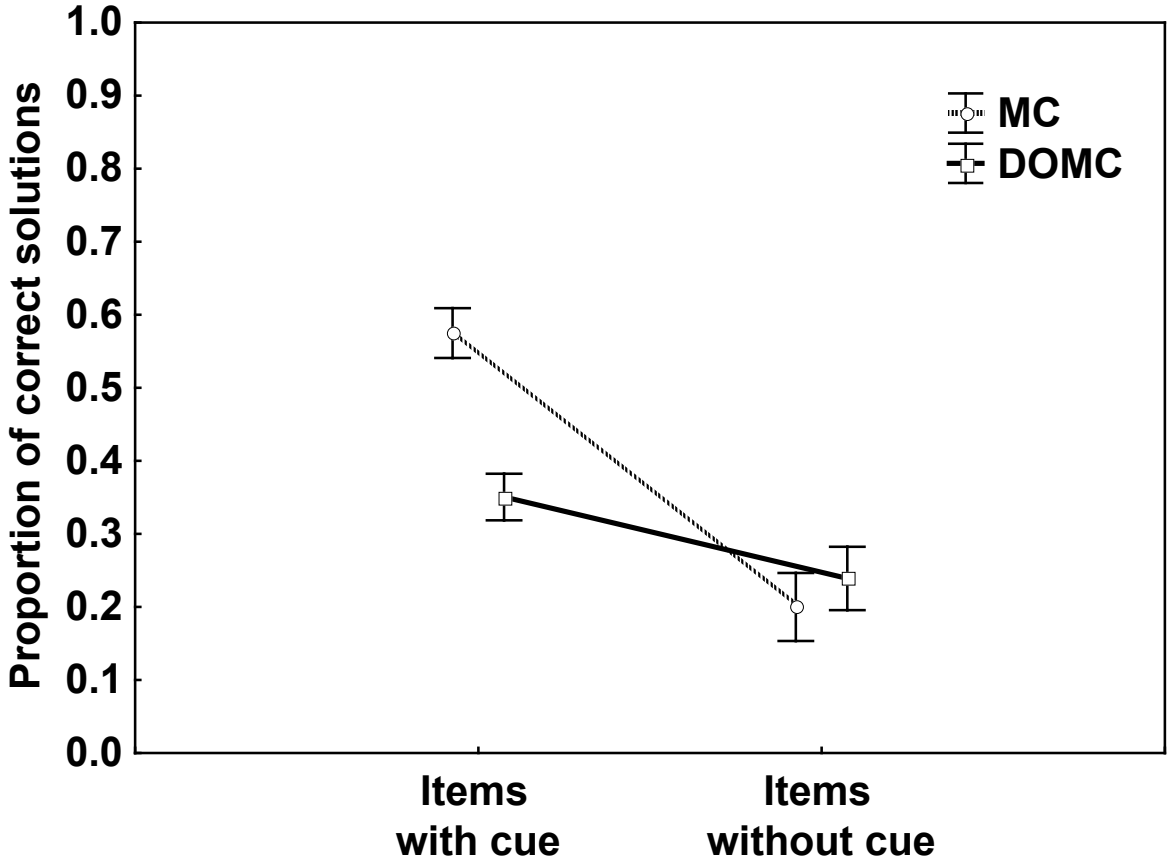

Figure 2. Proportions of correct solutions and their standard errors as a function of answer format and cue availability in Experiment 2.

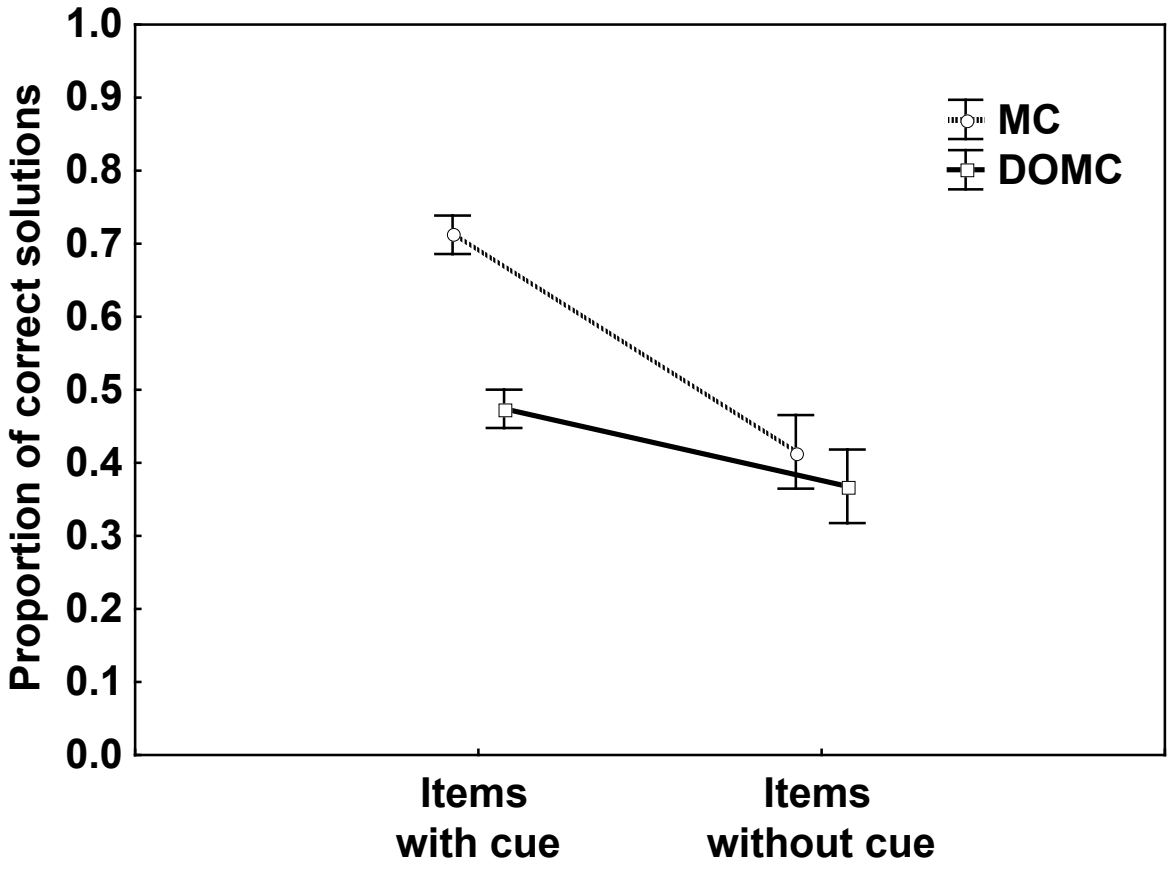

Figure 3. Proportions of correct solutions and their standard errors as a function of answer format and cue availability in Experiment 3.

Sequentially Presented Answer Options Prevent the Use of Testwiseness Cues in Multiple-Choice Testing

Martin Papenberg, Sonja Willing, and Jochen Musch

Department of Experimental Psychology

University of Düsseldorf

Correspondence address:

Sonja Willing

University of Duesseldorf

Institute of Experimental Psychology

Building 23.03

Universitaetsstr. 1

D-40225 Düsseldorf

Germany

Email: sonja.willing@hhu.de

Abstract

Background. The multiple-choice (MC) test format is one of the most popular testing formats for the assessment of knowledge. However, testwiseness – the ability to find subtle cues to the solution by comparing all available answer options – threatens the validity of MC tests.

Aim. We investigated discrete-option multiple-choice (DOMC) testing, an alternative testing format that has recently been proposed by Foster and Miller (2009). In DOMC testing, answer options are presented sequentially rather than simultaneously.

Sample. Participants were 181 psychology students.

Method. A test consisting of items that included cues to their solutions was constructed to test whether DOMC testing allows for a better control of testwiseness than MC testing.

Results. Although test items were generally more difficult in the DOMC than in the MC format, the availability of item cues led to an increase in test scores that was considerably larger in the MC condition. DOMC was thus shown to allow for a better control of testwiseness than MC. DOMC testing also reduced the number of answer options that had to be presented.

Conclusion. The DOMC format seems to deserve further study as an interesting alternative to traditional MC testing.

Keywords: discrete-option multiple-choice, item cues, sequential item presentation, testwiseness, multiple-choice test

Sequentially Presented Answer Options Prevent the use of Testwiseness Cues in Multiple-Choice Testing

Multiple-choice testing is one of the most popular testing formats for the assessment of knowledge. It is widely used in diverse settings including school tests, university exams, vocational aptitude tests, and even TV quiz shows. In its standard form, a multiple-choice (henceforth MC) item consists of a stem and a set of three to five answer options, one of which is the solution (Foster & Miller, 2009). The stem is the beginning part of an item and presents the question that has to be answered. Next to the stem, all possible answer options are presented. The examinee's task is to choose the correct answer from among this set of options. Usually, all options (i.e., the solution and the distractors) are presented simultaneously to the test taker.

MC testing of this kind provides an efficient way to reliably measure cognitive ability. Unlike other test formats such as open questions or essays, MC tests can be scored easily, objectively, and even in an automated manner, rendering the testing of large groups feasible (Tamir, 1991). Considering the approximately 90 years of research on MC tests, Downing (2006) concluded that there is strong evidence for the validity of MC testing across a wide range of areas.

Critics, however, have remarked that the selection of an answer option for an MC item does not directly reveal the actual knowledge of a respondent, but rather indicates the alternative that the respondent considers to be most likely to be true (Holmes, 2002). This choice is based on a comparison that is performed by a taking all available options into account simultaneously. Therefore, a drawback of the MC test format is that cues that indicate which solution is correct may be derived or identified by comparing the various answer options.

Gibb (1964) defined testwiseness as the ability to find and to make use of such extraneous cues in MC items. Item cues have been shown to make MC items less difficult, and testwise persons who are capable of making use of item cues may use these cues to increase their test scores (Allan, 1992). Rost and Sparfeldt (2007) surprisingly found that by comparing all available answer options, pupils could often identify the correct solution without even knowing the question.

Item cues that can be used to identify the correct answer also reduce the construct validity of MC items if individual differences in testwiseness that need not necessarily be related to the examinee's knowledge (Millman, Bishop & Ebel, 1965) add construct-irrelevant variance to MC test scores (Haladyna & Downing, 2004; Rost & Sparfeldt, 2007). In principle, items on carefully constructed tests should not be solvable by simply using testwiseness strategies if guidelines for good item writing practices are followed (Haladyna, 2004). However, many MC items are created under time pressure and by authors who have little experience with test development (Downing, 2006). Accordingly, Brozo, Schmelzer, and Spires (1984) found that even in a sample of 1,220 MC items that had been used in real college examinations, 44% of the items contained one of 10 different kinds of item cues. On average, for these flawed items, using the available cues almost tripled the probability of a correct solution as compared to a baseline of random guessing. In a more recent study, Tarrant and Ware (2008) analyzed 10 tests that had been used for high-stakes assessments in a nursing program. He also found that between 28 - 75% of the MC test items contained flaws, most of which favored testwise students.

Testing formats that control for the application of testwiseness are therefore desirable, but none have been available. Recently, however, Foster and Miller (2009) proposed an alternative testing format that prohibits the comparison of the available answer options by presenting them sequentially, rather than simultaneously. They called this variant of MC

testing, which typically has to be performed using a computer for practical reasons, *discrete-option multiple-choice* (henceforth DOMC) testing.

Like a standard MC item, a DOMC item consists of a stem and a number of answer options, one of which is the solution (Foster & Miller, 2009). The difference from standard MC items is that answer options are not presented simultaneously, but one at a time in a random order. For each single option, the test taker therefore has to make a decision about whether it is the correct solution or not. Unlike MC items, DOMC items are usually answered before all answer options have been presented. This is because in DOMC testing, the presentation of an item ends when one of the following conditions is met: (a) the solution has been correctly identified as such (in this case, no more answer options need to be presented); (b) the solution has incorrectly been rejected, or (c) a distractor has incorrectly been accepted. In the latter two cases, there is also no need to present additional answer options because the item has already been answered incorrectly. In other words, the presentation of a DOMC item ends as soon as it has been answered correctly or incorrectly. After the presentation of a DOMC item ends, none of the remaining answer options is shown; instead, the next question is presented. This feature of DOMC testing may help to reduce testing time in spite of the sequential presentation, and Foster and Miller (2009) indeed observed that, compared to MC, DOMC reduced testing time by about 10%. Foster and Miller (2009) also identified the limited exposure of the various answer options as another advantage of the new format they proposed. If an answer option is never presented to a participant, he or she cannot recall it or give it away to future participants. Test security is thus enhanced, and the reuse of DOMC items on future exams is made easier. Taken together, these potential advantages of DOMC testing make it worthy of further exploration.

In a first test of their new format, Foster and Miller (2009) found that DOMC questions were more difficult than standard MC questions. This finding was replicated in a

subsequent study using a larger sample (Kingston, Tiemann, Miller, & Foster, 2012). A likely explanation for this higher difficulty is that in the DOMC format, it is no longer possible to compare the plausibility of all available answer options; rather, the examinee repeatedly has to make decisions on the basis of the limited information that is provided by each single option. To make correct decisions in sequential DOMC testing, the examinee therefore has to be able to assess the correctness of each answer option separately, whereas in MC testing, all answer options can be considered simultaneously to identify the correct solution. Foster and Miller (2009) surmised that DOMC testing might therefore motivate deeper learning because the solution has to be identified by the learner without the help of accompanying distractors. Most important for the present investigation, however, is that not being able to compare sequentially presented answer options may help to prevent the use of item cues. Both Foster and Miller (2009) and Kingston et al. (2012) have therefore argued that DOMC may help to control for the application of testwiseness. Although this assertion is plausible, direct evidence for the alleged improved control of testwiseness is lacking in what is still a small body of research on the properties of the DOMC answer format at this time. In the present study, we therefore wanted to investigate whether DOMC testing does indeed help to control for testwiseness better than the traditional MC format. To this end, we presented examinees with a test that contained cues about the correct solution in each item and checked whether these cues could be used less easily in DOMC testing.

Several tests have been constructed to measure the ability of individuals to take advantage of the existence of item cues (e.g., Gibb, 1964; Diamond & Evans, 1972). A test of testwiseness needs to fulfill the following criteria: First, the test questions must be rather difficult; participants should normally not have much knowledge that would allow them to answer the questions. Second, each question must contain an item cue, which, if used cleverly, will allow the test taker to identify the correct solution or at least to increase the

person's probability of identifying the correct solution. If these criteria are met, an item on a test of testwiseness can be solved if the item cue is recognized and applied by the test taker. The number of items that can be solved correctly can then be used as an index of the examinee's testwiseness. Unfortunately, to the best of our knowledge, no test of testwiseness has ever been published in the German language. Because the content of existing instruments is often rather culture-specific, we therefore constructed a new test for the present study, the details of which are provided below in the Method section. After constructing this test of testwiseness, we also created a parallel control test by removing all cues from the testwiseness test items. In our experiment, we were thus able to create a condition in which participants were asked to solve items that did not contain any cues (no cue condition) or in which they were asked to solve items containing such cues (cue condition). To establish an additional group that would take a test that was even more susceptible to the use of item cues, we asked a third group of participants to work on a test that also contained item cues, and we additionally informed the participants in this group about the presence and the nature of these cues (informed cue condition). We created this third condition to examine whether DOMC can reduce the use of testwiseness even when examinees are explicitly informed about the presence of cues. We randomly assigned participants to each of the three groups, and within these groups, we randomly assigned the participants to either the MC or the DOMC condition.

Our main hypothesis was that with the increasing availability of item cues, the difference in test scores between the DOMC and MC conditions would increase because the DOMC format was expected to allow for a much better control of testwiseness than the MC format. In particular, we expected that the susceptibility of items to the use of testwiseness would be lowest in the no cue condition, would be larger in the cue condition, and would be largest in the informed cue condition. If DOMC allows for a better control of testwiseness than the MC format, this should lead to an interaction between the cue condition and the

answer format such that the difference between MC and DOMC test scores would be larger when item cues were present and would be largest when item cues were not only present but when their presence was also made known to the respondents to make sure that the cues were noticed. In the informed cue condition, we therefore expected MC participants to profit considerably from the available item cues, whereas we expected DOMC testing to hinder participants from making a similarly extensive use of the item cues. In addition to the predicted interaction, we also expected a possible main effect of the testing format as both Foster and Miller (2009) and Kingston et al. (2012) had observed that MC items are typically easier to answer than sequentially presented DOMC items. For this reason, a difference between the scores in the MC and the DOMC conditions was expected to arise even when no cues were present to be taken advantage of.

A secondary purpose of the present study was to investigate the efficiency of the new DOMC answer format. This was done by calculating the reduction in the number of answer options that needed to be presented to the examinee by using the DOMC format and by determining the decrease in testing time that could thus be achieved.

Method

Participants. We conducted the experiment using a sample consisting of 181 psychology students (85.6% female) between the ages of 19 and 35 years (M = 22.79, SD = 2.80). All participants were recruited via announcements in psychology student groups in the German social network "studiVZ." The data of an additional 23 participants who did not finish the questionnaire had to be discarded; the number of dropouts did not differ between the response format conditions, $\chi^2(1) = 1.83$, *ns.* At the end of the test, participants were debriefed and thanked and were provided with the answers to all test questions.

Materials. We constructed a German test of testwiseness that was based on the comprehensive taxonomy of testwiseness cues published by Millman et al. (1965). It consisted of items containing one of the following four cues that were also described by Gibb (1964) and Brozo et al. (1984):

1. *Direct Opposites* (Brozo et al., 1984). When two alternatives are directly opposite in meaning, one of them is usually correct. An example item we constructed using this cue read:

Dissolving ammonium nitrate in water leads to

a) an increase in temperature

b) a clouding of the water

c) *a decrease in temperature*

d) a blue color change

Using the direct opposites test cue, even a completely naïve test taker can increase the probability of guessing the correct solution from 25% to 50%. In their analysis of a sample of 1,220 MC items that had actually been used in real college examinations, Brozo et al. (1984) found that 151 of these items (12.4%) contained this cue.

2. Longest Alternative (Gibb, 1964; Brozo et al., 1984). Many teachers tend to take more care in elaborating the real solution than when formulating distractors. If one alternative is more verbose than other alternatives, it is therefore often the solution. When constructing items using this cue, we followed Brozo et al.'s (1984) recommendation and operationally defined this cue as the situation in which one alternative is one line of print longer than the other alternatives. In their analysis of a sample of 1,220 MC items that had been used in real college

examinations, Brozo et al. (1984) found that 54 of these items (4.4%) contained this cue. This is an example we used on our test:

Zombia…

a) was a Mongolian emperor of the 12th Century.

b) is a relatively short fan palm discovered on the island Hispaniola with clustered stems and a very distinctive appearance caused by its persistent spiny leaf sheaths.

c) is a horror movie from the 70s.

d) is a Romanian mythical creature.

3. *Middle Value* (Brozo et al., 1984). Given a list of alternatives that can be ordered from small to large, one of the middle values rather than one of the extreme values is typically the correct solution. In their analysis of 1,220 sample items that had been used in real college examinations, Brozo et al. (1984) found that in 65 out of 79 (82.3%) items that had rank-ordered alternatives, one of the middle values was the solution. This is an example of an item we constructed for our test containing this cue:

When did the Roman emperor Septimius Severus die?

a) 480 AD

b) 395 AD

c) 211 AD

d) 103 AD

4. *Categorical Exclusives (Gibb, 1964)*. In an attempt to make distractors wrong, teachers often construct distractor items by including overgeneralizations based on words such as

"never," "always," or "absolutely." According to Gibb (1964), the solution is often more general and can therefore often be found by looking for answer alternatives that do not include one of these overgeneralizing qualifiers. This is an example of an item we constructed containing this cue:

The Austrian composer Alban Berg (1885 - 1935)

a) never created a composition for the violin.

b) lost all of his seven children to typhus.

c) exclusively set music to Theodor Fontane's work.

d) was born in Vienna and also died there.

We constructed six items for each of the above four cues; the final test thus consisted of 24 items. Each item consisted of a stem and four answer options with one correct solution. The content of the items was taken from a number of different domains of general knowledge including history, sports, mineralogy, and botany, among others. All questions were rather difficult and typically could not be solved using personal knowledge; instead, each item contained exactly one cue that could be used to infer the solution.

For each of the 24 testwiseness items, a twin item was created in which the item cue was removed. For example, to avoid the direct opposites cue, one of the direct opposites was removed from the set of available answer options and replaced with a new answer alternative. To remove the longest alternative cue, we either shortened the solution, lengthened the distractors, or both. The middle value cue was removed by making one of the extreme alternatives the solution. Finally, the categorical exclusives cue was avoided by removing overgeneralizing qualifiers such as "never" or "always."

All items were presented in an online questionnaire using the software *Unipark* (Version 7.1, Global Park AG, Germany). The sequence of the items was arranged in a random order in both the MC and the DOMC conditions. Answer options were also presented in a random order. In the MC condition, one item was presented per page along with all of the possible answer options. In the DOMC condition, answer options were presented sequentially.

Design. The study used a 2 x 3 between-subjects design. The first factor consisted of the *testing format* and compared the two levels MC and DOMC. The second factor consisted of the *availability of testwiseness cues*. This factor had three levels to establish the (a) no cue, (b) cue, and (c) informed cue conditions. The susceptibility of the items to the application of testwiseness cues increased from the first to the last level of this factor.

Procedure. At the beginning of the questionnaire, participants were asked to indicate their age, sex, and education. They were then randomly assigned to one the six experimental conditions that resulted from crossing the 2x3 levels of the two experimental factors. Participants were first introduced to the testing format that was used on the test. As the DOMC format was expected to be less familiar, its description had to be more detailed. The DOMC procedure was explained using a sample item, and participants were informed about the stopping criteria employed in the sequential presentation procedure. Participants worked on test items that did not contain any item cues in the no cue condition. In the cue condition, all participants worked on items that contained such cues. In the informed cue condition, participants worked on items containing cues and were additionally informed about the presence and the nature of these cues before the test began. To this end, each of the four cues was described and an example of an item containing the cue was given.

Results

For each participant, all responses were recorded, and a total test score for the 24 items was computed. Additionally, we recorded the time needed to read the instructions and to complete all items. For the statistical tests, an alpha level of .05 was used. Effect sizes for the difference between two means were calculated using Cohen's *d*. ANOVA effect sizes were computed using the classical eta-squared (η^2), indicating the proportion of the variance explained by each factor or interaction.

Testwiseness Scores. To compare the testwiseness scores across conditions, a 2 x 3 (testing format [DOMC, MC] x availability of testwiseness cues [no cue, cue, informed cue]) ANOVA was computed. Participants in the MC condition solved more items (M = 10.90, SD = 5.43) than participants in the DOMC condition (M = 7.27, SD = 3.51). This difference was statistically significant, $F(1, 175) = 53.56$, $p < .001$, $\eta^2 = 0.10$. Test scores also increased as a function of the availability of item cues. Participants in the no cue condition obtained lower scores (M = 5.87, SD = 2.46) than participants in the cue condition (M = 7.63, SD = 3.21) and participants in the informed cue condition (M = 14.20, SD = 4.39). This effect of the cue availability factor was significant, $F(2, 175) = 120.52$, $p < .001$, $\eta^2 = 0.45$. However, a significant interaction showed that participants in the MC condition were more successful in making use of an increased availability of item cues than participants in the DOMC condition, $F(2, 175) = 12.87$, $p < .001$, $\eta^2 = 0.05$ (see Figure 1).

Additional t tests were computed to explore the nature of the interaction. All t tests were one-tailed because of the directed nature of our hypotheses, which predicted that the availability of items cues would make items easier and that the sequential presentation of answer options would make items more difficult. We found that participants obtained higher scores when cues were available than when they were not available. This was true both in the

MC condition, $t(60) = 2.61$, $p < .01$, $d = 0.66$, and in the DOMC condition, $t(56) = 2.26$, $p = .01$, $d = 0.59$. As compared to the cue condition, test scores were further increased by informing participants of the cues in the informed cue condition. Again, this was true both in the MC condition, $t(64) = 10.68$, $p < .001$, $d = 2.26$, and in the DOMC condition, $t(50) = 4.49$, $p < .001$, $d = 1.24$. Additional t tests also revealed that regardless of the availability of cues, participants who were given items in the MC format scored higher than participants who were given items in the DOMC format. This was true in the no cue condition, $t(61) = 2.23$, $p = .03$, $d = 0.56$, the cue condition, $t(55) = 2.33$, $p = .01$, $d = 0.62$, and in the informed cue condition, $t(59) = 7.61$, $p < .001$, $d = 1.96$.

Number of Answer Options Presented in the DOMC Condition. In the DOMC condition, the presentation of answer options stopped whenever a distractor was erroneously accepted as the solution. Moreover, the presentation always stopped after the presentation of the solution because the solution could only be correctly accepted or wrongly rejected, and both of these outcomes rendered it unnecessary to present additional answer options. The position of the solution was randomly varied. The stopping criteria reduced the average number of answer options that were presented to the test takers in the DOMC condition. Because the solution was presented in each of the four possible positions with equal probability, a perfectly knowledgeable test taker who never incorrectly accepted a distractor could be expected to complete each item with an equal probability (p = .25) after each of the four answer options. Thus, on average, a perfect test taker could be expected to see 2.5 out of the 4 possible answer options in the DOMC condition. For a less than perfect test taker, the presentation of a smaller number of answer options had to be expected because in the DOMC condition, the presentation of the answer items stopped whenever a distractor was wrongly accepted as the solution. Taken together, this resulted in a positively skewed distribution of the average number of options that were presented to the test takers in the DOMC condition.

In particular, we found that in 40.5% of cases, the item presentation ended after the presentation of the very first option. In 24.3% of cases, this option happened to be the solution, and in 16.2% of cases, this option was a distractor that was wrongly accepted as the solution. The item presentation ended after the second, third, and fourth answer options were presented for 32.0%, 20.4%, and 7.1% of all items, respectively. On average, this resulted in an end to the item presentation after 1.94 out of the four possible answer options ($SD = 0.94$).

Testing Times. A *t* test was computed to compare the testing times between the DOMC and MC conditions. Participants in the DOMC condition ($M = 358.58$ s, $SD = 147.56$) finished the test significantly faster than participants in the MC condition ($M = 454.52$ s, $SD = 209.44$), $t(179) = 3.50$, $p < .001$, $d = 0.53$. Thus, due to the smaller number of answer options that had to be presented in the DOMC condition, the time needed to answer all items was reduced by 21% when the answer options were presented sequentially. However, participants needed longer to read the extended instructions in the DOMC condition ($M = 82.78$, $SD = 50.11$ vs. $M = 20.44$, $SD = 9.30$), $t(179) = 12.08$, $p < .001$, $d = 1.73$. When the time needed to read the instructions was added to the total testing time, the total time needed for the test was no longer significantly different between the MC ($M = 474.96$, $SD = 212.13$) and DOMC conditions ($M = 441.36$, $SD = 174.44$), $t(179) = 1.12$, *ns*.

Discussion

The present experiment shows that the DOMC answer format is capable of preventing the use of item cues better than the traditional MC format. Even though the availability of item cues led to an increase in test scores in both conditions, this increase was larger in the MC condition. Although items were generally more difficult in the DOMC than in the MC format, this effect was strongest when item cues were present and participants knew about

these cues. As compared to the uninformed control condition, knowledge about the presence of item cues allowed participants to correctly answer an additional eight out of 24 questions in the MC condition. In the DOMC condition, the improved control of the use of testwiseness cues that resulted from the sequential presentation of the answer options reduced this advantage to only four items. Thus, the DOMC format allowed for a considerably better control of testwiseness than the MC format. However, it is also true that this control was less than perfect, considering that the test scores profited from the availability of item cues even in the DOMC condition. This was most likely because some item cues could be used even in the DOMC condition; for example, in those cases in which all answer options were presented before one of the stopping criteria was met. Nevertheless, the DOMC format allowed for an improved control of testwiseness that was greatly superior to that of the MC condition.

Kingston et al. (2012) found that DOMC items were more difficult than MC items and surmised that this might be due to the better control of testwiseness that is afforded by the DOMC answer format. We found that even in the no cue condition, participants scored lower when given the test items in the DOMC format. This suggests that a higher item difficulty might be a stable property of the DOMC format that cannot be attributed solely to a better control of testwiseness.

An analysis of the number of answer options that was presented in the DOMC condition helped us understand why this format is more efficient in controlling for testwiseness than MC. In most cases (40.5%), the presentation of DOMC items ended after the presentation of only one of the four possible answer options. Only 1.94 options had to be shown on average, and in only 7.1% of all items were all four answer options presented to the test taker. This large reduction in the number of answer options that were available for comparison made it difficult for test takers to take full advantage of the item cues in the DOMC condition. Moreover, even when all four answer options were presented, the memory

load required to take advantage of the available item cues was still considerably larger in the DOMC condition, owing to the sequential presentation of the answer options. Test security was also enhanced because many answer options were not presented at all; the reuse of DOMC items in future examinations was thus made easier.

A reduction in test time may be seen as an additional advantage of the DOMC answer format. Even though this reduction was no longer significant when the time needed for the extended instructions was taken into account in the present investigation, there is little doubt that instructions can be shortened considerably once the test takers are familiar with the new format.

In summary, there seem to be three important characteristics of the new DOMC format. First, our experiment showed that the DOMC format allows for a better control of testwiseness than traditional MC testing. Second, DOMC testing reduces the number of answer options that are presented to the test taker and that are available for comparison when trying to arrive at the correct solution. This enhances both test difficulty and test security. Third, DOMC seems to have the potential to reduce testing time, at least once the test takers get accustomed to the new format and no longer need lengthy instructions. DOMC testing therefore seems to offer a promising alternative to the traditional MC format, and it seems worthwhile to further explore the usefulness of this new testing procedure.

References

Allan, A. (1992). Development and validation of a scale to measure testwiseness in EFL/ESL reading test takers. *Language Testing*, *9*, 101-119. doi:10.1177/026553229200900201

Brozo, W. G., Schmelzer, R. V., & Spires, H. A. (1984). A study of testwiseness clues in college and university teacher-made tests with implications for academic assistance centers *(Technical Report 84-0*1). Georgia State University: College Reading and Learning Assistance. ERIC database (ED240928).

Diamond, J. J., & Evans, W. J. (1972). An investigation of the cognitive correlates of testwiseness. *Journal of Educational Measurement*, *9*, 145-150. doi:10.1111/j.1745-3984.1972.tb00771.x

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum.

Foster, D., & Miller, H. L. (2009). A new format for multiple-choice testing: Discrete-option multiple-choice. Results from early studies. *Psychology Science Quarterly, 51*, 355-369.

Gibb, B. G. (1964). *Testwiseness as secondary cue response* (Doctoral dissertation). Stanford University, Ann Arbor, Michigan: University Microfilms, 1964. No. 64-7643.

Haladyna, T. M. (2004). *Developing and validating multiple choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*, 17-27. doi:10.1111/j.1745-3992.2004.tb00149.x

Holmes, P. (2002). *Multiple evaluation versus multiple choice as testing paradigm* (Unpublished doctoral dissertation). Twente University, Enschede.

Kingston, N. M., Tiemann, G. C., Miller, H. L., & Foster, D. (2012). An analysis of the discrete-option multiple-choice item type. *Psychological Test and Assessment Modeling, 54*, 3-19.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of testwiseness. *Educational and Psychological Measurement*, *25*, 707-726. doi:10.1177/001316446502500304

Rost, D. H., & Sparfeldt, J. R. (2007). Leseverständnis ohne Lesen? Zur Konstruktvalidität von multiple-choice-Leseverständnistestaufgaben [Reading comprehension without reading? On the construct validity of multiple-choice reading comprehension test items]. *Zeitschrift für Pädagogische Psychologie, 21,* 305-314. doi:10.1024/1010-0652.21.3.305

Tamir, P. (1991). Multiple-choice items: How to gain the most out of them. *Biochemical Education, 19*, 188-192. doi:10.1016/0307-4412(91)90094-O

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, *42*, 198-206. doi:10.1111/j.1365-2923.2007.02957.x

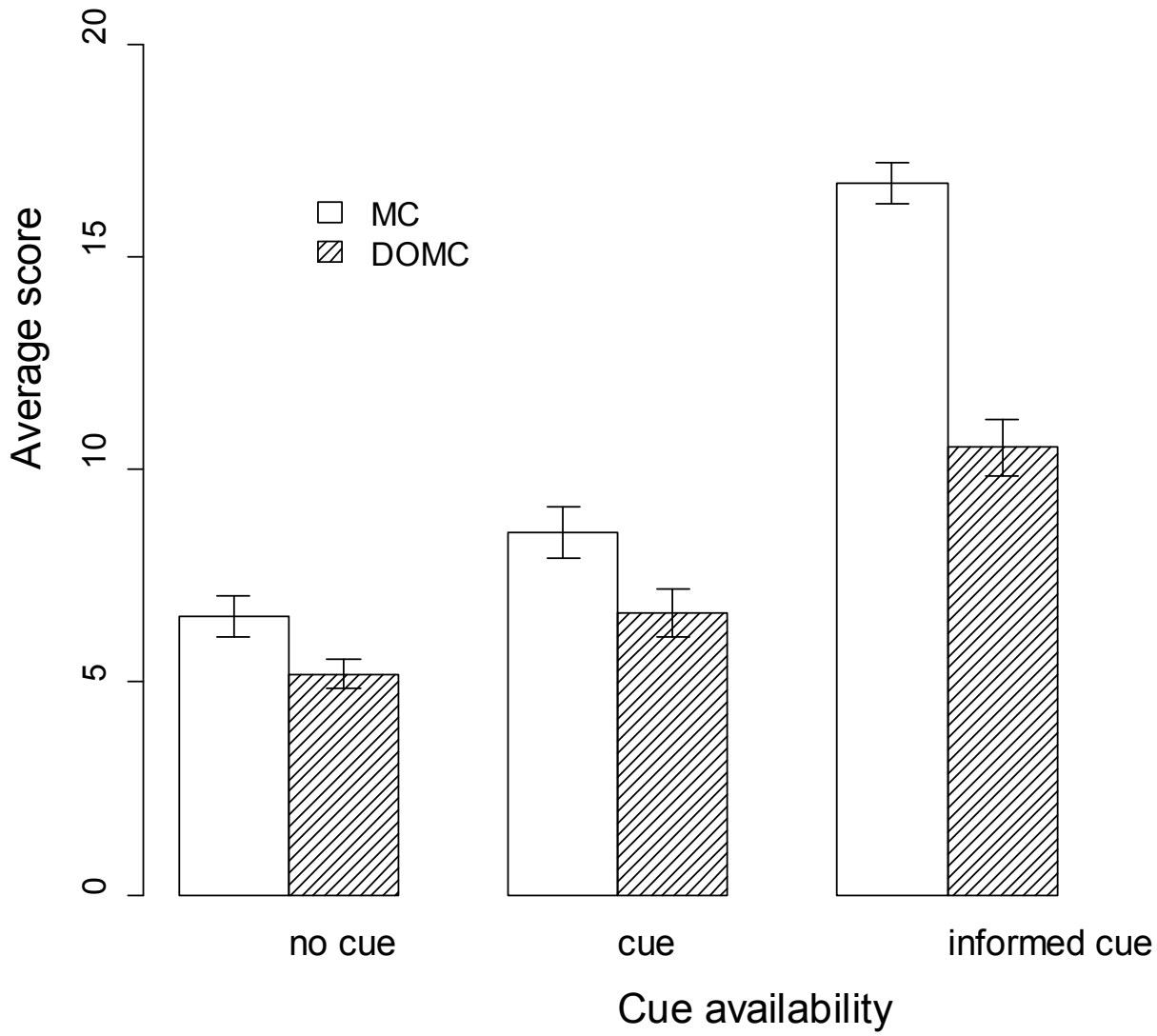

Figure 1. Test scores and their standard errors as a function of testing format and cue availability.

The Political Knowledge Test (PWT):

Development and Validation of a New Instrument for Assessing Political Knowledge

Sonja Willing and Jochen Musch

Institute of Experimental Psychology

University of Duesseldorf

Author Note

Correspondence concerning this article should be addressed to Sonja Willing, University of Duesseldorf, Institute of Experimental Psychology, Building 23.03, Universitaetsstr. 1, D-40225 Düsseldorf, Germany, Email: sonja.willing@hhu.de, Phone: +49-211-81-11524, Fax: +49-211-81-11753

Abstract

Political knowledge is a necessary precondition for successful participation in public life and for the development of political culture in democracies. However, in the German language, a validated measure to assess an individual's level of political knowledge has been missing. Therefore, the current paper presents the development and validation of the German political knowledge test (Politikwissenstest; PWT). In a pretest, 31 multiple-choice items were selected to construct the final form of the PWT. In two cross-validation studies, the PWT showed good psychometric properties and a clear one-factor loading structure. Evidence for the convergent validity of the PWT was obtained in six additional validation studies by relating the test to (a) self-ratings of political knowledge, interests, and media use, (b) political knowledge items taken from general knowledge tests, and (c) various measures of intelligence. Discriminant validity was established with regard to the scientific knowledge subscales of a general knowledge test. Taken together, we found evidence that the PWT has high construct validity; thus, it is able to provide a reliable and valid assessment of an individuals' level of political knowledge.

Keywords:

political knowledge, knowledge test, knowledge assessment, crystallized intelligence, validation

The Political Knowledge Test (PWT):

Development and Validation of a New Instrument for Assessing Political Knowledge

Since 1949, democracy has been a constitutional principle in Germany according to Article 20, Paragraph 1 of the Basic Law for the Federal Republic of Germany. For their continued existence, democracies are dependent on politically mature citizens who are capable of democratic acting (Galston, 2001; Schmid & Watermann, 2010). Therefore, political knowledge on the part of the electorate is a necessary requirement for a successful and well-functioning democracy (Delli Carpini & Keeter, 1996; Wanka, 2001).

Knowledge includes facts, concepts, principles, and procedures that can be memorized or understood (Haladyna, 2004). It is frequently organized into different domains (Haladyna, 2004), one of which represents the field of politics (Hossiep, Schulte, & Frieg, 2010; Mickel, 2005). Political knowledge can be regarded as knowledge about the institutions, structures, people, and events of national and international politics. Political knowledge encourages democratic principles and political participation (Delli Carpini & Keeter, 1996; Galston, 2001); using political knowledge, a person can participate more successfully in public life. Politically informed people are better able to understand political events and public policy and are more likely to successfully advocate their interests through political actions (Delli Carpini & Keeter, 1996) and elections (Popkin & Dimock, 1999). Political knowledge also influences political judgments by facilitating the incorporation of new politically related facts (Gilens, 2001) and by increasing the consistency of people's views across issues and time (Delli Carpini & Keeter, 1996; Gilens, 2001). Politically ignorant citizens are less able to follow discussions on political issues, tend to judge representatives by their personal character instead of their political performance, and are significantly less inclined to participate in politics at all (Popkin & Dimock, 1999). Thus, taken together, political knowledge is a precondition for a democratic political culture and for effective political participation in

public life (Ingrisch, 1997; Popkin & Dimock, 1999). To obtain German citizenship by naturalization, a minimum knowledge of German politics, history, and culture is therefore required and tested to ensure that a potential new citizen meets the preconditions for successful political participation (Wilhelm, Hülür, Köller, & Radalewski, 2010).

Political knowledge can be conceptualized as part of an individual's crystallized intelligence. Whereas fluid intelligence is comprised of a person's primary thinking skills and the innate ability to adapt to new situations and new problems, Cattell (1963) described crystallized intelligence as the part of a person's acquired knowledge that can be used to solve problems. Unlike fluid intelligence, crystallized intelligence has been shown to increase over much of the life span (Horn & Cattell, 1967). Political knowledge may be thought of as comprising those aspects of crystallized intelligence that refer to the domain of politics.

Measurement of Political Knowledge

Empirical studies addressing political knowledge have been conducted primarily in the area of electoral analysis (e.g., Zaller, 1990) and in the communication sciences (e.g., Allen & Spilich, 1997; Bennett, Rhine, & Flickinger, 2000; Berkowitz & Pritchard, 1989; Conway, Lyckoff, Feldbaum, & Ahern, 1981; Lukesch, 1991). In particular, the influence of media use on political knowledge and the emergence of political knowledge among adolescents have been investigated in youth and socialization studies in Germany (Ingrisch, 1997; Oesterreich, 2003). This led to several tests that were not validated by the scientific standards of test development (Hossiep, Schulte, Frieg, & Schardien, 2010; Downing, 2006). Consequently, there is both an interest and a gap in the reliable and valid assessment of the construct. As political knowledge is highly culture-specific, a German political knowledge test is needed for the valid assessment of political knowledge in Germany. Surprisingly, no carefully validated test for measuring an individual's level of political knowledge has ever been published in the

German language with the exception of three short subscales from German general knowledge tests. These subscales are included in (a) the Differential Knowledge Test (DWT; Jäger & Fürntratt, 1968), (b) the Bochum Knowledge Test (BOWIT; Hossiep & Schulte, 2008), and (c) the SPIEGEL students PISA test (Trepte & Verbeet, 2010) as described below.

First, the *Differential Knowledge Test* (DWT) was introduced by Jäger and Fürntratt (1968) to assess an individual's general state of knowledge. The development of the test was primarily oriented toward the school subjects taught in middle and high schools. Hence, the DWT consists of 11 knowledge areas, one of which is politics. Each subscale contains 20 multiple-choice (MC) items with four answer options. At the subscale level, factor analyses indicated a five-factor solution explaining 64.8% of the total variance. The sport, finance, and politics subscales did not load on the two major factors that were identified as (a) knowledge of literature and arts and (b) scientific-technical knowledge. Most DWT items are now rather dated and require a critical review because for example, no item refers to European politics. Moreover, since its construction in 1968, several items from the political knowledge subscale of the DWT have become obsolete (e.g., Switzerland implemented women's suffrage in 1971).

Second, Hossiep and Schulte (2008) developed the *Bochum Knowledge Test* (BOWIT), which they primarily designed for students, graduates, and academically trained professionals and executives. According to factor analyses, the 11 subscales of the BOWIT form the two factors (a) knowledge in social sciences, which also includes the subscale "society and politics," and (b) knowledge in the natural and technical sciences (Hossiep & Schulte, 2008). Two parallel test versions exist, both consisting of 11 subscales containing 14 multiple choice (MC) items for which there are always five answer options, including the option *none of the above*. A test-retest reliability of r_{tt} = .95 and an internal consistency of α = .95 were computed for the total test; for the subscale "society and politics," an internal consistency of

α = .73 was observed (Hossiep & Schulte, 2008). The thematic assignment of the BOWIT items to the 11 subscales was criticized as being somewhat arbitrary (Liepmann & Beauducel, 2010). In particular, the subscale "society and politics" mixes two different areas of knowledge, making it difficult to assess these two domains separately.

Third, the *SPIEGEL students PISA test* (Trepte & Verbeet, 2010) was developed for the assessment of German students' general knowledge in 2009. The construct of general knowledge was defined by the following five areas of knowledge: politics, history, economy, culture, and science. For each knowledge area, four item sets were designed. Each item set consisted of nine items about half of which were MC items with four answer options. The reliability of the subscales was rather low and has been criticized as insufficient; Cronbach's α coefficients ranged from .58 - .62 (Hossiep, Schulte, Frieg, et al., 2010). Many items referred to current political events that are subject to rapid change and in part, have already become obsolete (e.g., "How many delegates are in the current 16th German Bundestag?"; "How many countries are members of the European Union?"). Data on the validity of the items are scarce (Hossiep, Schulte, Frieg, et al., 2010).

To summarize, items from the political knowledge subscales of present general knowledge tests (DWT, BOWIT, and SPIEGEL students PISA test) are not sufficient for a reliable assessment of the construct of political knowledge because they are too few in number, partly unreliable or obsolete and not sufficiently validated, and sometimes refer to domains other than politics. Therefore, the purpose of the present studies was to address these problems and to develop and validate a German political knowledge test that will enable a reliable and valid assessment of an individual's level of political knowledge.

Development of the Political Knowledge Test (PWT)

The PWT (the German abbreviation for "Politikwissenstest," translating to "political knowledge test") was intended to assess general political knowledge. Following the definition of political knowledge presented above, this was done by collecting items pertaining to institutions, structures, procedures, historical figures, and events of national and international politics. In a first step, 99 MC items were generated. Items were partly self-developed and partly taken or adapted from a variety of reference and quiz books containing introductions to and questions about general and political knowledge (e.g., Duden, 2007; Walther & Hayo, 2004). For standardization, all items were adapted, and where necessary, extended to a fixed number of four answer options each. In contrast to the DWT and the SPIEGEL students PISA test items, we considered only items pertaining to content that is not subject to rapid change. The resulting initial set of 99 items was administered online to a sample of 258 native German speakers (115 female, 44.6%) ranging in age from 15 to 79 years (M = 32.19, SD = 15.31). For a subsequent empirical item selection, the following criteria were considered: item difficulty (proportion of examinees correctly answering an item), item discrimination (part-whole correlation between item and test performance), and loadings on the main factor that resulted from a principal components analysis. For the final version, 31 items with an item discrimination > .35 (M = .47, SD = .07, range: .36 - .61) were retained. The mean item difficulty of the selected items was .71 (SD = .14, range: .39 - .95). The internal consistency of the 31 items according to Cronbach's α was .91. Based on a principal components analysis with a varimax rotation, the resulting distribution of eigenvalues (8.484, 1.636, 1.375, 1.301, 1.232, 1.088, 1.045, 1.019) indicated the existence of a single main factor that explained 27.4% of the total variance. Figure 1 shows an example item from the PWT.[1]

--- Insert Figure 1 here ---

[1] The PWT items are freely available by request for research purposes.

Hypotheses for Validation

To investigate the psychometric properties of the new test, the 31-item version of the PWT resulting from the empirical item selection was validated in eight studies. It was expected that the PWT would provide a reliable and valid assessment of an individual's level of political knowledge. We derived the following hypotheses to address (a) the convergent and discriminant validity of the PWT and (b) expected group differences in the PWT test score.

Hypothesis 1: The PWT measures a unitary construct; therefore, we expected it to show a one-factor loading structure and a high internal consistency that - even after a Spearman-Brown correction (Spearman, 1910) - would be significantly higher than the reliability of previous political knowledge subscales on general knowledge tests.

Hypothesis 2: We expected that the PWT would display a high test-retest reliability, thus indicating a high temporal stability. However, due to the reduced variance, a reduced reliability in selected samples taken from the upper ability range was expected.

Hypothesis 3: With regard to construct validity (Cronbach & Meehl, 1955), we expected that the PWT would show high correlations with external criteria that are relevant to the construct of political knowledge; in particular, (a) political subscales of general knowledge tests, (b) self-ratings of political knowledge and interest, (c) measures of political media consumption, and (d) verbal intelligence.

Hypothesis 4: For the same reason, we also expected that the PWT would display higher correlations with political knowledge subscales than with more general measures of crystallized and fluid intelligence, supporting the notion that the construct of political knowledge cannot simply be equated with the broader construct of general intelligence.

Hypothesis 5: In a factor analysis, Hossiep and Schulte (2008) found a two-factor loading structure for the BOWIT subscales (a) knowledge in the social sciences - including

political knowledge - and (b) knowledge in the natural and technical sciences. As evidence of its discriminant validity, we expected that the PWT would show higher correlations with political knowledge subscales of general knowledge tests than with scientific knowledge subscales. Accordingly, we also expected that among pupils, the PWT would show higher correlations with grades in political science than with grades in science subjects that are supposedly unrelated to political knowledge.

Hypothesis 6: According to findings summarized in Delli Carpini and Keeter (1996), differences in PWT test performance were expected for subgroups differing in gender and education. First, as in previous investigations (e.g., Trepte & Verbeet, 2010), differences were expected in favor of male participants. Second, political science students were expected to achieve better scores than psychology students. Moreover, differences were expected in favor of pupils taking an advanced social science course (German "Leistungskurs") rather than a more basic social science course (German "Grundkurs").

Method

Participants

To investigate its factor structure, construct validity, and presumed group differences, the PWT was administered to eight validation samples consisting of 984 participants. The demographic characteristics of the respective samples are shown in Table 1.

--- Insert Table 1 here ---

Sample 1 consisted of 98 native adults who were recruited via an online panel. The panel consisted of participants from previous studies conducted by our department that were unrelated to politics.

Sample 2 comprised 342 native adults who participated online and were recruited through a quiz portal.

Sample 3 was used to examine the convergent and discriminant validity of the PWT. For this purpose, a heterogeneous sample was used, consisting of 83 pupils from a grammar school in Neuss (North Rhine-Westphalia), 63 psychology students from the University of Duesseldorf, and 21 working adults.

Sample 4 consisted of 119 pupils from Grades 10 and 11 from various schools in North Rhine-Westphalia (47 pupils from a grammar school in Duesseldorf, 49 pupils from a middle school in Monheim, and 23 pupils from a comprehensive school in Duesseldorf).

Sample 5 consisted of 176 pupils from Grades 11 to 13 from a local grammar school in Duesseldorf.

Sample 6 consisted of 20 students in the master program "Political Communication" and 20 psychology students from the University of Duesseldorf who had been matched on age, sex, and grade in the final secondary school examination (the German "Abitur").

Sample 7 consisted of 30 pupils from Grades 11 and 12 from a comprehensive school in Duesseldorf. These pupils were also retested after 4 weeks to determine the retest stability of the PWT.

Sample 8 was used for validation with an extreme group and consisted of 12 doctoral students from a German party-affiliated scholarship system. These doctoral students were also retested after 11 weeks.

Procedure

For Samples 1 and 2, all questions were delivered online using the software Unipark (Globalpark AG, 2009). For Samples 3 to 8, the PWT was delivered in group settings as a paper-and-pencil test. All pupils participated in a regular 45-min class lesson. For each

correctly solved item, 1 point was awarded. The total test score was calculated by adding the points obtained across all 31 items.

Material

To determine the convergent validity of the PWT, we used several external criteria. In particular, we included (a) self-ratings of political knowledge, interests, and media use, (b) political knowledge items taken from general knowledge tests, and (c) various measures of intelligence. Discriminant validity was assessed using the scientific subscales of a general knowledge test. Measures of academic performance were used to examine convergent and discriminant validity as described below.

Self-ratings of general and political knowledge. To obtain two measures of cognitive ability, participants in Samples 1 to 3 were asked to estimate their general and their political knowledge on a 7-point Likert scale ranging from 1 (*very poor*) to 7 (*very good*).

Self-reported measures of interest in politics and participation in political discussions. Hossiep and Schulte (2008) expected and found a relation between a person's knowledge in a given field and his or her interests in this field. In particular, a correlation of $r = .45$ was observed between self-reported political interest and the BOWIT (Hossiep & Schulte, 2008) subscale "society/politics." We assessed participants' interest in politics in all eight samples using a 7-point Likert scale ranging from 1 (*very low*) to 7 (*very high*).

Political events, issues, and controversies are often reflected in discussions with friends and family members. Such political debates and discussions can also be viewed as a form of political participation (Bennett, Flickinger, & Rhine, 2000) that is expected to be associated with an increase in political knowledge (Gesellschaft für Politikdidaktik und politische Jugend- und Erwachsenenbildung, 2004). Therefore, respondents were asked to indicate how often they participated in political discussions in their circle of family and friends using a 7-point Likert scale ranging from 1 (*very rarely*) to 7 (*very often*).

Political media use. Media are an integral part of social and political communication (Besand, 2005). We therefore expected that the use of mass media would have a positive impact on a person's level of political knowledge. Bennett, Rhine, et al. (2000) found that an individual's level of knowledge can be best predicted by his/her reading behavior. Likewise, Hossiep and Schulte (2008) demonstrated a correlation between knowledge and reading frequency of $r = .61$. In addition to print media, however, television and the Internet are increasingly being used to obtain political knowledge. Newspaper consumption is usually lower among younger people who tend to attribute less importance to print media. The Internet has become the main source of news for 16-24 year-old people, but people between the ages of 25-34 still read more online news than younger people (OECD, 2010). In 2008, the percentage of German people between the ages of 16-74 who read online newspapers or magazines was 21% (OECD, 2010). Moreover, as an opportunity "to inform themselves what is going on in the world," 29% of young people use the Internet every day and another 21% use it regularly every week (Albert, Hurrelmann, Quenzel, & TNS Infratest Sozialforschung, 2010). Consequently, participants' use of political media was assessed with three items. First, participants in Samples 1 to 7 were asked to report how often they read the 10 most widely circulated national newspapers and magazines (Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern, 2010) on a 7-point Likert scale ranging from 1 (*very rarely*) to 7 (*very often*). Second, participants in Samples 1 to 3 were requested to indicate how often they used (a) political television programs and (b) the Internet for political information each week.

Political knowledge items from the BOWIT. The BOWIT (Hossiep & Schulte, 2008) subscale "society/politics" includes 14 MC items on each of the two test versions. Of these 28 items, one item had to be excluded because it was identical to one DWT political knowledge item ("Who elects the Federal President?"). In order to restrict the assessment to only political

knowledge, nine additional items were excluded because they referred to topics other than politics (e.g., genealogy). To use the BOWIT as an external criterion, the remaining 18 MC items were presented to Sample 3.

Political knowledge items from the DWT. The DWT (Jäger & Fürntratt, 1968) subscale "politics" includes 20 MC items that differ between the two parallel test versions only in the order in which they are presented. Three of these items had to be excluded for the purpose of validation. This was because two items had become obsolete, and one item overlapped in content with a PWT item. Therefore, 17 of the 20 political knowledge items from the DWT were used as an external criterion in Sample 3.

Political knowledge items from the SPIEGEL students PISA test. The subscale "politics" from the SPIEGEL students PISA test (Trepte & Verbeet, 2010) consists of four item sets with nine items each. In Sample 4, we used two of these item sets, for which Hossiep, Schulte, Frieg, et al. (2010) demonstrated the highest internal consistency. In their study, for item set 2, Cronbach's α was .62 for students and .70 for non-students. For item set 4, they reported a Cronbach's α coefficient of .62 for students and of .71 for non-students. Two items containing content that would otherwise have been obsolete were updated. Nine items were presented in their original MC format with four answer options. Four of the remaining items required test-takers to specify a number to make an assignment on a map, whereas five items asked for a short answer in a free-response field.

Vocabulary test. A standardized vocabulary test (WST) by Schmidt and Metzler (1992) was used to provide an economical, reliable, and valid measurement of crystallized intelligence. Each of the 42 items on the WST consists of six word formations including one real word and five nonsense words. Participants are asked to identify the real word. The score is determined as the sum of the correctly answered items (Sample 4).

10-Minute-Test. The 10-Minute-Test is an unpublished short test for the assessment of general intelligence (Musch et al., 2009). It comprises items for the assessment of fluid intelligence (deductive reasoning) and crystallized intelligence (general knowledge, vocabulary) for which a total test score that is characterized by a high g-loading (Musch et al., 2009) is computed.

Trapnell Smart Scale. In Samples 4, 5, and 7, the Smart Scale (Trapnell, 1994) was used to assess self-reported intelligence. Self-reported intelligence has been found to correlate with crystallized intelligence in a number of studies using a variety of instruments; however, correlations rarely exceed .30 (Paulhus, Lysy, & Yik, 1998). Consisting of only four items, Trapnell's Smart Scale is a particularly economic but still reliable (Cronbach's α = .86 - .88) and valid (r = .24 - .25) instrument for assessing self-reported intelligence. On a 9-point Likert scale ranging from 1 (*low*) to 9 (*high*), participants are asked to indicate their degree of agreement with the following four statements: "I'm considered exceptionally or unusually intelligent," "I'm considered a very brainy or scholarly person," "I'm considered extremely 'gifted' or talented at academic things," and "My school grades have usually been near the top of every class."

Scientific knowledge subscales. To examine discriminant validity, we used the BOWIT scientific knowledge subscales that are supposed to be unrelated to the construct of political knowledge. Therefore, the subscales mathematics/physics and biology/chemistry of BOWIT test version A were administered to Sample 3. Each subscale consisted of 14 items with five answer options. For the two subscales, Hossiep and Schulte (2008) reported an international consistency of Cronbach's α = .82 (mathematics/physics) and α = .70 (biology/chemistry).

Academic performance. In the school and student samples (Samples 3 to 8), academic performance was used to determine convergent and discriminant validity. To this end,

participants were asked to provide self-reports of their grades ranging from 1.0 (*highest grade*) to 5.0 (*lowest grade*) in the following school subjects: political and social sciences, German, mathematics, and English. In Sample 3, participants were also asked to indicate their grades in scientific subjects (i.e., biology, chemistry, and physics).

Results

An alpha level of .05 was used for all statistical tests. Effect sizes for the difference between two means were calculated using Cohen's *d*. According to Cohen (1988), an effect of $d \geq 0.20$ may be considered small, an effect of $d \geq 0.50$ medium, and an effect of $d \geq 0.80$ large. ANOVA effect sizes were computed using eta-squared (η^2), which can be interpreted as the proportion of variance explained by each factor or interaction. $\eta^2 \geq 0.01$ implies a small effect, $\eta^2 \geq 0.06$ a moderate effect, and $\eta^2 \geq 0.14$ a large effect (Cohen, 1988).

Item analysis. In the total data set including all 984 participants, the 31 items of the PWT showed a mean item difficulty of .59 (*SD* = .16, range: .26 - .89) and a mean item discrimination of .47 (*SD* = .08, range: .34 - .66). Thus, the PWT items were found to be sufficiently difficult and discriminating.

Factor structure. To review the dimensionality of the PWT, we performed a principal components analysis with a varimax rotation. In Sample 1, Kaiser's eigenvalue criterion (Guttman, 1954) indicated a solution of 11 factors with eigenvalues > 1.0 (eigenvalues: 5.262, 2.413, 2.083, 1.780, 1.736, 1.586, 1.406, 1.361, 1.302, 1.217, 1.089). Together, the 11 factors accounted for 68.5% of the variance, with Factor 1 contributing 17.0% of the total variance. However, the distribution of the eigenvalues revealed only one main factor prior to the break in the scree plot (Cattell, 1966). In Sample 2, a principal components analysis revealed the presence of nine factors with eigenvalues exceeding 1.0 (i.e., 6.625, 1.476, 1.320, 1.301, 1.199, 1.127, 1.092, 1.057, 1.048). The nine factors explained 52.4% of the variance, with the

first factor contributing 21.4% of the variance. Again, an inspection of the scree plot revealed a clear break after the first factor. These findings support the notion that the PWT mainly consists of one major factor on which all items showed strong factor loadings. This main factor was also replicated in the subsequent cross-validation studies.

Reliability. Cronbach's α coefficient (Cronbach, 1951) was calculated as an estimate of the reliability of the 31 PWT items. In the present samples (Samples 1, 2, 3, 6, and 7), the internal consistency of the PWT ranged from α = .80 to α = .87 (see Table 1). Only in two rather homogeneous samples of pupils (Samples 4 and 5) and in a small homogeneous sample of 12 doctoral students (Sample 8) was a reduced α observed due to the reduction in variance (α = .55 - .66). For the total sample (N = 984), Cronbach's α was .91. Thus, satisfactory to high reliabilities were obtained for the final scale.

The program AlphaTest by Lautenschlager and Meade (2008) was used to test for differences in coefficient α between the 31 items of the PWT and previous political knowledge subscales, for which reliabilities were Spearman-Brown-corrected to allow for a fair comparison. In Sample 3, we found that the internal consistency of the PWT (α = .87) was significantly higher than the Spearman-Brown-corrected reliability of (a) the BOWIT political knowledge items (α = .79), $\chi^2(1) = 9.44$, $p < .01$, and (b) the DWT political items (α = .77), $\chi^2(1) = 12.33$, $p < .001$. This result was also replicated in Sample 5. The internal consistency of the PWT (α = .66) was significantly higher than the Spearman-Brown-corrected reliability of the BOWIT political knowledge items (α = .55), $\chi^2(1) = 3.21$, $p = .04$. However, when the reliability of the political knowledge items of the SPIEGEL students PISA test were Spearman-Brown-corrected for their small number, their internal consistency (α = .79) was higher than the internal consistency of the PWT (α = .66), $\chi^2(1) = 9.42$, $p < .01$.

In summary, the PWT items were characterized by a satisfactory reliability that was - even after a Spearman-Brown correction - significantly higher than the reliability of the short

political knowledge subscales of the BOWIT and the DWT. Hypothesis 1 was thus confirmed with the exception of the SPIEGEL students PISA test.

To determine retest reliability, participants in Sample 7 completed the PWT again 4 weeks after the first administration of the test. Test-retest reliability was $r_{tt} = .93$. As expected, a somewhat lower test-retest reliability ($r_{tt} = .66$) was observed for the subsample of 12 doctoral students (Sample 8), arguably due to the reduced variance in this sample. Hypothesis 2 was thus confirmed.

Validity. Convergent and discriminant validity was determined using Pearson correlations to examine the association between the PWT test score and the external validation criteria (see Table 2).

--- Insert Table 2 here ---

For statistical comparisons, a *z*-test for comparing dependent correlations was applied (Steiger, 1980). The PWT showed a higher correlation with self-reported political knowledge ($r = .51$, $r = .42$, $r = .48$) than with self-reported general knowledge ($r = .13$, $r = .27$, $r = .34$), $z = 4.14$, $p < .001$; $z = 3.57$, $p < .001$; $z = 2.14$, $p = .02$ (Samples 1 to 3).

In Sample 4, the PWT was more strongly associated with the political knowledge test items of the SPIEGEL students PISA test ($r = .59$) than with (a) the vocabulary test as a measure of crystallized intelligence ($r = .31$, $z = 3.19$, $p < .01$) and (b) the Trapnell Smart Scale as a measure of self-reported intelligence ($r = .36$, $z = 2.38$, $p = .02$). Furthermore, the PWT showed a higher correlation with the BOWIT political knowledge items ($r = .48$) than with the 10-Minute-Test as a measure of crystallized and fluid intelligence ($r = .24$), $z = 2.78$, $p < .01$ (Sample 5).

In addition, the PWT was also more strongly associated with the BOWIT political knowledge items (r = .77) than with the BOWIT subscale mathematics/physics (r = .38), z = 6.38, p < .001 (Sample 3). The PWT also showed a higher correlation with the BOWIT political knowledge items (r = .77) than with the BOWIT subscale biology/chemistry (r = .32), z = 7.01, p < .001. This result also held for the DWT political knowledge items: The PWT showed a higher correlation with the DWT political knowledge items (r = .71) than with both (a) the BOWIT subscale mathematics/physics (r = .38, z = 5.16, p < .001) and (b) the BOWIT subscale biology/chemistry (r = .32, z = 5.73, p < .001).

Additionally, the PWT was also more strongly correlated with school grades in the social or political sciences (r = -.46) than with grades in scientific and language subjects (Sample 3):

- mathematics: r = -.23, z = -2.99, p < .01,
- physics: r = -.16, z = -3.65, p < .001,
- biology: r = -.09, z = -5.12, p < .001,
- chemistry: r = -.11, z = -3.11, p < .01,
- German: r = -.28, z = -3.04, p < .01, and
- English: r = -.30, z = -2.45, p = .01.

To summarize, the PWT showed high correlations with (a) political knowledge items contained in general knowledge tests (BOWIT, DWT, SPIEGEL students PISA test), (b) self-ratings of political knowledge and interests, and (c) self-ratings of participation in political discussions. Second, the PWT displayed higher correlations with political knowledge items than with measures of intelligence. Third, the PWT was more strongly associated with political knowledge items from general knowledge tests than with scientific knowledge subscales; and the PWT was more strongly associated with school grades in politically related

subjects than with grades in scientific subjects and languages. This pattern supports the discriminant validity of the PWT. Hypotheses 3 to 5 were thus confirmed.

Group differences. In the total sample, male participants (M = 19.96, SD = 7.58) achieved, on average, higher test scores than female participants (M = 16.74, SD = 7.08), $t(982) = 6.88$, $p < .001$, $d = 0.44$. This gender difference was also observed in each subsample. A similar gender difference was observed for self-reported interest in politics: Male participants (M = 4.62, SD = 1.65) reported a significantly higher interest in politics than female participants (M = 3.95, SD = 1.54), $t(967) = 6.49$, $p < .001$, $d = 0.42$. Political science students (M = 25.90, SD = 2.69) obtained significantly higher PWT test scores than a matched sample of psychology students (M = 19.20, SD = 4.41) comparable in age, gender, and grade in their secondary school diploma (the German "Abitur"), $t(38) = 5.80$, $p < .01$, $d = 1.83$. Finally, pupils participating in an advanced social science course (German "Leistungskurs"; M = 15.21, SD = 3.02) scored higher on the PWT than pupils participating in the basic social science course (German "Grundkurs"; M = 11.93, SD = 4.03), $t(135) = 2.96$, $p < .01$, $d = 0.92$.

A one-way between-subjects analysis of covariance was conducted to compare the test scores for the political science students in Sample 6a and the adult participants in Sample 1. After adjusting for participants' age, there was no significant difference in the mean test score between the sample of political science students (M = 27.12, SE = 0.87) and the sample of adults (M = 25.73, SE = 0.37), $F(1,115) = 2.08$, $p = .15$, $\eta^2 = .02$. However, there was a significant relation between the age covariate and the PWT test score, $F(1,115) = 15.17$, $p < .001$, $\eta^2 = 0.12$. For the total sample of 984 participants, a correlation of $r = .54$ ($p < .001$) was observed between age and PWT test scores.

To summarize, Hypothesis 6 was confirmed: The PWT was able to reliably discriminate between groups with known or presumed differences in ability. In particular, group differences were demonstrated in favor of (a) male over female participants, (b) political

science students over psychology students, and (c) pupils in advanced social science courses over pupils in more basic social science courses. Moreover, test performance was significantly correlated with age.

Discussion

To date, no test that had been exclusively developed and validated to assess individual differences in political knowledge was available in the German language. To remedy this deficiency, we developed and validated the German political knowledge test (PWT).

Across a considerable number of validation studies, the PWT showed good psychometric properties. The PWT was found to be sufficiently difficult; it was also able to distinguish between politically knowledgeable and less politically informed individuals. The clear one-factor structure that was found for the PWT during empirical item selection was confirmed. Hence, the PWT total score provides a reliable overall measurement of political knowledge. The internal consistency of the PWT was satisfactory. Even after a Spearman-Brown correction, the PWT was significantly more reliable than the political knowledge subscales of two general knowledge tests (BOWIT, DWT). A satisfactory test-retest reliability of the PWT indicated that individual differences in political knowledge seem to be stable over time and can be reliably assessed using the PWT.

Evidence for the convergent validity of the PWT was obtained by relating the PWT to political knowledge subscales of general knowledge tests, self-ratings of political knowledge and interests, and self-ratings of participation in political discussions. Similar to other measures of crystallized intelligence, PWT test performance was significantly correlated with age. However, PWT scores showed higher correlations with self-reported political knowledge than with self-reported general knowledge. This finding suggests that political knowledge cannot simply be equated with general knowledge or general intelligence. Rather, political

knowledge turned out to be correlated with the use of political media. As in the studies by Bennett, Rhine, et al. (2000) and Hossiep and Schulte (2008), reading political newspapers and magazines strongly predicted an individual's level of political knowledge.

Discriminant validity of the PWT was established with regard to the scientific knowledge subscales of a general knowledge test and grades in scientific subjects that were unrelated to the construct of political knowledge. These results are in good agreement with the two-factor solution of general knowledge comprising (a) knowledge in the social sciences and (b) knowledge in the natural and technical sciences as reported by Hossiep and Schulte (2008).

Significant differences in political knowledge were also found between groups. Male participants showed a higher level of political knowledge than female participants. Additionally, political science students achieved higher test scores than psychology students even though the latter had been matched on age, gender, and final school grades. Pupils in the advanced social science course scored higher than pupils in the basic social science course. In summary, the PWT allowed for reliable discriminations between groups with known or presumed differences in ability.

Taken together, the result pattern demonstrates that the PWT has a clear one-factor loading structure, a high reliability that even after a Spearman-Brown correction surpasses the reliability of most previously existing political knowledge items, and a high construct validity. Thus, the PWT allows for a reliable and valid assessment of an individuals' level of political knowledge. In developing the PWT - in contrast to the procedures followed for the DWT and the SPIEGEL students PISA test - we selected only items that are not subject to rapid change over time. The PWT therefore offers a reliable tool for future inquiries into the field of political knowledge.

References

Albert, M., Hurrelmann, K., Quenzel, G., & TNS Infratest Sozialforschung (2010). *Jugend 2010. 16. Shell Jugendstudie* [*Youth 2010. 16th Shell Youth Study*]. Frankfurt am Main: Fischer Taschenbuch Verlag.

Allen, G. L., & Spilich, G. J. (1997). Children's political knowledge and memory for political news stories. *Child Study Journal, 27,* 163-177.

Bennett, S. E., Flickinger, R. S., & Rhine, S. L. (2000). Political talk over here, over there, over time. *British Journal of Political Science, 30,* 99-119.

Bennett, S. E., Rhine, S. L., & Flickinger, R. S. (2000). Reading's impact on democratic citizenship in America. *Political Behavior, 22,* 167-195.

Besand, A. (2005). Medienerziehung [Media education]. In W. Sander (Ed.), *Handbuch politische Bildung* [*Civic Education Manual*] (3nd edition, pp. 419-429). Schwalbach: Wochenschau Verlag.

Berkowitz, D., & Pritchard, D. (1989). Political knowledge and communication resources. *Journalism Quarterly, 66*, 697-702.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1-22.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245-276.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action.* Amsterdam: Elsevier.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Conway, M. M., Lyckoff, M. L., Feldbaum, W., & Ahern, D. (1981). The news media in children's political socialization. *Public Opinion Quarterly, 45*, 164-178.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Delli Carpini, M. X., & Keeter, S. (1996). *What Americans know about politics and why it matters.* New Haven: Yale University Press.

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.*), Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum.

Duden (2007). *Allgemeinbildung. Testen Sie ihr Wissen!* [*General education: Test your knowledge!*] (2nd edition). Mannheim: Duden Verlag.

Galston, W. A. (2001). Political knowledge, political engagement, and civic education. *Annual Review of Political Science, 4,* 217-234.

Gesellschaft für Politikdidaktik und politische Jugend- und Erwachsenenbildung [Society for civic education didactics and civic youth and adult education] (2004). *Anforderungen an nationale Bildungsstandards für den Fachunterricht in der Politischen Bildung an Schulen* [*Requirements for national education standards for school lessons in civics education*]. Schwalbach: Wochenschau Verlag.

Gilens, M. (2001). Political ignorance and collective policy preferences. *American Political Science Review, 95,* 379-396.

Globalpark AG (2009). *Enterprise Feedback Suite. EFS Survey.* Available from: http://www.unipark.info

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrica, 18,* 277-296.

Haladyna, T. M. (2004). *Developing and validating multiple choice test items* (3rd Ed). Mahwah: Lawrence Erlbaum.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26*, 107-129.

Hossiep, R., & Schulte, M. (2008). *BOWIT - Bochumer Wissenstest* [*BOWIT - Bochum knowledge test*]. Göttingen: Hogrefe.

Hossiep, R., Schulte, M., & Frieg, P. (2010). Was ist Wissen und wie lässt es sich messen [What is knowledge and how can it be measured]? In S. Trepte & M. Verbeet (Eds.), *Allgemeinbildung in Deutschland. Erkenntnisse aus dem SPIEGEL-Studentenpisa-Test [General knowledge in Germany. Findings from the SPIEGEL students PISA test]* (pp. 39-54). Wiesbaden: VS Verlag.

Hossiep, R., Schulte, M., Frieg, P., & Schardien, P. (2010). Wie gut misst der Studentenpisa-Test [How valid and reliable is the students PISA test]? In S. Trepte & M. Verbeet (Eds.), *Allgemeinbildung in Deutschland. Erkenntnisse aus dem SPIEGEL-Studentenpisa-Test* [*General knowledge in Germany. Findings from the SPIEGEL students PISA test*] (pp. 71-86). Wiesbaden: VS Verlag.

Ingrisch, M. (1997). *Politisches Wissen, politisches Interesse und politische Handlungsbereitschaft bei Jugendlichen aus den alten und neuen Bundesländern* [*Political knowledge, political interest and political willingness to act among young people from the old and new federal states*].Regensburg: Roderer.

Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern [Audit Bureau of Circulation] (2010). *Quartalsauflagen Printmedien* [*Quarterly circulation print media*]. Available from: http://daten.ivw.eu/index.php

Lautenschlager, G. J., & Meade, A. W. (2008). AlphaTest: A Windows program for tests of hypotheses about coefficient alpha. *Applied Psychological Measurement, 32,* 502-503.

Liepmann, D., & Beauducel, A. (2010). BOWIT - Bochumer Wissenstest [Bochum knowledge test]. *Zeitschrift für Arbeits- und Organisationspsychologie, 54,* 39-45.

Lukesch, H. (1991). Inzidentelles oder systematisches Lernen durch das Fernsehen? Fernsehnutzung und politisches Wissen bei Kindern und Jugendlichen aus Ost und West [Inzidentelles or systematic learning through television? Television use and political knowledge among children and young people from East and West]. *Report Psychologie, 16*, 14-21.

Jäger, A. O., & Fürntratt, E. (1968). *Differentieller-Wissens-Test - DWT* [*Differential Knowledge Test - DWT*]. Göttingen: Hogrefe.

Mickel, W. W. (2005). Politische Bildung in der Europäischen Union [Civic Education in the European Union]. In W. Sander (Ed.), *Handbuch politische Bildung* [*Civic Education Manual*] (3nd edition, pp. 635-651). Schwalbach: Wochenschau Verlag.

Musch, J., Ostapczuk, M., Hilbig, B. E., Auer, T. S., Brandt, M., Cüpper, L., Erdfelder, E., & Undorf, M. (2009). *10-Minuten-Test* [*10-Minutes-Test*]. Unpublished test, University of Duesseldorf.

OECD (2010). *The evolution of news and the internet*. Report of the OECD's Directorate for Science, Technology and Industry. Available from: http://www.oecd.org/sti/ieconomy/45559596.pdf

Oesterreich, D. (2003). Politische Bildung von 14-Jährigen in Deutschland. Übersicht zu Ergebnissen des Civic-Education-Projekts der IEA [Civic Education of 14-year-olds in Germany. Overview of results of the Civic Education Project of the IEA.]. *Unsere Jugend, 55,* 396-401.

Paulhus, D. L., Lysy, D. C., & Yik, M. S. M. (1998). Self-report measures of intelligence: Are they useful as proxy IQ tests. *Journal of Personality, 66*, 525-554.

Popkin, S. L., & Dimock, M. A. (1999). Political knowledge and citizen competence. In S. L. Elkin & K. E. Soltan (Eds.), *Citizen competence and democratic institutions* (pp. 117-146). University Park: Pennsylvania State University Press.

Schmid, C., & Watermann, R. (2010). Demokratische Bildung [democratic education]. In R. Tippelt & B. Schmidt (Eds.), *Handbuch Bildungsforschung* [*Manual Educational Research*] (3nd edition, pp. 881-894). Wiesbaden: VS Verlag.

Schmidt, K.-H., & Metzler, P. (1992). *Wortschatztest* [*Vocabulary Test*]*(WST)*. Weinheim: Beltz.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251.

Trapnell, P. D. (1994). Openness versus intellect: a lexical left turn. *European Journal of Personality, 8,* 273-290.

Trepte, S., & Verbeet, M. (Eds.). (2010). *Allgemeinbildung in Deutschland. Erkenntnisse aus dem SPIEGEL-Studentenpisa-Test* [*General knowledge in Germany. Findings from the SPIEGEL students PISA test*]. Wiesbaden: VS Verlag.

Walther, L., & Hayo, M. (2004). *Teste deine Allgemeinbildung: Politik & Gesellschaft. Begriffe, Daten, Fakten* [*Test your general education: Politics and Society. Concepts, data, facts, 2nd Edition*]. Baden-Baden: Humboldt.

Wanka, R. (2001). *Das Bildungsziel: „Mündiger Bürger“ - Eine reale Utopie?!* [*The educational aim: "mature citizens" - A real utopia!*]. Mühldorf: Selbstverlag

Wilhelm, O., Hülür, G., Köller, O., & Radalewski, M. (2010). Empirische Grundlagen zum Einbürgerungstest [Empirical basics for the naturalization test]. In G. Weißeno (Ed.), *Bürgerrolle heute, Migrationshintergrund und politisches Lernen [Citizen role today,*

immigrant backgrounds, and political learning]. Bonn: Bundeszentrale für politische Bildung.

Zaller, J. (1990). Political awareness, elite opinion leadership, and the mass survey response. *Social Cognition, 8,* 125-153.

Table 1

Demographic Information on the Validation Samples, the PWT Achievements, and the Respective Reliabilities

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6		Sample 7		Sample 8	
						Sample 6a	Sample 6b	t1	t2	t1	t2
Study aim	Cross-validation	Cross-validation	Construct validity	Construct validity	Construct validity	Group difference		Test-retest reliability		Test-retest reliability	
Participants	Adults	Adults	83 pupils, 63 psychology students, 21 adults	Pupils in 10th and 11th grade	Pupils in 11th to 13th grade	Political science students	Psychology students	Pupils in 11th and 12th grade		PhD students	
N	98	342	167	119	176	20	20	30		12	
Gender											
Male	51	141	64	56	83	8	8	13		9	
Female	47	201	103	63	93	12	12	17		3	
Age											
M	40.88	25.43	22.40	16.55	17.42	24.65	24.65	17.73		27.08	
SD	15.78	12.57	9.97	0.89	1.36	2.23	3.59	0.91		1.38	
Range	17-80	15-88	15-61	15-20	15-21	22-30	21-33	17-20		25-30	
PWT											
M	25.98	22.17	15.08	11.92	12.47	25.90	19.20	11.87	13.03	29.17	25.00
SD	4.02	6.02	6.54	4.19	4.17	2.69	4.41	5.76	6.04	1.64	2.34
Range	10-31	4-31	4-31	4-27	2-26	20-30	11-30	5-29	5-30	26-31	21-28
α	.80	.87	.87	.66	.66	.81		.84	.85	.55	.62
r_{tt}								.93[a]		.66[b]	

Note. *N* = sample size, *M* = mean, *SD* = standard deviation, α = Cronbach's alpha, r_{tt} = test-retest reliability, t1 = test date 1, t2 = test date 2,
[a] retest interval: 4 weeks, [b] retest interval: 11 weeks.

Table 2

Validation Constructs and Instruments and their Correlations with PWT Test Scores

	Sample 1		**Sample 2**			**Sample 3**			**Sample 4**			**Sample 5**		**Sample 6**			**Sample 7**	
	M (*SD*)	*r*(PWT)	*M* (*SD*)	*r*(PWT)	α	*M* (*SD*)	*r*(PWT)	α	*M* (*SD*)	*r*(PWT)	α	*M* (*SD*)	*r*(PWT)	*M* (*SD*)	*r*(PWT)	α	*M* (*SD*)	*r*(PWT)
Age	40.88 (15.78)	**.36****	25.43 (12.57)	**.40****		22.40 (9.97)	**.50****		16.55 (0.89)	-.11		16.55 (0.89)	**.18***	24.65 (2.94)	.17		17.73 (0.91)	**.47****
Self-rating of general knowledge	5.09 (0.99)	.13	4.70 (1.23)	**.27****		4.03 (1.11)	**.34****											
Self-rating of political knowledge	4.01 (1.28)	**.51****	4.11 (1.39)	**.42****		3.02 (1.34)	**.48****											
Self-rating of interest in politics	4.63 (1.47)	**.40****	4.73 (1.54)	**.35****		3.75 (1.52)	**.35****		3.74 (1.63)	**.33****		3.95 (1.51)	**.32****	4.73 (1.60)	**.65****		3.20 (1.99)	**.72****
Self-rating of participation in political discussions	4.13 (1.26)	**.40****	4.36 (1.38)	**.22****		3.48 (1.23)	**.27****		3.43 (1.33)	**.26****		4.00 (1.28)	**.20****	4.40 (1.49)	**.59****		3.13 (1.55)	**.50****
Frequency of reading political newspapers and magazines	2.22 (0.86)	**.34****	2.27 (0.95)	**.18****		1.91 (0.89)	**.16***		1.82 (0.80)	**.28****		2.12 (0.82)	**.16***	2.06 (0.77)	**.61***			
Frequency of viewing TV with political content	3.35 (3.29)	**.36****	3.65 (4.23)	**.12***		2.62 (3.50)	.05											
Frequency of using the Internet to obtain political information	2.95 (3.67)	.20	3.84 (5.81)	**.24****		1.65 (3.46)	**.27****											
BOWIT[a] political knowledge items (18 items)					.68	8.40 (3.32)	**.77****				.41	7.05 (2.91)	**.48****					
DWT[b] political knowledge items (17 items)					.65	6.02 (3.05)	**.71****											
SPIEGEL students PISA Test[c] / political knowledge items (18 items)								.69	5.03 (3.04)	**.59****								

	Sample 1		Sample 2		Sample 3			Sample 4			Sample 5		Sample 6		Sample 7		
	M (*SD*)	*r*(PWT)	*M* (*SD*)	*r*(PWT)	α	*M* (*SD*)	*r*(PWT)	α	*M* (*SD*)	*r*(PWT)	α	*M* (*SD*)	*r*(PWT)	*M* (*SD*)	*r*(PWT)	α	*M* (*SD*)	*r*(PWT)
BOWIT[a] subscale mathematics/physics (14 items)					.52	4.76 (2.54)	**.38****											
BOWIT[a] subscale biology/chemistry (14 items)					.57	5.10 (2.44)	**.32****											
Vocabulary Test[d] (42 items)								.89	26.35 (7.10)	**.31****								
10-Minutes-Test[e] (32 items)											.67	15.28 (3.56)	**.24****					
Trapnell Smart Scale[f] (4 items)								.75	20.88 (5.91)	**.36****	.85	20.57 (6.25)	.02			.77	20.30 (5.65)	**.48****
Grade in social or political sciences						2.52 (1.01)	**-.46****		2.93 (0.88)	-.17		2.71 (0.93)	-.15	2.06 (0.66)	-.37		2.58 (1.02)	**-.52****
Grade in German						2.25 (1.03)	**-.28****		2.73 (0.71)	-.17		2.64 (0.90)	-.13	1.95 (0.93)	.04		2.75 (0.68)	-.21
Grade in English						2.48 (1.01)	**-.30****		2.85 (0.90)	-.17		2.84 (0.90)	-.11	2.02 (0.95)	.17		2.91 (0.83)	-.28
Grade in mathematics						2.56 (1.14)	**-.23****		2.96 (1.06)	-.08		2.93 (1.13)	-.02	2.28 (1.01)	**.32***		3.03 (1.00)	-.22
Grade in biology						2.34 (0.96)	-.09											
Grade in chemistry						2.03 (0.69)	-.11											
Grade in physics						2.95 (0.94)	-.16											

Note. α = Cronbach's alpha. Cronbach's α could not be computed for grades and for self-assessments based on a single item. In Sample 6, correlations with external criteria were determined at t1. Due to some missing data, sample sizes ranged from 139 to 167 for Sample 3, from 117 to 119 for Sample 4, from 168 to 176 for Sample 5. Significant correlations are written in bold ($p < .05$).

[a] Hossiep & Schulte, 2008; [b] Jäger & Fürntratt, 1968; [c] Trepte & Verbeet, 2010; [d] Schmidt & Metzler, 1992; [e] Musch et al., 2009; [f] Trapnell, 1994.

* $p < .05$. ** $p < .01$.

What is the Maastricht Treaty?

a) A European treaty on subsidies in agriculture

b) A treaty for the realization of an economic and monetary union in Europe

c) An international treaty regulating the position of Germany in the Western community of states

d) A contract for the abolition of border controls within Europe

Figure 1. Example item from the PWT. The correct answer option is written in bold.