On how testwiseness and acceptance reluctance influence the validity of sequential knowledge tests

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Martin Papenberg
aus Düren

Düsseldorf, Juni 2018

aus dem Institut für Experimentelle Psychologie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:
1. Prof. Dr. Jochen Musch
2. Prof. Dr. Axel Buchner

Tag der mündlichen Prüfung: 10.07.2018

Contents

Abstract

Multiple-choice tests are widely used to assess cognitive ability, scholastic achievement and knowledge. A multiple-choice (MC) item consists of a question stem that is shown along with the correct answer and one or several distractors. From among these response options, test-takers have to identify the solution. Test scores derived from MC tests have been shown to satisfy high psychometric standards if guidelines for good item writing are followed. However, investigations of MC tests in practice revealed a high prevalence of flawed item writing. By comparing the response options that are shown simultaneously, testwise test-takers can capitalize on such flaws to increase their test scores, making MC tests biased against less testwise test-takers. As a consequence, MC test scores may contain construct-irrelevant variance due to individual differences in testwiseness. Discrete-option multiple-choice (DOMC) testing has been proposed as a means to reduce such construct-irrelevant variance. In DOMC items, the question stem is presented together with only one response option at a time; additional options are shown only after the correctness of the present option has been assessed in a yes/no decision. Thus, response options are presented sequentially rather than simultaneously, precluding the possibility to compare the relative plausibility of all alternatives. Previous investigations indicated that DOMC testing makes items more difficult, reduces the number of response options that have to be shown, and reduces testing time as compared to traditional MC testing. The present thesis expands the yet small body of research on the DOMC test format. First, it is shown that DOMC is capable of reducing the effectiveness of testwiseness strategies, as indicated by the observation that the gap between DOMC and MC test scores increases as the level of testwiseness increases. Next, by experimentally establishing different levels of knowledge, we find that MC and DOMC test scores are

equally capable of representing known ability levels, suggesting that MC and DOMC test scores are equally valid. In exploratory analyses, we also find that DOMC test-takers very rarely select the correct answer when they have no knowledge. This gives surprising evidence to the notion that DOMC testing reduces test-takers' chances to randomly guess a solution. The remaining part of this thesis proposes and investigates a new model of response behavior in sequential tests (MORBIST). On the basis of signal detection theory, MORBIST describes the processes that lead to DOMC item responses. Simulations based on MORBIST predict that DOMC test scores do not only depend on test-takers' ability, but also on their knowledge-independent individual preference to choose late rather than early response options in DOMC items. We first demonstrate the existence of such individual differences in *acceptance reluctance* in a correlational investigation. Acceptance reluctance is thereby identified as an important cognitive trait in the domain of sequential knowledge tests. Consistent with MORBIST's prediction, the results of the correlational study also show that higher acceptance reluctance is related to better test scores in a DOMC knowledge test. Additional evidence of MORBIST's usefulness is obtained by reproducing several known properties of the DOMC test format in a computer simulation. In a concluding experimental study, we establish low versus high acceptance reluctance using a payoff manipulation. The results confirm that high acceptance reluctance causally leads to increased DOMC test scores. Another computer simulation also shows that MORBIST can adequately reproduce the observed test score gap between high and low acceptance reluctance. Taken together, the present research shows first evidence to the notion that DOMC test scores contain some construct-irrelevant variance. To obtain an integrated judgment on the viability of DOMC testing, this drawback has to be weighed up against the advantages of the DOMC test format that include a reduced susceptibility to testwiseness and random guessing.

Zusammenfassung

Zur Erfassung von Wissen, kognitiver Fähigkeit und schulischer Leistung werden häufig Multiple-Choice-Tests verwendet. Eine Multiple-Choice (MC) Aufgabe besteht aus einem Fragestamm, der gemeinsam mit der richtigen Antwort und mindestens einem Distraktor vorgegeben wird. Aus den verschiedenen Antwortoptionen sollen Testnehmer die korrekte Lösung auswählen. Die psychometrischen Eigenschaften von MC-Tests genügen hohen Ansprüchen, wenn anerkannte Richtlinien zum Erstellung der Testaufgaben befolgt werden. Empirische Untersuchungen zeigen jedoch, dass in der Praxis vorkommende Aufgaben oft Mängel aufweisen. Durch den Vergleich aller Antwortoptionen können testschlaue Personen solche Mängel ausnutzen, um ihren Testwert zu verbessern. Aus diesem Grund können individuelle Unterschiede in Testschläue konstrukt-irrelevante Varianz in MC-Punktzahlen erzeugen, was deren Validität gefährdet. Das Discrete-Option Multiple-Choice (DOMC) Verfahren wurde vorgeschlagen, um diese Art konstrukt-irrelevanter Varianz zu reduzieren. In DOMC-Aufgaben wird der Fragestamm nur mit einer einzigen Option vorgegeben; weitere Antwortoptionen folgen erst, nachdem die Korrektheit der vorherigen Optionen bewertet wurde. Die bisherige Forschung konnte zeigen, dass DOMC-Aufgaben im Vergleich zu traditionellen MC-Aufgaben schwieriger sind, die Präsentation von weniger Antwortoptionen erfordern und in kürzerer Zeit bearbeitet werden können. Die vorliegende Dissertation berichtet weiterführende Untersuchungen zu den psychometrischen Eigenschaften des DOMC-Antwortformats. Eine erste Untersuchung zeigt, dass die sequentielle Vorgabe der Antwortoptionen Testnehmer wirksam daran hindert, nur auf der Basis ihrer Testschläue die richtige Lösung zu identifizieren. Dies belegt der Umstand, dass die Differenz von MC- und DOMC-Punktzahlen umso größer ausfällt, je testschlauer Testteilnehmer sind. Danach wird mithilfe einer experimentellen

Manipulation des Wissens von Studienteilnehmern gezeigt, dass MC- und DOMC-Punktzahlen bekannte Wissenszustände vergleichbar valide abzubilden vermögen. In einer explorativen Analyse wird außerdem festgestellt, dass DOMC-Testnehmer nur selten zufällig die korrekte Lösung auswählen, wenn sie über kein Wissen verfügen. Der darauffolgende Teil untersucht ein hier neu vorgeschlagenes Modell des Antwortverhaltens in sequentiellen Tests (MORBIST: a model of response behavior in sequential tests). MORBIST beschreibt mithilfe der Signalentdeckungstheorie die Prozesse, die beobachtbaren Antworten in DOMC-Aufgaben vorausgehen. MORBIST sagt vorher, dass das Abschneiden in DOMC-Tests nicht nur von der Fähigkeit der Testnehmer abhängt, sondern auch von ihrer wissensunabhängigen Neigung, eher späte als frühe DOMC-Antwortalternativen zu akzeptieren. Die Existenz derartiger individueller Unterschiede im *Festlegungszögern*, die in der vorliegenden Arbeit erstmals als für die sequentielle Wissenstestung bedeutsamer kognitiver Trait belegt werden, wird in einer ersten korrelativen Studie gezeigt. In Übereinstimmung mit der Vorhersage von MORBIST zeigt sich darin, dass hohes Festlegungszögern mit höheren DOMC-Punktzahlen einhergeht als niedriges Festlegungszögern. Eine Computersimulation vermag weitere empirische Beobachtungen der korrelativen Untersuchung erfolgreich zu reproduzieren und erbringt dadurch weitere Belege für die Bedeutung des Festlegungszögerns und für die Nützlichkeit von MORBIST. Eine abschließende Studie erzeugt hohes und niedriges Festlegungszögern experimentell durch eine Manipulation der Auszahlungsmatrix. Die Ergebnisse bestätigen, dass hohes Festlegungszögern kausal zu höheren Punktzahlen führt als niedriges Festlegungszögern. In einer weiteren Simulation kann MORBIST die Punktzahldifferenz zwischen Teilnehmern mit hohem und niedrigem Festlegungszögern adäquat reproduzieren. Insgesamt belegen die durchgeführten Untersuchungen, dass DOMC-Testwerte mit konstrukt-irrelevanter Varianz kontaminiert sind. Um die Brauchbarkeit des DOMC-Verfahrens abschließend zu beurteilen, muss

dieser Nachteil gegen die empirisch belegten Vorteile des Verfahrens abgewogen werden, zu denen eine bessere Kontrolle von Testschläue und Rateprozessen gehört.

1 Scientific background

> [...] a century of experience with test construction and analysis clearly shows that it is very hard to find out where the scores are coming from if tests are not constructed on the basis of a theory of item response processes in the first place. [...] No table of correlations, no matter how big, can be a substitute for knowledge of the processes that lead to item responses. (Borsboom, Mellenbergh, & van Heerden, 2004, p. 1067f)

1.1 Multiple-choice testing

Multiple-choice (MC) testing is widely used to assess knowledge, scholastic achievement and cognitive ability. In its common form, an MC item consists of a question stem and a set of response options. One of the response options is the correct answer that needs to be identified by the test-takers. Incorrect options are called distractors. This version of MC testing is sometimes called "single-choice" testing to highlight a distinction from response formats that may include multiple correct answers (Dressel & Schmid, 1953; Kubinger, Holocher-Ertl, Reif, Hohensinn, & Frebort, 2010; Štěpánek & Šimková, 2013). Whereas there is always only one solution in MC items, the number of distractors may vary between items. Traditionally, measurement textbooks advised test creators to write at least three or four distractors to reduce the chance of randomly guessing the correct solution (Owen & Froman, 1987). However, researchers now generally agree that creating two distractors in addition to the solution is usually sufficient. Three options are comparably easy to write and empirical investigations have found that this number offers a sufficiently high test quality (Edwards, Arthur, & Bruce, 2012; Haladyna & Downing, 1993; Rodriguez, 2005). Even only two response options may suffice if the one accompanying distractor is of high quality (Papenberg & Musch, 2017).

The application of MC tests is motivated by pragmatic considerations. In particular, MC items are scored easily by comparing a respondent's answer to the *key* – i.e., the correct solution – that is already determined during the item writing process. Therefore, MC items can be scored efficiently, objectively and even in an automized manner (Lindner, Strobel, & Köller, 2015; Tamir, 1991). This efficiency is especially important when a large number of students is to be tested. MC testing scales well with an increasing size of examinees. In this sense, MC testing stands in stark contrast to other classical testing formats such as essays or oral examinations that require more effort in scoring. Such test formats also necessarily rely on a more subjective assessment of student achievement (Malouff, Emmerton, & Schutte, 2013). In contrast, MC testing, by design, offers the possibility to score test items objectively. Arguably, objectivity in scoring can be regarded as the major strength of MC testing (Haladyna, 2004). In particular, objective scoring procedures suppress conscious and unconscious biases by human scorers who may, for example, fall prey to the "halo effect" (E. L. Thorndike, 1920) when assessing a student's achievement. Malouff et al. (2013) showed that a written essay received lower scores when graders were led to believe the writer had previously given a bad oral presentation, as opposed to a good oral presentation. This effect had a considerable size of $d = 0.53$ even though the same performance was rated in both cases (also see Malouff & Thorsteinsson, 2016). In contrast, objective scoring based on applying the same standards for everyone is a precondition for fair testing.

A long line of research also shows that test scores derived from MC tests warrant reliable and valid conclusions on the knowledge and ability levels of test-takers (Downing, 2006). The key to a high psychometric quality is good item writing. During the last decades, a vast amount of research has resulted in many item writing guidelines, that today's test creators can rely on to write MC test questions of high

quality (Downing, 2002; Ebel, 1971; Farley, 1989; Haladyna, 2004; Haladyna, Downing, & Rodriguez, 2002; Lindner et al., 2015; Siroky & Di Leonardi, 2015). Unfortunately however, MC items in practice often do not follow these good item writing guidelines. A large number of investigations into the quality of practically applied MC tests have revealed an unsettling proportion of flawed items (Brozo, Schmelzer, & Spires, 1984; Downing, 2002; Hijji, 2017; Hughes, Salvia, & Bott, 1991; Jozefowicz et al., 2002; Metfessel & Sax, 1958; Rogers & Bateson, 1991; Tarrant & Ware, 2008; Tarrant, Knierim, Hayes, & Ware, 2006; Tomkowicz & Rogers, 2005; Willing, Ostapczuk, & Musch, 2015). This contrast between practical experience and theoretical guidance may explain why MC testing sometimes has a bad reputation among scholars and laymen alike (Frederiksen, 1984). It also highlights a flip side to the efficiency of MC testing: whereas administering and scoring MC tests can be done efficiently, the creation of high-quality items is a time-consuming and challenging task (Farley, 1989). In particular, the creation of functioning distractors poses difficulties to many test creators and teachers (Haladyna & Downing, 1993; Lee & Winke, 2013).

Poor item writing reduces the validity of MC test scores by introducing construct-irrelevant variance to the measurement (Downing, 2005). As opposed to unsystematic random variation in test scores due to unreliability, construct-irrelevant variance is a systematic distortion of test scores that will consistently put some test-takers at disadvantage, leading to unfair testing (Haladyna & Downing, 2004; Messick, 1989). As an example, Downing (2005) reported that flawed MC items were associated with lower pass-rates on medical school tests. Thus, some students' test scores were unfairly impaired as a result of poor item writing rather than their own lack of knowledge.

1.2 Testwiseness

If some students capitalize on item flaws more effectively than others, they obtain an unfair advantage over their less cunning peers. The clever exploitation of superficial flaws in MC item writing has been discussed under the umbrella term of *testwiseness*. Testwiseness is conceptually independent from the attribute under investigation[1] and has been identified as the most prominent source of construct-irrelevant variance in MC test scores (Foster & Miller, 2009; Gibb, 1964; Millman, Bishop, & Ebel, 1965; Sarnacki, 1979; Thoma & Köller, 2018). Millman et al. (1965) proposed a comprehensive taxonomy of testwiseness that still provides an important frame of reference for this construct. Their work encompasses a wide spectrum of test-taking strategies that also include general aspects like efficient time usage in testing situations. However, most aspects of testwiseness are specific to MC testing and concern the ability to recognize superficial cues to the solution that test creators unwillingly introduce when writing items (Sarnacki, 1979). Starting with Gibb (1964), researchers interested in measuring individual differences have therefore highlighted the ability to use item cues as the defining feature of testwiseness (e.g., Diamond & Evans, 1972; Thoma & Köller, 2018). Testwise students are able to recognize such unintended cues to artificially increase their test scores. If test-takers use cues to identify the solution, their choice of the correct answer does not reveal their factual knowledge, but only their testwiseness. Individual differences in testwiseness may therefore lead to construct-irrelevant variance in MC test scores (Allan, 1992).

Often, cues arise as a consequence of idiosyncrasies in how item writers phrase the question stem and the response options. For example, item writers are often tempted to elaborate the solution more carefully than the accompanying distractors

[1] It has however been argued that some partial knowledge of test content helps to better exploit testwiseness cues (Rogers & Yang, 1996).

because they find it difficult to formulate a concise solution. This often leads to a solution that is visibly longer than the distractors, or more specific in meaning (Sarnacki, 1979). Another problem is that the solution is sometimes more general in meaning because all distractors include specific determiners such as “never”, “all”, or “exclusively”. Such overqualifing determiners easily falsify a statement and are therefore often used in distractors, but less often in solutions (Slakter, Koehler, & Hampton, 1970). With regard to response alternatives that can be sorted in a numeric order, it has been shown that test writers tend to key one of the middle values (Brozo et al., 1984). By selecting one of these middle values, even test-takers who are completely unknowing – but testwise – may increase their chances of selecting 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. It has also been found that when two response options are directly opposite in meaning, one of them is often correct. This direct opposites cue occurs rather frequently in real examinations, presumably because creating a distractor that simply reverses the correct response option lightens the work load of a test writer. In the investigation by Brozo et al. (1984), this cue was the most prevalent of nine testwiseness cues they identified in American college examinations; 12.4% of all MC items contained two response options that directly opposed each other.

The use of most testwiseness cues is made possible by the direct comparison of the response options that are shown at the same time (Willing et al., 2015). For example, the longest alternative can only be recognized when the length of all response options can be compared. The effectiveness of such meta-cognitive, comparative strategies is therefore a drawback of the MC format that enables the simultaneous comparison of the different response options. As test creators are not interested in measuring how

effectively test-takers exploit item cues – but in assessing actual knowledge –, they may wish to employ a test format that prevents the direct comparison of all response options.

1.3 Discrete-option multiple-choice (DOMC)

In a study of what they called discrete-option multiple-choice (DOMC) testing, Foster and Miller (2009) discussed that a sequential presentation of response options might prevent the usage of testwiseness cues. In DOMC items, response options are presented one after another in random order. Therefore, test-takers cannot directly compare the plausibility of the different response options using a relative judgment. Instead, the correctness of each answer has to be evaluated in a separate yes/no decision using an absolute judgment. For practical reasons, the DOMC item type can only be administered using a computer (Kubinger, 2009). Figure 1 illustrates a DOMC item.

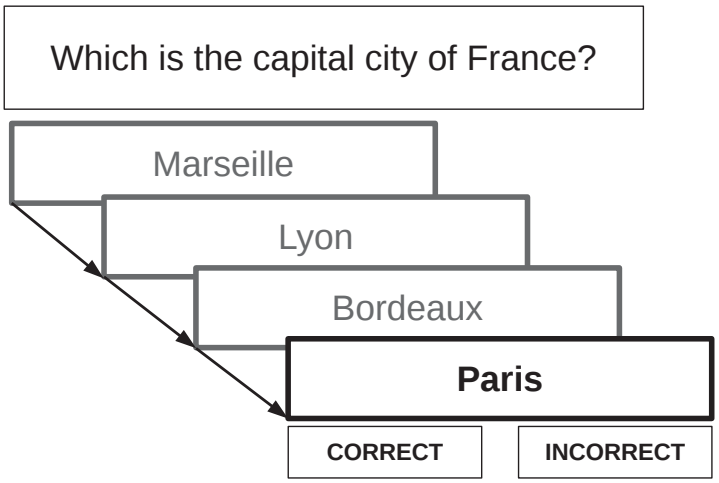

Figure 1. Illustrative example of a DOMC item in which the solution ("Paris") is presented as the fourth response option. The solution is shown only if all of the three distractor options ("Marseille", "Lyon", and "Bordeaux") have been rejected. A point is awarded only if all distractors that have been presented prior to the solution are rejected, and if the solution is accepted when it is shown.

The basic elements characterizing the MC item type remain in DOMC testing. That is, DOMC items also contain a question stem and a fixed number of response options, one of which is the solution. DOMC items are, however, usually answered before all response 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; (b) the solution has incorrectly been rejected, or (c) a distractor has incorrectly been accepted. In each of these cases, the correctness of the response is determined, rendering the presentation of additional options unnecessary. Instead, the next DOMC item is presented. Due to the random order of the response options and the application of stopping rules, different test-takers are presented with different subsets of all possible response options. Test-takers only see all alternatives if the solution is the last option and if they are willing to reject all distractors shown before.

That fewer response options have to be shown helps to reduce testing time because test-takers have to read less item text. Foster and Miller (2009) observed that compared to parallel MC items, DOMC items led to a reduction in testing time of about 10%. Willing et al. (2015) even observed a reduction in testing time as large as 30%. Foster and Miller (2009) also argued that when test-takers are presented with fewer response options, test security is enhanced. If test-takers are never presented with a response option in the first place, they cannot give it away to participants of future examinations.

The DOMC test format has consistently been shown to increase item difficulties in comparison to traditional MC testing (Foster & Miller, 2009; Kingston, Tiemann, Miller, & Foster, 2012; Willing, 2013; Willing et al., 2015). A likely reason for this observation is that DOMC test-takers have to base their decisions on only a subset of the information that is available to MC test-takers. Moreover, performance in yes/no

tasks such as DOMC items has long been known to be worse than performance in forced-choice tasks such as MC items – the comparative availability of the distractors facilitates the identification of the target (Jang, Wixted, & Huber, 2009). However, as long the test score range allows to discriminate between test-takers of low and high ability, item difficulty in itself is not an indicator of psychometric quality. The more important criteria of reliability and validity also have to be evaluated in order to allow judgments on the comparative viability of MC and DOMC testing.

When Foster and Miller (2009) first investigated the DOMC response format, they observed that parallel MC tests and DOMC tests showed comparable psychometric functioning with regard to item discrimination and thus, internal consistency. However, they argued that DOMC tests offer an important psychometric advantage over MC tests by preventing testwiseness strategies. Thus, they expected that DOMC testing reduces construct-irrelevant variance due to individual differences in testwiseness, thereby increasing test score validity. They surmised that DOMC tests better prevent usability of testwiseness cues because their usage usually relies on the simultaneous availability of all response options, that is precluded by the sequential presentation of response options in DOMC testing. Whereas Foster and Miller (2009) only provided a logical argument for their claim, Willing et al. (2015) found evidence for the notion that DOMC reduces cue usability by examining test items used in a continuing medical education test. Of the ten items under investigation, eight items contained cues to their solution. It was shown that cued items were easier than items that did not contain cues. Furthermore, items in DOMC format were more difficult than items in MC format. This increased difficulty was however only observed for the items containing cues, which was evident from an interaction of cue occurrence and test format on item difficulty. Hence, it seemed that cue usage was hindered by the sequential presentation of response options. This notion however relied on a quasi-experimental design as

cue availability varied between items of different content. Therefore, cue availability and item content were confounded, and orthogonal experimental manipulations of cue availability and item content were still needed. The present thesis therefore reports an experimental study that manipulates cue usability independently from item content to unambiguously investigate whether DOMC testing reduces cue usability (see Appendix: Original Research Articles, Article 1).

Two other investigations also examined the psychometric functioning of parallel MC and DOMC tests (Kingston et al., 2012; Willing, 2013). In these studies, estimates of test score reliability, item discrimination and criterion-related validity indicated that test quality was comparable for both formats. Therefore, even though DOMC testing may prevent the usability of testwiseness cues, direct investigations of psychometric quality indicate that DOMC test scores are no more valid than MC test scores. It should be noted, however, that all of the previous comparisons of reliability and validity of MC and DOMC test scores relied on correlational study designs. To investigate test validity, correlational validation procedures approximate the attribute under investigation by presenting a test that serves as a criterion with which the to-be-validated test is correlated. This procedure yields an estimate of "criterion-related validity" (Newton & Shaw, 2014). For example, an intelligence test may be validated using an older version of the same test, which is in fact a standard approach. Correlational studies however suffer from interpretational problems that preclude strong conclusions. Borsboom et al. (2004) even surmised that any correlational conception of validity might be "hopeless". That is because correlational validation studies evoke at least two problems. First, validation must start at one place – how do we know that the criterion itself is a valid indicator of the attribute that is intended to be measured (cf. Wechsler, 1944)? Moreover, correlations between a test and a criterion do not necessarily indicate the degree to which a test actually measures

a construct because correlations capture all systematic variance that is shared by two measures. Such shared variance may be due to factors other than the construct under investigation if both the test and the criterion are contaminated with the same construct-irrelevant variance, such as test-takers' anxiety, motivation, testwiseness or propensity to take risks (e.g., Diamond & Evans, 1972; Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011; Richardson & Norgate, 2015; Rowley & Traub, 1977). Disentangling the influence of different, possibly intercorrelated factors is a difficult – if not impossible – task in correlational research (Westfall & Yarkoni, 2016). The observation that MC and DOMC tests scores exhibit the same correlation with some external criterion therefore does not provide compelling evidence that the two test formats have comparable validity.

A promising way to overcome the problems associated with correlational validation is to employ an experimental approach (Diedenhofen & Musch, 2017; Erdfelder & Musch, 2006; cf. Poizner, Nicewander, & Gettys, 1978). In experimental validation studies, test-takers' ability levels are known – and not confounded by any other variables –, because the information available to test-takers is manipulated experimentally (cf. Poizner et al., 1978). The accuracy with which test scores reflect the test-takers' experimentally manipulated level of information therefore provides a more reliable index of test validity than a mere correlation with some related test or construct. Due to the advantages associated with experimental approaches, an experimental validation of DOMC test scores is reported in the present thesis (see Appendix: Original Research Articles, Article 2).

1.4 Response sets

DOMC testing might be capable of reducing construct-irrelevant variance due to testwiseness. DOMC however is associated with a change in response format and

it is therefore an open question whether DOMC testing introduces other method-specific variance. This question is addressed for the first time in the present thesis. DOMC test scores may contain their own share of construct-irrelevant variance if there are additional factors beyond ability that affect test-takers' item responses. This notion is not implausible: DOMC item responses are given in an uncertain situation characterized by incomplete information. Uncertain and ambiguous testing situations tend to invoke individual differences in *response sets* if a test format allows test-takers to make use of qualitatively different response categories, such as responding "yes" or "no" in a true-false test (Berg, 1955; Cronbach, 1946).

Response sets arise when test-takers systematically differ in the degree to which they make use of the available response categories. They can be thought of as an influence of personality on test responses that operates above and beyond of what is to be expected based on the actual trait under investigation. The observation that personality variables affect test responses has been made early (Swineford, 1938, 1941; R. L. Thorndike, 1938; Wiley & Trimble, 1936). In more recent years, this area of research has also been given attention (e.g., Kantner & Lindsay, 2012; Wetzel, Lüdtke, Zettler, & Böhnke, 2016; Ziegler, 2015). The ongoing research interest is most likely explained by the variety of response formats in which response sets occur, as has first been systematized by Cronbach (1946). Cronbach also concluded that the presence of response sets threatens test validity. As the most prominent example of a response set, acquiescence is the tendency "to agree with test items, regardless of their content" (McGee, 1962, p. 229). That is, a preference to say "yes" when faced with a question that requires a yes/no decision. In achievement testing, acquiescent persons may more often agree with true-false statements when in doubt about their response, whereas less acquiescent persons might more often disagree—even if they do not differ in ability. Individual differences in acquiescence can introduce construct-irrelevant variance to

test scores if a preference for responding “no” or “yes” systematically influences test scores. This might happen if the test creator tends to present more false or more true statements on the test, leading to biased test results as a consequence of individual differences in response set rather than ability.

A similar kind of response set can be investigated in the domain of recognition memory where old and new words have to be distinguished. In such tasks, signal detection theory can be used to distinguish ability from bias, i.e., the tendency to conclude that an item is known from a previous learning phase rather than calling it “new”. Kantner and Lindsay (2012) recently investigated the stability of such response biases in old/new recognition memory tasks. In several experiments, they observed high correlations between measures of response bias across experimental sessions. They therefore concluded that “some people require more memory evidence than do others before they are willing to call an item ‘old’ (p. 1163)”. Due to the stability of individual differences in response bias, Kantner and Lindsay (2012) also coined the term “cognitive trait” to describe response biases that reliably and systematically influence observed behavior in any kind of cognitive tasks. They used this label to distinguish biases from more general personality traits like extraversion and agreeableness, even though they did not claim that cognitive traits have to be regarded as independent concepts. For the present thesis, the concepts of response set and cognitive trait will both be regarded as influences of personality on test responses, and no particular distinction will be made between the two. As compared to the term response set, the label cognitive trait may place a special focus on the reliability and stability of individual differences, highlighting a trait-like character.

As compared to testwiseness, response sets constitute a different kind of unwanted influence on test responses. Testwiseness is based on a deliberate analysis of the superficial properties of MC items (Rogers & Yang, 1996). In contrast, response

sets are based on an inherent personal preference for choosing a certain response type (Sarnacki, 1979). For example, a testwise examinee may select "yes" more often in a true-false test, but not because of a personal preference for responding "yes", but because she or he knows that the test creator tends to use more true than false statements. Whereas testwiseness is seen as a threat to the validity of MC test scores, it has been argued that MC testing is comparably free of response sets (Cronbach, 1950). That is because in MC testing, there are no different types of responses – the selection of each option belongs to the same qualitative response category. Note that MC tests may allow for a response set when a penalty for false answers is included in the scoring scheme (Sherriffs & Boomer, 1954). To reduce the success rate of random guesses in MC tests – that is often considered to be a major disadvantage of MC testing – test administrators often apply negative scores when test-takers select a distractor. If such penalty scoring is applied and test-takers are uncertain about the correct solution, they have to choose – based on their personal propensity to take risks – if they nevertheless select one of the options or if they omit the response. Unfortunately – with most scoring schemes –, the dominant strategy is to always select an answer even if penalty scores are applied (cf. Dirkzwager, 1996; Frary, 1988; Rowley & Traub, 1977). Risk averse test-takers are thereby unfairly put at a disadvantage. In particular, female test-takers are discriminated against because they tend to omit items more often (Baldiga, 2013; Ben-Shakhar & Sinai, 1991). Another possibility to allow for a response set in MC testing is to include response options such as "none of the above is true" or "all of the above is true" (Frary, 1991; Haladyna, 2004). In this case, test-takers may differ in their willingness to use one of these "catch-all" options. Maybe for this reason, test writers are generally discouraged to use such options in MC tests (Haladyna & Downing, 1989).

In the standard version of MC testing discussed here, the only plausible response

set is one of positional bias. Test-takers may prefer to select response options in dependence of their spatial location. For example, they might prefer one of the middle options (Attali & Bar-Hillel, 2003; Wevrick, 1962). Research shows that test-takers seem to have a small preference for selecting an option that is shown at the top of the list of alternatives. This leads to slightly increased item difficulties when the key is one of the later options (Hohensinn & Baghaei, 2017; Tellinghuisen & Sulikowski, 2008). However, the evidence for individual differences in positional bias is weak; test-takers do not seem to vary substantially in their propensity to select options as a function of their spatial position (Cronbach, 1950). Furthermore, test creators seem to take care to place the correct solution evenly among the possible positions (Brozo et al., 1984). This suggests that even if individual differences in positional bias existed, they would not introduce substantial variance to test scores.

Whereas DOMC testing may reduce construct-irrelevant variance due to test-wiseness, it is well possible that it introduces other construct-irrelevant variance as a consequence of a positional bias. In DOMC testing, positional bias would not be concerned with spatial location, but with the sequential presentation order of the response options. Some test-takers may prefer to select response options that are presented early, whereas others may prefer late positions. Differences in positional preference may be fueled by the uncertainty that is inherent to DOMC testing: even though test-takers know that one of the response options will eventually be the correct solution, they are clueless with regards to *when* the correct answer will be shown. In this ambiguous situation, some test-takers may feel rushed to accept a plausible response option early, while other respondents may be inclined to wait longer before they accept an answer. No previous study has investigated whether DOMC item responses are affected by response sets. The present thesis reports first evidence of the existence of a positional response set that will be called *acceptance reluctance* (see

Appendix: Original Research Articles, Article 3). Acceptance reluctance influences test-takers proclivity to accept early DOMC response options and is identified as a potentially important cognitive trait in the domain of sequential knowledge tests. If individual differences in knowledge-independent acceptance reluctance affect DOMC test scores, DOMC test scores would be contaminated with construct-irrelevant variance. This would raise serious questions regarding the viability of DOMC testing as an alternative to MC testing.

2 Summary of new contributions

To extend the yet small body of research on the viability of DOMC testing, the present thesis presents four empirical studies and a theoretical model of the processes that lead to item responses in DOMC tests. The two major aspects of validity in psychological assessment are investigated in the domain of DOMC testing (cf. Messick, 1989). That is: (a) how well do DOMC test scores capture the relevant *signal*, i.e., how well do they represent the construct under investigation, and (b) how well do DOMC test scores suppress *noise*, i.e., to what degree are they susceptible or robust to incorporating construct-irrelevant variance. The studies reported here do not employ a traditional correlational approach to validation – correlating test scores with some external criteria (Newton & Shaw, 2014) –, but instead focus on the *causal* processes that lead to DOMC item responses. Thus, recent recommendations on the validation of testing procedures were followed (e.g. Borsboom et al., 2004; Embretson, 2007; Lissitz & Samuelsen, 2007). A particular emphasis was given to experimental methods that traditionally have been underutilized in research on psychological assessment (Erdfelder & Musch, 2006). Experimental methods are however necessary to investigate the causal influence of attributes that are hypothesized to affect test achievement (cf. Borsboom et al., 2004; Messick, 1993). In particular, they are needed to analyze the extent to which construct-irrelevant factors – such as testwiseness or response sets – affect test scores, because these unwanted factors may be correlated with the construct under investigation (Rowley & Traub, 1977). For example, more testwise test-takers may also be more intelligent, making it difficult to assess the unique contribution of testwiseness to test scores in purely correlational research (Diamond & Evans, 1972; Scruggs & Lifson, 1985; cf. Westfall & Yarkoni, 2016).

In a first study, we extend recent findings showing that DOMC is capable of

preventing the usability of testwiseness cues (see Appendix: Original Research Articles, Article 1). By varying test-takers' testwiseness experimentally – thereby improving the design over the previous quasi-experimental designs –, it is shown that DOMC testing reduces cue usability particularly well when test-takers know about their presence and actively search for them. Thus, DOMC testing reduces the test score gap between more and less testwise test-takers and therefore has the potential to reduce construct-irrelevant variance due to individual differences in testwiseness. As a secondary finding – supporting the notion that DOMC prevents the usability of testwiseness cues – it is shown that the reduction in cue usage is moderated by cue validity: the more reliably cues hint towards the solution, the better will DOMC prevent their usage.

Second, an experimental procedure is employed to compare the validity of DOMC and MC test scores (see Appendix: Original Research Articles, Article 2). By experimentally inducing different levels of information, the ability levels of test-takers are known and serve as a strong validation criterion. Based on a comparison of the accuracy with which test scores predict information levels, a Bayes factor gives strong evidence to the notion that MC and DOMC test scores are equally valid. Furthermore, we find no difference in internal consistency as measured by Cronbach's α (Cronbach, 1951). By closely exploring the response behavior of completely uninformed test-takers, we furthermore surprisingly find that DOMC testing reduces the chance to luckily guess the correct solution in comparison to MC testing. When DOMC test-takers have no knowledge, they only choose the correct answer in 5% of all cases, as compared to a success rate of 17% in MC format. When information is provided, test-takers choose the correct solution in 87% of all MC item presentations and in 80% of all DOMC item presentations. Thus, the gap between MC and DOMC test scores is even larger when test-takers are uninformed, indicating that DOMC testing makes it

particularly hard for uninformed test-takers to randomly guess a solution. Given that the susceptibility to random guessing is often considered to be the major weakness of MC testing (Kubinger et al., 2010), this result may encourage researchers and practitioners to try out DOMC testing as an alternative procedure.

The remaining part of this thesis investigates a newly developed model of response behavior in sequential tests (*MORBIST*; see Appendix: Original Research Articles, Article 3). We propose MORBIST as the first formalization of the processes that lead to DOMC item responses. The mathematical basis of MORBIST is the Gaussian model of signal detection, which is a standard model of decision making in many areas of psychological research (Kellen, Klauer, & Singmann, 2012; Macmillan & Creelman, 2005; Pastore, Crawley, Berens, & Skelly, 2003). MORBIST assumes that DOMC item responses are not only determined by the ability of test-takers, but also by their knowledge-independent acceptance reluctance. Acceptance reluctance is a positional response set that is based on test-takers' willingness to choose late rather than early response options in DOMC items.[2] A high acceptance reluctance is characterized by the application of a stricter and more conservative response criterion for early options – and consequently, a preference for choosing later response options –, whereas a low acceptance reluctance is characterized by the application of a more liberal response criterion that favors the acceptance of early answers.

In a correlational study, we report first evidence of the existence of individual differences in acceptance reluctance. Whereas some test-takers consistently accept response options early when faced with questions to which they cannot know the answer, others consistently decide to reject more options. Thus, test-takers reliably show low or high acceptance reluctance, respectively. To investigate potential correlates of acceptance reluctance, we also assess other measures of response style and personality

[2] As a German translation of acceptance reluctance we chose *Festlegungszögern*.

such as risk-taking propensity (Dohmen et al., 2011; Lejuez et al., 2002), perfectionism (Stoeber, Otto, Pescheck, Becker, & Stoll, 2007), need for cognition (Bless, Wänke, Bohner, Fellhauer, & Schwarz, 1994; Cacioppo, Petty, Feinstein, & Jarvis, 1996), and overclaiming (Paulhus, Harms, Bruce, & Lysy, 2003; Ziegler, Kemper, & Rammstedt, 2013). Acceptance reluctance is not related to any of these variables, indicating that it might be a cognitive trait of its own (Kantner & Lindsay, 2012). We observe that a higher acceptance reluctance is related to better test scores in a DOMC knowledge test. This relationship remains significant even when statistically controlling for overall knowledge, suggesting that DOMC test scores may contain some construct-irrelevant variance due to individual differences in knowledge-independent acceptance reluctance.

Next, we find that a simulation based on the MORBIST model can reproduce several results that have been observed in the correlational study. Thus, first evidence of MORBIST's usefulness is obtained. The reproduced results include (a) a higher difficulty of DOMC items as compared to MC items, (b) a higher number of false acceptances of distractors in comparison to missed solutions in DOMC tests, (c) that DOMC items are more difficult when the solution is shown later, and (d) that in DOMC tests, fewer response options have to be shown than in MC tests. Another simulation shows that MORBIST can also account for the observation that higher acceptance reluctance is associated with better DOMC test scores. By predicting that DOMC test scores depend on individual acceptance reluctance, MORBIST suggests that DOMC test scores contain a share of construct-irrelevant variance. MORBIST thereby exemplifies how the investigation of internal response processes may help to better understand test validity, as it reveals a direct link between these internal processes and construct-irrelevant variance in DOMC test scores (cf. Borsboom et al., 2004).

Last, using a within-participants manipulation, we establish high versus low

acceptance reluctance to provide a final empirical test of the claim that acceptance reluctance causally influences DOMC test scores. In the experiment, a payoff manipulation encourages the selection of either early or late response options to induce low and high acceptance reluctance, respectively. Consistent with MORBIST's predictions and the correlational results, we find that a high acceptance reluctance payoff leads to significantly increased test scores. Therefore, the experiment confirms that DOMC test scores contain construct-irrelevant variance, raising questions on the viability of DOMC testing as an alternative to traditional MC testing. On the basis of another simulation, we compute a Bayes factor that evaluates the relative evidence the data provide for the MORBIST model in comparison to a null hypothesis that assumes no effect of acceptance reluctance. The Bayes factor indicates that MORBIST describes the observed gap between test scores under high and low acceptance reluctance substantially better than the null model. Thereby, MORBIST creates a direct link between psychological theory – as captured in the MORBIST simulation – and statistical inference. Such analyses that connect statistics and theory are desirable (Vanpaemel, 2010). Unfortunately, however, there is usually no clear link between statistical testing and substantive theory in psychological research (Kass, 2011; Rouder, Haaf, & Aust, 2018).

Because the experimental manipulation of acceptance reluctance was conducted within participants, it was possible to compare test-takers' performance in two successive DOMC tests. The comparison indicates that regardless of the level of acceptance reluctance, test-takers tend to score better in the second DOMC test than in their first. This result suggests that test-takers become better at solving DOMC items when increasing their familiarity with the new response mode.

3 Discussion

Taken together, the present research presents a mixed picture on the viability of DOMC testing. While successfully reducing the usability of testwiseness cues, DOMC testing introduces its own share of construct-irrelevant variance due to individual differences in acceptance reluctance. The finding that DOMC test scores contain some construct-irrelevant variance was suggested by a correlational investigation, causally confirmed in an experimental study, and is theoretically underpinned by the newly proposed MORBIST model that provides a link between DOMC response processes and DOMC test scores. By investigating the response processes in DOMC testing more closely than has been done before, our findings also identified knowledge-independent acceptance reluctance as a new cognitive trait that deserves further research attention in the future. Such research should for example investigate the temporal stability and the domain specifity of acceptance reluctance. Other research domains that may profit from a closer look at individual differences in acceptance reluctance include consumer product choice (e.g. Bearden & Connolly, 2007) and eyewitness performance in police lineups (e.g., Meisters, Diedenhofen, & Musch, 2018; Mickes, Flowe, & Wixted, 2012).

The present thesis reports the first experimental validation of DOMC test scores and therefore provides a methodologically particularly strong contribution to the yet small body of research on DOMC testing. The experimental validation strongly indicates that DOMC test scores are no less valid than MC test scores. While this finding may be considered a success for the relatively new DOMC format – after all, MC testing is the "gold standard" that has withstood long years of critical scientific enquiries –, it raises the important question whether the technically more demanding DOMC format should be employed in the first place. While MC tests can be administered relatively easy in paper-and-pencil settings, the DOMC test format

can only be administered using a computer (cf. Kubinger, 2009). If DOMC testing does not improve the reliability and validity of test scores – and even introduces its own method-specific variance – the question is raised whether there are any circumstances that justify the increased effort that is necessary to administer DOMC tests. Indeed, there may be applications that could benefit from the advantages that DOMC tests offer. For example, high-stakes tests that require high test security may profit from the fewer response options shown in DOMC items (Foster & Miller, 2009). Moreover, previous research indicated that DOMC can reduce testing time considerably, which may also be a desired characteristic in some settings, for example online research. The time savings may also allow to employ more test items in the same time, thus potentially increasing reliability. Given that the best established finding on the DOMC test format is an increased item difficulty, DOMC testing also offers a rather simple way to increase the difficulty of MC items, which may be an interesting application in some settings.

Moreover, the reduced susceptibility to testwiseness and random guessing counters one of the most prevalent criticisms of MC testing, namely that test-takers may obtain unduly high test scores even in the absence of any factual knowledge (Foster & Miller, 2009; Kubinger, 2009; Kubinger et al., 2010; cf. Owen & Froman, 1987). The reduced susceptibility to guessing was first observed in our experimental validation, establishing that the test score gap between MC and DOMC test scores is larger when test-takers have no information about the correct answer. Due to the presence of this interaction, the very low chance of guessing the solution (success rate: 5%) could not be explained by a mere higher difficulty of DOMC items. It is thus suggested that uninformed test-takers have strong difficulties to identify the correct solution in DOMC items, which is therefore another advantage of DOMC testing. However, as this result was surprising and was established in exploratory analyses, further

research has to investigate the guessing success of DOMC test-takers. If the reduced susceptibility to random guessing turns out to be a stable characteristic, an important additional benefit of DOMC testing has been identified. This potential benefit alone makes future investigations of the DOMC response format worthwhile.

To pass an integrated judgment on the viability of DOMC testing, several additional aspects also have to be considered. First, whereas we observe individual differences in acceptance reluctance that lead to unwanted variation in DOMC test scores, it is possible that the magnitude of such individual differences would decrease once test-takers get acquainted to the new test format. The experimental manipulation of acceptance reluctance successfully influenced test-takers' response criteria, indicating that test-takers are capable of strategically adapting their response behavior in DOMC tests. Test-takers also tended to perform better in their second DOMC test. Hence, it is possible that test-takers learn to employ a better level of acceptance reluctance after gathering more experience with DOMC testing. This might negate the problem of individual differences in acceptance reluctance in the long run. However, more research is needed to investigate the prevalence and the variation of individual differences in acceptance reluctance. Such research should also assess how these individual differences may change over time and across repeated testing.

When judging the usefulness of DOMC testing, we also should not forget that the body of research on the DOMC format is still small. MC testing on the other hand looks back on an entire century of scholarly research and practical experience. Moreover, MC testing has accumulated a large amount of criticism in that time, similar to the criticism that DOMC test scores contain some construct-irrelevant variance due to acceptance reluctance. In fact, there are many scholars and practitioners who completely object to the use of MC testing (see Frederiksen, 1984). Importantly, the viability of MC testing rests on the availability of item writing guidelines that

summarize the large amount of empirical findings. Given a long line of research, test creators *now* have all the tools they need to write effective MC test questions. Ninety years back however, such information was not available. This should be considered when judging the viability of DOMC testing that was only proposed in 2009 (Foster & Miller, 2009). Nine years later, there is still only little experience with the DOMC test format. It is also noteworthy that previous research – including the research presented here – constructed DOMC items by simply presenting response options of available MC items sequentially. Even though this is a convenient procedure, it is possible that there is a need for DOMC specific item writing guidelines. Up to date, no research has addressed which test items work well in DOMC format and which do not. It is therefore possible that all of the previous investigations did not observe the best possible performance of DOMC testing, simply because it was not known what characteristics good DOMC items should possess. Maybe these characteristics are different from the properties that well-written MC items have. I therefore suggest that future research on the DOMC test format should focus on how to maximize the quality of DOMC test items, instead of focusing on comparative studies of DOMC and MC testing. The investigation of new MORBIST model was a first step in this direction by focusing on the response processes in DOMC items specifically.

Given that both MC testing and DOMC testing have their own disadvantages—generally, it seems to be unavoidable that a testing procedure has some flaws—, it is plausible that the decision to employ the DOMC response format should incorporate situational requirements. If the positive aspects of DOMC testing – that include better control of testwiseness and guessing, and increased test security compared to conventional MC tests – are deemed desirable in a given application, DOMC testing can and should be employed.

Our results encourage the application of experimental validation procedures in

future studies. The reduced susceptibility to random guessing could only be identified because – due to the experimental design – it was known whether respondents had received the critical information that was necessary to answer items correctly. In general, experimental validation studies have important advantages over correlational studies because of the strict control over the validation criterion. Due to the improved study design, we thereby obtained strong evidence to the notion that DOMC test scores are as valid MC test scores. Importantly, the applicability of experimental validation studies is not limited to the comparison of different variants of multiple-choice testing; in principle, many testing procedures could be validated using an experimental approach. Even test formats such as essays or oral examinations can be investigated using experimental validations.

Finally, the proposed model of response behavior in sequential tests (MORBIST) was found to provide a promising foundation for investigating the response processes involved in answering DOMC questions. In particular, MORBIST successfully reproduced several empirically observed properties of the DOMC test format and also accounted for the observation that high acceptance reluctance leads to better test scores than low acceptance reluctance. As the present findings indicate that it is worthwhile to further investigate the properties of DOMC testing, the MORBIST formulations may offer a helpful guidance for all researchers who are interested in studying DOMC tests in the future.

References

Allan, A. (1992). Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers. *Language Testing*, *9*(2), 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*(2), 109–128.

Baldiga, K. (2013). Gender differences in willingness to guess. *Management Science*, *60*(2), 434–448.

Bearden, J. N., & Connolly, T. (2007). Multi-attribute sequential search. *Organizational Behavior and Human Decision Processes*, *103*(1), 147–158.

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement*, *28*(1), 23–35.

Berg, I. A. (1955). Response bias and personality: The deviation hypothesis. *Journal of Psychology*, *40*, 61–72.

Bless, H., Wänke, M., Bohner, G., Fellhauer, R. F., & Schwarz, N. (1994). Need for cognition: Eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben [Need for cognition: A scale measuring engagement and happiness in cognitive tasks]. *Zeitschrift für Sozialpsychologie*, *25*, 147–154.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071.

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). Atlanta, GA: Georgia State University: College Reading and Learning Assistance. Retrieved from http://eric.ed.gov/?id=ED240928

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197–253.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475–494.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*(3), 3–31.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Diamond, J. J., & Evans, W. J. (1972). An investigation of the cognitive correlates of test-wiseness. *Journal of Educational Measurement*, *9*(2), 145–150.

Diedenhofen, B., & Musch, J. (2017). Empirical option weights improve the validity of a multiple-choice knowledge test. *European Journal of Psychological Assessment*, *33*(5), 336–344.

Dirkzwager, A. (1996). Testing with personal probabilities: 11-year-olds can correctly estimate their personal probabilities. *Educational and Psychological Measurement*, *56*(6), 957–971.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral conse-

quences. *Journal of the European Economic Association*, *9*(3), 522–550.

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine*, *77*(10), 103–104.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, *10*(2), 133–143.

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: Erlbaum.

Dressel, P. L., & Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement*, *13*(4), 574–595.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, *108*(19), 7716–7720.

Ebel, R. L. (1971). How to write true-false test items. *Educational and Psychological Measurement*, *31*(2), 417–426.

Edwards, B. D., Arthur, W., & Bruce, L. L. (2012). The three-option format for knowledge and ability multiple-choice tests: A case for why it should be more commonly used in personnel testing. *International Journal of Selection and Assessment*, *20*(1), 65–81.

Embretson, S. E. (2007). Construct validity: A universal validity system or just

another test evaluation procedure? *Educational Researcher*, *36*(8), 449–455.

Erdfelder, E., & Musch, J. (2006). Experimental methods of psychological assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 205–220). Washington, D.C.: American Psychological Association.

Farley, J. K. (1989). The multiple-choice test: Writing the questions. *Nurse Educator*, *14*(6), 10–12.

Foster, D., & Miller, H. (2009). A new format for multiple-choice testing: Discrete-option multiple-choice. results from early studies. *Psychology Science Quarterly*, *51*(4), 355–369.

Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice*, *7*(2), 33–38.

Frary, R. B. (1991). The none-of-the-above option: An empirical study. *Applied Measurement in Education*, *4*(2), 115–124.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, *39*(3), 193–202.

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

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

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-

writing rules. *Applied Measurement in Education*, *2*(1), 37–50.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, *53*(4), 999–1010.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–333.

Hijji, B. M. (2017). Flaws of multiple choice questions in teacher-constructed nursing examinations: A pilot descriptive study. *Journal of Nursing Education*, *56*(8), 490–496.

Hohensinn, C., & Baghaei, P. (2017). Does the position of response options in multiple-choice tests matter? *Psicológica*, *38*(1), 93–109.

Hughes, C. A., Salvia, J., & Bott, D. (1991). The nature and extent of test-wiseness cues in seventh- and tenth-grade classroom tests. *Assessment for Effective Intervention*, *16*(2-3), 153–163.

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*(2), 291–306.

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic*

Medicine, *77*(2), 156–161.

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition*, *40*(8), 1163–1177.

Kass, R. E. (2011). Statistical inference: The big picture. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *26*(1), 1–9.

Kellen, D., Klauer, K. C., & Singmann, H. (2012). On the measurement of criterion noise in signal detection theory: The case of recognition memory. *Psychological Review*, *119*(3), 457–479.

Kingston, N. M., Tiemann, G. C., Miller, H., & Foster, D. (2012). An analysis of the discrete-option multiple-choice item type. *Psychological Test and Assessment Modeling*, *54*(1), 3–19.

Kubinger, K. D. (2009). Psychologische Computerdiagnostik [computerized diagnostics in psychology]. *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie*, *57*(1), 23–32.

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment*, *18*(1), 111–115.

Lee, H., & Winke, P. (2013). The differences among three-, four-, and five-option-item formats in the context of a high-stakes english-language listening test. *Language Testing*, *30*(1), 99–123.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G.

L., ... Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The balloon analogue risk task (bart). *Journal of Experimental Psychology: Applied*, *8*(2), 75–84.

Lindner, M. A., Strobel, B., & Köller, O. (2015). Multiple-Choice-Prüfungen an Hochschulen? Ein Literaturüberblick und Plädoyer für mehr praxisorientierte Forschung [Are multiple-choice exams useful for universities? A literature review and argument for a more practice oriented research]. *Zeitschrift für Pädagogische Psychologie*, *29*, 133–149.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437–448.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Malouff, J. M., & Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education*, *60*(3), 245–256.

Malouff, J. M., Emmerton, A. J., & Schutte, N. S. (2013). The risk of a halo bias as a reason to keep students anonymous during grading. *Teaching of Psychology*, *40*(3), 233–237.

McGee, R. K. (1962). The relationship between response style and personality variables: The measurement of response acquiescence. *The Journal of Abnormal and Social Psychology*, *64*(3), 229–233.

Meisters, J., Diedenhofen, B., & Musch, J. (2018). Eyewitness identification in

simultaneous and sequential lineups: An investigation of position effects using receiver operating characteristics. *Memory. Advance Online Publication.*

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Messick, S. (1993). Foundations of validity: Meaning and consequences in psychological assessment. *ETS Research Report Series*, *1993*(2), 1–18.

Metfessel, N. S., & Sax, G. (1958). Systematic biases in the keying of correct responses on certain standardized tests. *Educational and Psychological Measurement*, *18*(1958), 787–790.

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied*, *18*(4), 361–376.

Millman, J., Bishop, C. H., & Ebel, R. L. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, *25*(3), 707–726.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment.* London: Sage.

Owen, S. V., & Froman, R. D. (1987). What's wrong with three-option multiple choice items? *Educational and Psychological Measurement*, *47*(2), 513–522.

Papenberg, M., & Musch, J. (2017). Of small beauties and large beasts: The quality of distractors on multiple-choice tests is more important than their quantity.

Applied Measurement in Education, *30*(4), 273–286.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review*, *10*(3), 556–569.

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology*, *84*(4), 890–904.

Poizner, S. B., Nicewander, W. A., & Gettys, C. F. (1978). Alternative response and scoring methods for multiple-choice items: An empirical study of probabilistic and ordinal response modes. *Applied Psychological Measurement*, *2*(1), 83–96.

Richardson, K., & Norgate, S. H. (2015). Does IQ really predict job performance? *Applied Developmental Science*, *19*(3), 153–169.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13.

Rogers, W. T., & Bateson, D. J. (1991). The influence of test-wiseness on performance of high school seniors on school leaving examinations. *Applied Measurement in Education*, *4*(2), 159–183.

Rogers, W. T., & Yang, P. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment*, *12*(3), 247–259.

Rouder, J. N., Haaf, J. M., & Aust, F. (2018). From theories to models to predictions: A bayesian model comparison approach. *Communication Monographs*, *85*(1),

41–56.

Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement*, *14*(1), 15–22.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. *Review of Educational Research*, *49*(2), 252–279.

Scruggs, T. E., & Lifson, S. A. (1985). Current conceptions of test-wiseness: Myths and realities. *School Psychology Review*, *14*(3), 339–350.

Sherriffs, A. C., & Boomer, D. S. (1954). Who is penalized by the penalty for guessing? *Journal of Educational Psychology*, *45*(2), 81–90.

Siroky, K., & Di Leonardi, B. C. (2015). Refine test items for accurate measurement: Six valuable tips. *Journal for Nurses in Professional Development*, *31*(1), 2–8.

Slakter, M. J., Koehler, R. A., & Hampton, S. H. (1970). Grade level, sex, and selected aspects of test-wiseness. *Journal of Educational Measurement*, *7*(2), 119–122.

Štěpánek, J., & Šimková, M. (2013). Design and implementation of simple interactive e-learning system. *Procedia-Social and Behavioral Sciences*, *83*, 413–416.

Stoeber, J., Otto, K., Pescheck, E., Becker, C., & Stoll, O. (2007). Perfectionism and competitive anxiety in athletes: Differentiating striving for perfection and negative reactions to imperfection. *Personality and Individual Differences*, *42*(6), 959–969.

Swineford, F. (1938). The measurement of a personality trait. *Journal of Educational Psychology*, *29*(4), 295–300.

Swineford, F. (1941). Analysis of a personality trait. *Journal of Educational Psychol-*

ogy, *32*(6), 438–444.

Tamir, P. (1991). Multiple choice items: How to gain the most out of them. *Biochemical Education*, *19*(4), 188–192.

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*(2), 198–206.

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education in Practice*, *6*(6), 354–363.

Tellinghuisen, J., & Sulikowski, M. M. (2008). Does the answer order matter on multiple-choice exams? *Journal of Chemical Education*, *85*(4), 572–575.

Thoma, G.-B., & Köller, O. (2018). Test-wiseness: Ein unterschätztes Konstrukt? [Test-wiseness: An underestimated construct?]. *Zeitschrift für Bildungsforschung*, 1–18.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25–29.

Thorndike, R. L. (1938). Critical note on the pressey interest-attitudes test. *Journal of Applied Psychology*, *22*(6), 657–658.

Tomkowicz, J., & Rogers, W. T. (2005). The use of one-, two-, and three-parameter and nominal item response scoring in place of number-right scoring in the presence of test-wiseness. *Alberta Journal of Educational Research*, *51*(3), 200–215.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the bayes

factor. *Journal of Mathematical Psychology*, *54*(6), 491–498.

Wechsler, D. (1944). *The measurement of aduld intelligence* (3rd ed.). Baltimore: Williams & Wilkins.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS One*, *11*(3), e0152719.

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment*, *23*(3), 279–291.

Wevrick, L. (1962). Response set in a multiple-choice test. *Educational and Psychological Measurement*, *22*(3), 533–538.

Wiley, L. N., & Trimble, O. C. (1936). The ordinary objective test as a possible criterion of certain personality traits. *School and Society*, *43*, 446–448.

Willing, S. (2013). *Discrete-option multiple-choice: Evaluating the psychometric properties of a new method of knowledge assessment.* (Doctoral dissertation, Heinrich-Heine University, Düsseldorf, Germany). Retrieved from http://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=27633.

Willing, S., Ostapczuk, M., & Musch, J. (2015). Do sequentially-presented answer options prevent the use of testwiseness cues on continuing medical education tests? *Advances in Health Sciences Education*, *20*(1), 247–263.

Ziegler, M. (2015). "F*** you, i won't do what you told me!" – response biases as threats to psychological assessment. *European Journal of Psychological Assessment*, *31*(3), 153–158.

Ziegler, M., Kemper, C., & Rammstedt, B. (2013). The vocabulary and overclaiming test (voc-t). *Journal of Individual Differences*, *34*(1), 32–40.

Eidesstattliche Erklärung
laut §5 der Promotionsordnung vom 06.12.2013
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

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.

....

Datum, Ort Martin Papenberg

Appendix: Original Research Articles

Article 1

Papenberg, M., Willing, S., & Musch, J. (2017). Sequentially Presented Response Options Prevent The Use Of Testwiseness Cues In Multiple-Choice Testing. *Psychological Test and Assessment Modeling, 59*(2), 245–266.

I was the main author of Article 1. To prepare the manuscript, I created the study materials and devised an experimental design that I implemented using Unipark software. I was also responsible for data analysis.

Article 2

Papenberg, M., Diedenhofen, B., & Musch, J. (2018). An Experimental Validation of Sequential Multiple-Choice Tests. Manuscript submitted for publication.

I was the main author of Article 2. To prepare the manuscript, I implemented the empirical study that is reported using Unipark software and analyzed the resulting data.

Article 3

Papenberg, M., & Musch, J. (2018). An Investigation Of Sequential Response Behavior In Discrete-Option Multiple-Choice Knowledge Tests. Manuscript submitted for publication.

I was the main author of Article 3. To prepare the manuscript, I (a) created and compiled study materials, (b) devised the designs of two empirical studies, (c) implemented the two studies using Unipark software, (d) analyzed the resulting data, (e) conceptualized a mathematical model, and (f) conducted computer simulations on the basis of this model.

Psychological Test and Assessment Modeling, Volume 59, 2017 (2), 245-266

Sequentially presented response options prevent the use of testwiseness cues in multiple-choice testing

Martin Papenberg[1], *Sonja Willing*[2] *& Jochen Musch*[3]

Abstract

Testwiseness — the ability to find subtle cues to the solution by comparing all available response options — threatens the validity of multiple-choice (MC) tests. Discrete-option multiple-choice (DOMC) is an alternative testing format in which response options are presented sequentially rather than simultaneously. 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. 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 response options that had to be presented. The DOMC format therefore seems to offer an interesting alternative to traditional MC testing.

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

[1]*Correspondence concerning this article should be addressed to:* Martin Papenberg, PhD, University of Duesseldorf, Department of Experimental Psychology, Universitätsstr. 1, Building 23.03, 40225 Duesseldorf, Germany; email: martin.papenberg@uni-duesseldorf.de

[2]University of Duesseldorf

[3]University of Duesseldorf

Introduction

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 response options, one of which is the solution (Foster & Miller, 2009). The stem is the core of an item, which presents the question that has to be answered. Next to the stem, all possible response options are presented. The examinee's task is to choose the correct answer from among this set of options. Sometimes this variant of MC testing is called "single-choice" testing because only one response option is the correct solution. 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 objectively 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 doubted that recording the mere selection of a MC response option adequately assesses higher order thinking skills (Hancock, 1994). The selection of an MC option may not reveal actual knowledge of a respondent, but simply indicate the alternative a respondent considers to be the most plausible (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 response 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 response options, pupils could often identify the correct solution without even knowing the question (cf. also Sparfeldt, Kimmel, Löwenkamp, Steingräber & Rost, 2012).

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 – add construct-irrelevant variance to MC test scores (Haladyna & Downing, 2004; Millman, Bishop & Ebel, 1965; 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. Several other investigations also showed a high prevalence of item flaws that allowed identifying the solution (e.g. Hughes et al., 1991; Metfessel & Sax, 1958; Tomkowicz & Rogers, 2005). 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.

Testing formats that control for the application of testwiseness are therefore desirable. Computerized alternatives to traditional MC tests allow more flexibility in presenting items, and presenting response options sequentially may help to control for guessing (Kubinger, 2009). A sequential presentation of response options was first used by Srp (1994; cf. Kubinger, 2009) in a test of logical reasoning. In a study of what they called discrete-option multiple-choice (henceforth DOMC) testing, Foster and Miller (2009) discussed that a sequential presentation of response options might help to prevent the use of testwiseness cues, because a sequential presentation precludes the simultaneous comparison of all available response options prior to answering.

Like a standard MC item, a DOMC item consists of a stem and a number of response options, one of which is the solution (Foster & Miller, 2009). The difference from standard MC items is that response 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 response 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 response 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 response 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 response 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 response options as another advantage of the DOMC format. If a response 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.

Foster and Miller (2009) found that DOMC questions were more difficult than standard MC questions. This pattern was replicated in a subsequent study using a larger sample (Kingston, Tiemann, Miller, & Foster, 2012), and was also observed in a study by Hansmann (2010) using items from Srp's (1994) sequential logical reasoning test. 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 response 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 response option separately, whereas in MC testing, all response 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 response 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, more direct evidence is needed to allow definitive conclusions regarding whether the DOMC answer format allows to improve the control of testwiseness. In the present study, we therefore investigated whether DOMC testing controls 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.

Previous investigations showed that item-total-score correlations and internal consistencies were comparable for MC and DOMC items. These findings were interpreted as showing that items were functioning equally well in both formats (Foster & Miller, 2009; Kingston et al., 2012). However, the internal consistency of item scores may be increased by the presence of construct-irrelevant response dimensions that affect all items simultaneously (Green, Lissitz & Mulaik, 1977). Hence, internal consistency does not provide an appropriate estimate of item functioning if item responses are influenced by additional factors such as testwiseness (cf. Cortina, 1993). To go beyond a correlational comparison and to establish an unambiguous and direct causal link between testwiseness and test scores, we experimentally manipulated the susceptibility of items to the use of item cues.

By examining the causal processes that precede behavioral test responses, we followed recent recommendations regarding the validation of testing procedures (e.g. Borsboom, Mellenbergh, & van Heerden, 2004; Embretson, 2007; Lissitz & Samuelson, 2007).

Thus, the present study offers an experimental contribution to the validation of the DOMC test format that has hitherto been tested using correlational (Foster & Miller, 2009; Kingston et al., 2012) or quasi-experimental designs (Willing, Ostapczuk, & Musch, 2015). Willing et al. (2015) compared the difficulties of items from a continuing medical education test that were either presented in MC or DOMC format. Some of the items under investigation contained item cues; cue availability and item content were therefore confounded. Possibly, the observed interaction of test format and cue availability on test scores was therefore the result of differences in item content rather than differences in cue availability. In the present study, we therefore experimentally manipulated the presence of item cues.

To properly manipulate the availability of item cues, a testwiseness test is required. 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 for the tested sample; 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 students 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 students to work on a test that also contained item cues, and we additionally informed the students 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 students to each of the three groups, and within these groups, we randomly assigned the students 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 response 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 (85.64 % female) between the ages of 19 and 35 years ($M = 22.79$, $SD = 2.80$). All students were recruited via announcements in social network student groups. The data of an additional 23 students 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*. The experiment was conducted in accordance with the ethical standards of psychological research. At the end of the test, students 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):

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 reads:

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.

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.

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

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 response 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. This was confirmed in a multiple choice pretest with 130 psychology students who were asked to indicate whether they were certain that they had selected the correct solution. For 20 of the 24 testwiseness questions, not a single student indicated to be certain of his or her answer; for the remaining 4 items, only one of the 130 students indicated to be certain of the answer. Thus, the students could not confidently identify a solution to these items. However, each item contained a cue that could be used to infer the solution with at least some certainty.

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 response 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. Response options were also presented in a random order. In the MC condition, one item was presented per page along with all of the possible response options. In the DOMC condition, response options were presented sequentially.

Design

The study used a 2 × 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, students were asked to indicate their age, sex, and education. They were then randomly assigned to one of the six experimental conditions that resulted from crossing the 2 × 3 levels of the two experimental factors.

Students 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 students were informed about the stopping criteria employed in the sequential presentation procedure.

The 57 students in the no cue condition worked on test items that did not contain any item cues. The 61 students in the cue condition worked on items that contained such cues. In the informed cue condition, another 63 students also worked on items containing cues; they were additionally informed about the presence and the nature of these cues before the test started. To this end, prior to the start of the test, each of the four cues was described using an example.

For DOMC items, the question stem was presented above the first, randomly drawn response option. Test takers decided whether they accepted this response option as the solution by clicking one of two buttons labeled “true” and “false”. When test takers decided to reject a response option that was a distractor, the next randomly determined response option was shown below the question stem that remained on display throughout. Response options were shown until a response was recorded, and there was no time limit on the test takers’ response decisions. After the correctness of one response option had

been assessed, it was not possible to go back to previous options, nor was it possible to go back to correct answers to previous items.

Data analysis

For each student, 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. We used R (3.3.3, R Core Team, 2016) and the R-packages *afex* (0.16.1, Singmann, Bolker, Westfall, & Aust, 2016), *effsize* (0.7.1, Torchiano, 2016), and *papaja* (0.1.0.9485, Aust & Barth, 2016) for all our analyses. For the statistical tests, an alpha level of .05 was used. To compare the testwiseness scores across conditions, a 2 $\times$ 3 (testing format [DOMC, MC] $\times$ availability of testwiseness cues [no cue, cue, informed cue]) ANOVA was computed. ANOVA effect sizes were computed using the generalized eta-squared η^2_G, indicating the proportion of the variance explained by each factor or interaction (Olejnik & Algina, 2003). Effect sizes for the difference between two means were calculated using Cohen's *d* (1988).

Results

Testwiseness scores

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, $F(1, 175) = 53.56$, $p < .001$, $\eta^2_G = .23$. 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, $F(2, 175) = 120.52$, $p < .001$, $\eta^2_G = .58$. 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, $F(2, 175) = 12.87$, $p < .001$, $\eta^2_G = .13$ (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 response 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 (8.53 [SD = 3.29] vs. 6.53 [SD = 2.74]), $t(60) = 2.61$,

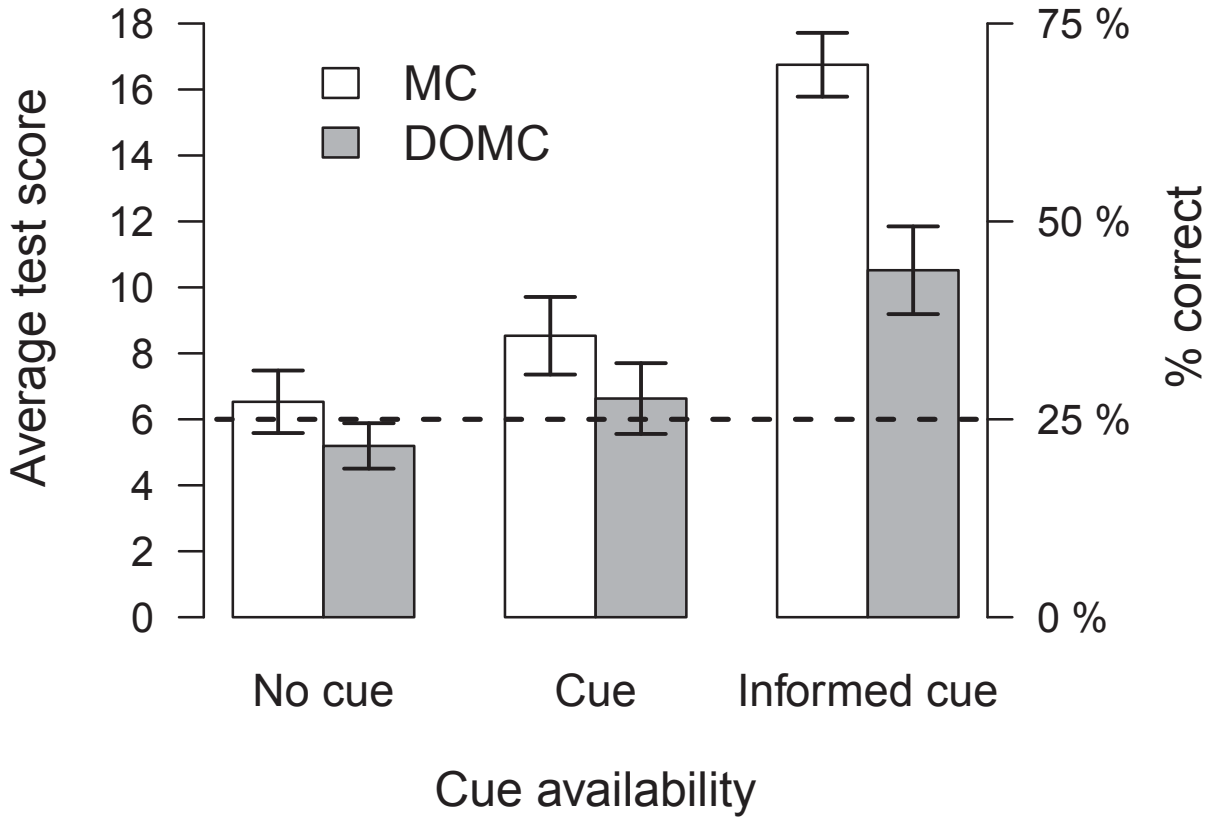

Figure 1: Test scores and their 95 % confidence intervals are shown as a function of (1) the two testing formats multiple-choice (MC) and discrete-option multiple-choice (DOMC), and (2) the availability of testwiseness cues. The dashed line indicates the chance level of 25 %, which is the expected test score for a random guessing strategy. The maximal possible test score was 24.

$p < .01$, $d = 0.66$, and in the DOMC condition (6.63 [*SD* = 2.84] vs. 5.19 [*SD* = 1.96]), $t(56) = 2.26$, $p < .05$, $d = 0.60$. 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 (16.75 [*SD* = 2.96] vs. 8.53 [*SD* = 3.29]), $t(64) = 10.68$, $p < .001$, $d = 2.64$, and in the DOMC condition (10.52 [*SD* = 3.39] vs. 6.63 [*SD* = 2.84]), $t(50) = 4.49$, $p < .001$, $d = 1.25$. 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 (6.53 [*SD* = 2.74] vs. 5.19 [*SD* = 1.96]), $t(61) = 2.23$, $p < .05$, $d = 0.56$, the cue condition (8.53 [*SD* = 3.29] vs. 6.63 [*SD* = 2.84]), $t(55) = 2.33$, $p < .05$, $d = 0.62$, and in the informed cue condition (16.75 [*SD* = 2.96] vs. 10.52 [*SD* = 3.39]), $t(59) = 7.61$, $p < .001$, $d = 1.98$.

To further explore how DOMC prevents the use of testwisenes cues, we tested whether the reduction in cue usage was moderated by the type of cue.[1] To this end, we repeated the ANOVA from above, but included the repeated measures factor cue type [direct opposites, categorical exclusive, middle value, longest alternative] in addition to the factors test format [MC, DOMC] and cue susceptibility [no cue, cue, informed cue]. Table 1 shows the results of this $4 \times 2 \times 3$ mixed ANOVA to which we applied a Greenhouse-Geisser correction for violation of sphericity (Greenhouse & Geisser, 1959). There was a significant main effect of cue type, $F(2.80, 490.15) = 37.98$, $p < .001$, $\eta^2_G = .12$. The longest alternative and categorical exclusive cue led to higher test scores than the middle value and the direct opposites cue (see Figure 2). This pattern is in accordance with the fact that the direct opposites cue and the middle value cue are not perfect predictors of the solution. Using these cues, however, allows to eliminate two of the four response options, and thereby improves the chance of guessing the correct solution from 25 % to 50 %. In contrast, both the categorical exclusive and the longest alternative cue directly point to the solution, and students made almost perfect use of these cues in the MC test when they had been informed of their presence. A significant two-way interaction between cue susceptibility and cue type indicated that informing students about the nature of the cues improved test scores more strongly for some cues than for others, $F(5.60, 490.15) = 26.47$, $p < .001$, $\eta^2_G = .17$, and the significant three-way interaction between cue susceptibility, cue type and test format, $F(5.60, 490.15) = 3.29$, $p < .01$, $\eta^2_G = .02$, indicated that the superior control of testwiseness in the DOMC test format was mainly due to a better prevention of the use of the categorical exclusive and the longest alternative cue.

[1] We are grateful to an anonymous reviewer for suggesting this additional analysis.

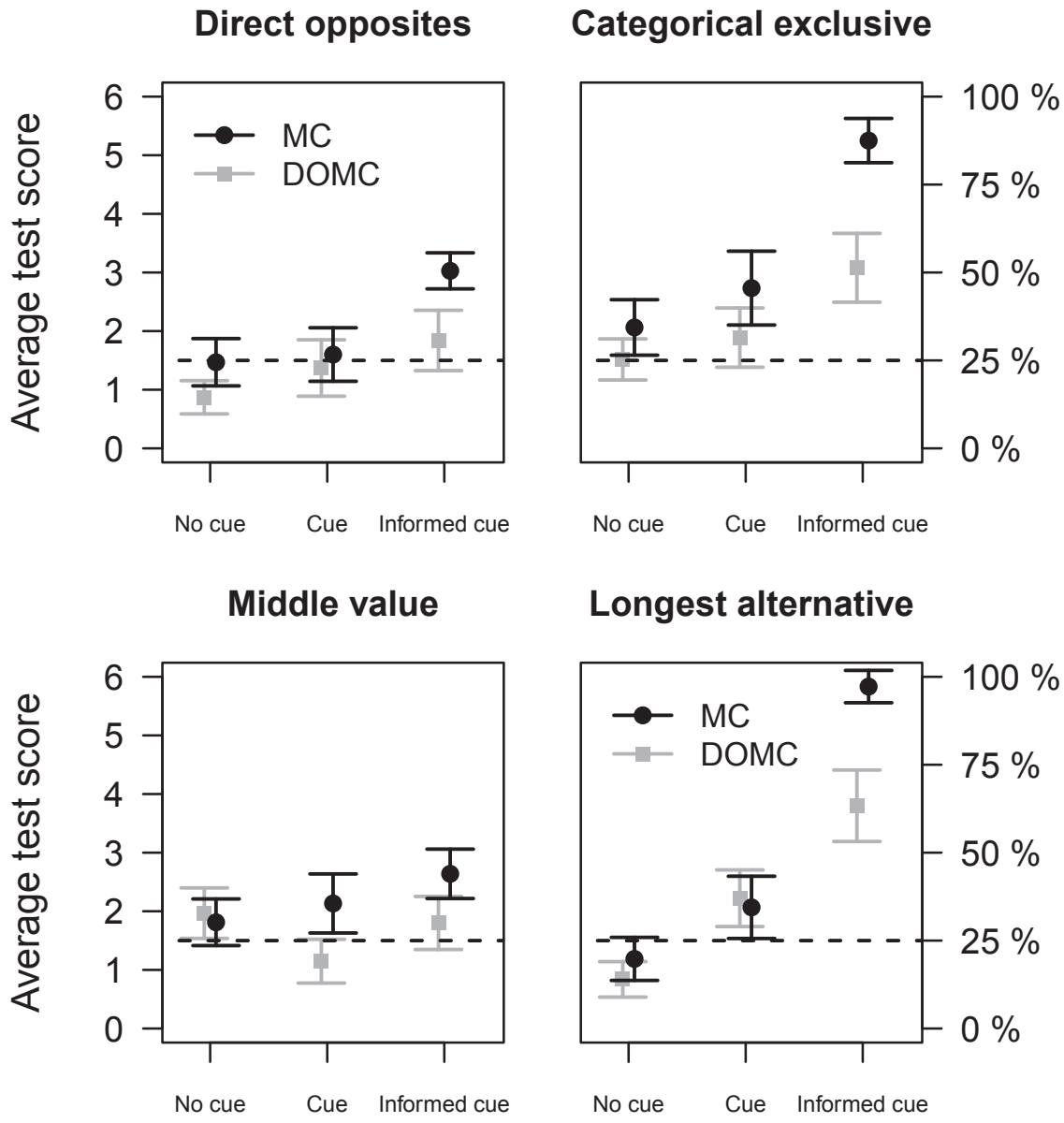

Figure 2: Average test scores and their 95 % confidence intervals are shown by cue type, test format, and cue availability. Each testwiseness cue was included in six testwiseness items. The dashed line indicates the chance level of 25 %.

Table 1: Results of a 4 × 2 × 3 mixed ANOVA investigating the influence of cue type, test format, and cue availability on testwiseness test scores

Effect	F	df_1^{GG}	df_2^{GG}	p	η^2_G
Cue availability	120.52	2	175	< .001	.322
Test format	53.56	1	175	< .001	.095
Cue type	37.98	2.80	490.15	< .001	.124
Cue availability × Test format	12.87	2	175	< .001	.048
Cue availability × Cue type	26.47	5.60	490.15	< .001	.165
Test format × Cue type	2.59	2.80	490.15	.057	.010
Cue availability × Test format × Cue type	3.29	5.60	490.15	.004	.024

Note. The degrees of freedom were corrected using the Greenhouse-Geisser correction. The cue type factor comprised the four testwiseness cues (direct opposites, categorical exclusive, middle value, and longest alternative).

Number of response options presented in the DOMC condition

In the DOMC condition, the presentation of response 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 response options. The position of the solution was randomly varied. The stopping criteria reduced the average number of response 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 response options. Thus, on average, a perfect test taker could be expected to see 2.5 out of the 4 possible response options in the DOMC condition. For a less than perfect test taker, the presentation of a smaller number of response 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 40.51 % of cases, the item presentation ended after the presentation of the very first option. In 24.35 % of cases, this option happened to be the solution, and in 16.16 % of cases, this option was a distractor that was wrongly accepted

Table 2: Distribution of the number of response options students were shown in the DOMC test

	N options					
	1	2	3	4	*M*	*SD*
no cue	43%	33%	20%	4%	1.85	0.87
cue	40%	31%	22%	7%	1.96	0.95
informed cue	38%	32%	18%	12%	2.04	1.02

Note. Percentages show how often one, two, three or all four options were shown to the test takers. The last two columns show the mean and the standard deviation of the number of options shown.

as the solution. The item presentation ended after the second, third, and fourth response options were presented for 31.98 %, 20.38 %, and 7.13 % of all items, respectively. On average, this resulted in an end to the item presentation after 1.94 out of the four possible response options ($SD = 0.94$).

When analyzing the number of response options participants were presented with separately for the three cue conditions, an interesting pattern emerged (see Table 2): participants tended to be presented with more response options if items were more susceptible to the use of testwiseness cues. In the no cue condition, test takers were presented with 1.85 (SD = 0.87) response options on average. In the cue condition, test takers were presented with 1.96 (SD = 0.95) response options on average, and in the informed cue condition, test takers were presented with 2.04 (SD = 1.02) response options, respectively. The most likely reason for this pattern is that there is a positive relationship between the number of correct responses and the number of response options test takers have to be presented with: when test takers are more apt at solving DOMC items correctly, they will produce less false alarms and therefore score higher. Consequently, test takers with higher scores – that is, test takers in the cue and in the informed cue condition – are presented with more response options than test takers that did not obtain any cues.

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), $t(174) = 3.60$, $p < .001$, $d = 0.52$. Thus, due to the smaller number of response options that had to be presented in the DOMC condition, the time needed to answer all items was reduced by 21 % when the response options were presented sequentially. However, participants needed longer to read the extended instructions in the DOMC condition ($M = 82.78$ s, $SD = 50.11$ vs. $M = 20.44$ s, $SD = 9.30$), $t(87) = 11.17$, $p < .001$, $d = 1.80$. 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 ($M = 474.96$ s, $SD = 212.13$) and DOMC conditions ($M = 441.36$ s, $SD = 174.44$), $t(179) = 1.17$, $p = .24$, $d = 0.17$.

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 response 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 response 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. However, even in the MC test, performance was never perfect. Students answered 16.75 of the 24 testwiseness items correctly when they had been informed about the presence of testwiseness cues. This less than perfect performance was not unexpected because only the longest alternative and the categorical exclusive cue were perfect predictors of the solution; the direct opposites and the middle value cue only improved the chance of guessing the correct solution from 25 % to 50 % by allowing to eliminate two of the four response options. Therefore, the expected test score assuming perfect cue usage was 18 rather than 24 (out of 24). The empirical results follow this expected pattern closely in the MC condition: when they were informed about the presence of these cues,

students scored almost perfectly for items that included the longest alternative (97 %) or categorical exclusive cue (88 %). Their performance was also very close to the expected 50 % for items containing a middle value or direct opposite cue (solution percentages for these item types were 44 % and 50 %, respectively). Thus, DOMC prevented cue usage most effectively for the item cues that most directly pointed towards the solution (the longest alternative and the categorical exclusive cue).

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 response 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 (40.51 %), the presentation of DOMC items ended after the presentation of only one of the four possible response options. Only 1.94 options had to be shown on average, and in only 7.13 % of all items were all four response options presented to the test taker. This large reduction in the number of response 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 response 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 response options. Test security was also enhanced because many response 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.

One obvious disadvantage, however, is that the DOMC format is technically more demanding and less easily implemented in school or university settings. The DOMC format requires a computerized presentation of test items (Kubinger, 2009), and DOMC exams are therefore not so easily administered and scored as traditional MC paper and pencil exams.

While DOMC was successful in controlling construct-irrelevant variance due to test-wiseness, it is possible that DOMC also introduces method-specific construct-irrelevant

variance if there are additional factors beyond ability that affect test takers' responses to DOMC items. Responses to DOMC items are given in a state of incomplete information, and individual differences in response style may influence test takers' decisions (cf. Cronbach, 1946). For example, anxious test takers may feel rushed to accept a plausible DOMC response option early, whereas more strong-nerved test takers might be willing to wait longer for a suitable response. Future research should address this question to rule out the possibility that DOMC responses are contaminated with individual differences in response style. In the case of traditional MC tests, some findings suggest that there might be differences in the willingness to guess between male and female test takers (Baldiga, 2013; Ben-Shakhar & Sinai, 1991). If such gender-dependent differences in response style occur, they might bias the results of DOMC tests. For this reason, it is desirable to more directly measure potential individual differences in response style in future studies of DOMC testing.

The present sample consisted of a rather selected group of mostly female psychology students who are most likely rather familiar with any kind of tests and response formats. Further research is therefore needed to explore whether the present results can be generalized to different samples of test takers. Another limitation of the current results should be addressed in future research. Although we established that DOMC helps to control the use of testwiseness cues, this result was shown via experimental manipulation and not by controlling individual differences in testwiseness. Therefore, to what extent DOMC is capable of reducing construct-irrelevant variance due to individual differences in testwiseness is still an open question. Another limitation is that based on the present results, we cannot judge the degree to which testwiseness impairs the interpretability of test scores in everyday testing situations. This is because the magnitude of potential problems associated with the presence of testwiseness cues depends on the prevalence of such cues. If items are well-written, testwiseness may not be a threat to the validity of MC tests at all. However, previous findings suggest that even in high-stakes assessments, a considerable portion of teacher-made MC items do contain cues to their solution (e.g., Brozo, Schmelzer, & Spires, 1984; Tarrant & Ware, 2008).

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 response 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 test-wiseness in efl/esl reading test takers. *Language Testing*, *9*(2), 101–119. doi: 10.1177/026553229200900201

Aust, F., & Barth, M. (2016). *Papaja: Create apa manuscripts with rmarkdown*. Retrieved from `https://github.com/crsh/papaja`

Baldiga, K. (2013). Gender differences in willingness to guess. *Management Science*, *60*(2), 434–448. doi:10.1287/mnsc.2013.1776

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement*, *28*(1), 23–35. doi:10.1111/j.1745-3984.1991.tb00341.x

Borsboom, D., Mellenbergh, G. J., & Heerden, J. van. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. doi:10.1037/0033-295X.111.4.1061

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 & Learning Assistance. ERIC database (ED240928). Retrieved from `http://eric.ed.gov/?id=ED240928`

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

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104. doi:10.1037/0021-9010.78.1.98

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475–494.

Diamond, J. J., & Evans, W. J. (1972). An investigation of the cognitive correlates of test-wiseness. *Journal of Educational Measurement*, *9*(2), 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: Erlbaum.

Embretson, S. E. (2007). Construct validity: A universal validity system or just another

test evaluation procedure? *Educational Researcher*, *36*(8), 449–455. doi: 10.3102/0013189X07311600

Foster, D., & Miller, H. (2009). A new format for multiple-choice testing: Discrete-option multiple-choice. Results from early studies. *Psychology Science Quarterly*, *51*(4), 355–369.

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

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, *37*(4), 827–838. doi:10.1177/001316447703700403

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112. doi:10.1007/BF02289823

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

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27. doi:10.1111/j.1745-3992.2004.tb00149.x

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of Experimental Education*, *62*(2), 143–157. doi:10.1080/00220973.1994.9943836

Hansmann, B. C. (2010). *About the psychometric quality of various multiple choice response formats in the context of cultural differences between Austria and the United States of America* (Unpublished diploma thesis). University of Vienna, Austria.

Holmes, P. (2002). *Multiple evaluation versus multiple choice as testing paradigm* (Unpublished PhD thesis). Twente University, Enschede, Netherlands. Retrieved from `http://doc.utwente.nl/38691/1/t0000017.pdf`

Hughes, C. A., Salvia, J., & Bott, D. (1991). The nature and extent of test-wiseness cues in seventh–and tenth-grade classroom tests. *Assessment for Effective Intervention*, *16*(2-3), 153–163. doi:10.1177/153450849101600310

Kingston, N. M., Tiemann, G. C., Miller, H., & Foster, D. (2012). An analysis of the discrete-option multiple-choice item type. *Psychological Test and Assessment Modeling*, *54*(1), 3–19.

Kubinger, K. D. (2009). Psychologische Computerdiagnostik [Computerized diagnostics

in psychology]. *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie*, *57*(1), 23–32. doi:10.1024/1661-4747.57.1.23

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437–448. doi:10.3102/0013189X07311286

Metfessel, N. S., & Sax, G. (1958). Systematic biases in the keying of correct responses on certain standardized tests. *Educational and Psychological Measurement*, *18*(1958), 787–790. doi:10.1177/001316445801800411

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of testwiseness. *Educational and Psychological Measurement*, *25*(3), 707–726. doi:10.1177/001316446502500304

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434–447. doi:10.1037/1082-989X.8.4.434

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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*(3/4), 305–314. doi:10.1024/1010-0652.21.3.305

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2016). *Afex: Analysis of factorial experiments*. Retrieved from https://CRAN.R-project.org/package=afex

Sparfeldt, J. R., Kimmel, R., Löwenkamp, L., Steingräber, A., & Rost, D. H. (2012). Not read, but nevertheless solved? Three experiments on pirls multiple choice reading comprehension test items. *Educational Assessment*, *17*(4), 214–232. doi: 10.1080/10627197.2012.735921

Srp, G. (1994). *Syllogismen.* Test: Software und Manual. Frankfurt/M, Germany: Swets Test Service.

Tamir, P. (1991). Multiple choice items: How to gain the most out of them. *Biochemical Education*, *19*(4), 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*(2), 198–206. doi:10.1111/j.1365-2923.2007.02957.x

Tomkowicz, J., & Rogers, W. T. (2005). The use of one-, two-, and three-parameter and nominal item response scoring in place of number-right scoring in the presence of test-wiseness. *Alberta Journal of Educational Research*, *51*(3), 200–215.

Torchiano, M. (2016). *Effsize: Efficient effect size computation*. Retrieved from https://CRAN.R-project.org/package=effsize

Willing, S., Ostapczuk, M., & Musch, J. (2015). Do sequentially-presented answer options prevent the use of testwiseness cues on continuing medical education tests? *Advances in Health Sciences Education*, *20*(1), 247–263. doi:10.1007/s10459-014-9528-2

An experimental validation of sequential multiple-choice tests

Martin Papenberg, Birk Diedenhofen and Jochen Musch

Word count (excluding abstract and references): 5282 words

Author Note:

Martin Papenberg, Birk Diedenhofen & Jochen Musch

Department of Experimental Psychology

University of Duesseldorf

Correspondence concerning this article should be addressed to:

Martin Papenberg

Department of Experimental Psychology

University of Duesseldorf

Universitaetsstrasse 1

Building 23.03

40225 Duesseldorf

Germany

E-mail: martin.papenberg@uni-duesseldorf.de

Abstract

The ability to recognize superficial cues pointing towards the solution ("testwiseness") may introduce construct-irrelevant variance to multiple-choice test scores. Presenting response options sequentially has been proposed as a potential solution to this problem. In an experimental validation, we determined the psychometric properties of a test based on the sequential presentation of response options. We created a strong validity criterion by providing participants with different levels of information on a domain about which they had no prior knowledge. Participants' knowledge was assessed using a traditional multiple-choice test or a sequential test. The sequential presentation of response options led to test scores that were as valid and reliable as multiple-choice test scores, but strongly decreased test-takers' probability of guessing the correct answer. We conclude that the sequential presentation of response options should be investigated more closely as a viable alternative to the traditional multiple-choice test format.

Keywords: discrete-option multiple-choice, sequential testing, experimental validation, test validity

An experimental validation of sequential multiple-choice tests

Multiple-choice (MC) testing is one of the most popular testing formats for assessing knowledge and aptitude. In its standard form, an MC item consists of a stem that poses the question to be answered, and a set of response options, one of which is the solution. Multiple-choice tests can be scored objectively and efficiently, and allow many areas to be covered in a short amount of time (Tamir, 1991). Although MC testing has been argued to be appropriate for many subject areas (Downing, 2006), critics continue to highlight the shortcomings associated with MC testing. Much of this criticism concerns the possibility of test-takers obtaining high test scores in the absence of substantive knowledge. Due to lucky guesses or testwiseness, test-takers may correctly solve MC items even if they do not have the knowledge the item writers intended to measure (Foster & Miller, 2009; Kubinger, Holocher-Ertl, Reif, Hohensinn, & Frebort, 2010). To address the problem of random guessing, researchers have proposed changing either the response format (e.g., Kubinger et al., 2010) or the scoring scheme of MC tests (e.g., Rowley & Traub, 1977).

Random guesses and individual differences in testwiseness tend to reduce the reliability and validity of MC tests because they add construct-irrelevant variance (Allan, 1992; cf. Messick, 1993; Millman, Bishop, & Ebel, 1965). Testwiseness is the ability to make advantageous use of superficial cues that point toward the solution (Gibb, 1964). Test-writers often inadvertently introduce such cues when creating items. For example, many test-writers tend to elaborate the correct answer more carefully than the distractors; therefore, the most verbose option is often the solution. Testwise test-takers may profit from using the “longest alternative” cue to infer the solution by comparing all response options on a superficial level. Several other testwiseness cues have also been identified (Millman, Bishop, & Ebel, 1965; Brozo, Schmelzer, & Spires, 1984; Thoma & Köller, 2018).

Individual differences in testwiseness may result in an unfair advantage for test-takers who are more testwise, but not necessarily more knowledgeable than their less testwise peers (Foster & Miller, 2009). As of yet, however, little research has addressed potential modifications of the MC test format that may help to reduce the influence of testwiseness. Instead, item writers are usually advised to closely follow guidelines of good item writing practices and to avoid including extraneous cues testwise respondents can capitalize on (Haladyna, 2004; Haladyna, Downing, & Rodriguez, 2002). Although there is no doubt regarding the usefulness of this advice, creating high-quality MC items is a challenging and time consuming task (Haladyna, 2004; Lee & Winke, 2013), and even in high-stake tests a considerable portion of MC items been found to contain cues that favor testwise students (Brozo et al., 1984; Hughes, Salvia, & Bott, 1991; Metfessel & Sax, 1958; Rogers & Bateson, 1991; Tarrant & Ware, 2008; Tomkowicz & Rogers, 2005).

To preclude participants from comparing response options and using testwiseness cues, Foster and Miller (2009) recommended a sequential test format, coining the term *discrete-option multiple-choice* (DOMC) for a test involving the sequential presentation of response options. DOMC items maintain all basic elements of traditional multiple-choice items; they also consist of a question stem and several response options. Unlike in traditional MC tests, however, the response options are presented one after another in random order (Foster & Miller, 2009). Therefore, test-takers cannot compare response options prior to answering. Instead, the correctness of each response option has to be evaluated in a separate yes/no decision. For practical reasons, the DOMC test format is usually delivered electronically in a computerized setting (Kubinger, 2009; Srp, 1994).

Unlike MC items, DOMC items are usually answered before all response 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; (b) the

solution has been wrongly rejected, or (c) a distractor has been wrongly accepted as the solution. In all of these cases, there is no need to present additional options to decide whether an item has been solved. Recent investigations have found that, in contrast to traditional MC items, less than half of the available response options are typically shown in a DOMC test, and the modal number of response options shown before an item presentation is terminated is just one (Papenberg, Willing, & Musch, 2017). As a result, three investigations have reported that DOMC testing reduces testing time by around 10-30% (Foster & Miller, 2009; Papenberg et al., 2017; Willing, Ostapczuk, & Musch, 2015). Foster and Miller (2009) convincingly argue that test security also profits from the reduced number of options that have to be conveyed in sequential testing.

A repeatedly reported finding is that DOMC items are more difficult than parallel MC items (Foster & Miller, 2009; Kingston, Tiemann, Miller, & Foster, 2012; Papenberg et al., 2017; Willing, 2013). A likely reason for this observation is that DOMC test-takers have to base all but their last decision on only a subset of the information that is available to MC test-takers. However, item difficulty is not in itself an indicator of item quality, provided that items cover a sufficiently large range of the ability being assessed. A more important potential reason for employing sequential tests is that they might reduce the influence of testwiseness because they prevent test-takers from comparing all response options prior to answering (Foster & Miller, 2009). This was indeed the case in several recent investigations; item difficulties between the MC and DOMC test formats differed more when they contained a larger number of testwiseness cues (Papenberg et al., 2017; Willing et al., 2015). On the one hand, these findings suggest that DOMC tests do not just make items more difficult; they specifically prevent the use of testwiseness cues. Hence, DOMC testing might be expected to decrease construct-irrelevant variance in test scores and thus improve test validity. On the other hand, little is known about the processes involved in answering DOMC items. Due to

the sequential decisions required, DOMC testing may add method-specific sources of variance that do not affect MC test scores (Papenberg et al., 2017). In the present study, we therefore investigated whether a change from the MC to the DOMC test format affects the psychometric properties of a test, particularly the validity of test scores.

Three correlational studies comparing item discrimination indices, reliability coefficients and criterion-related validity coefficients reported similar psychometric qualities for MC and DOMC tests (Foster & Miller, 2009; Kingston et al., 2012; Willing, 2013). Although these studies reported no difference in validity between MC and DOMC test scores, correlational studies suffer from interpretational problems that preclude strong conclusions. In a correlational validation study, a test is validated by determining its correlation with an external criterion, which is used as a proxy for an attribute of interest (e.g., the test-taker's knowledge). In such a design, it is not necessarily obvious what the test to be validated actually measures, and it is uncertain whether the criterion captures the same or a similar construct (Newton & Shaw, 2014). Moreover, correlations between a test and a criterion do not necessarily indicate the degree to which a test actually measures a construct because correlations capture all systematic variance that is shared by two measures. Such shared variance may be due to factors other than the construct under investigation if both the test and the criterion are contaminated with the same construct-irrelevant variance, such as test-takers' anxiety, motivation, or propensity to take risks (e.g., Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011; Rowley & Traub, 1977). The observation that MC and DOMC tests scores exhibit the same correlation with some external criterion therefore does not provide compelling evidence that the two test formats have comparable validity. Similarly, item discrimination and internal consistency also rise when item responses are influenced by additional factors, such as testwiseness; they therefore do not unequivocally reflect the degree to which an item or a test measures a construct of interest (Green, Lissitz, & Mulaik, 1977).

A promising way to overcome the problems associated with correlational validation is to employ an experimental approach (Diedenhofen & Musch, 2017; Erdfelder & Musch, 2006; cf. Poizner, Nicewander, & Gettys, 1978). To experimentally validate a knowledge test, participants with no prior knowledge in a domain are provided with previously unknown information. Different knowledge groups can be established by varying the amount of information provided to different groups. Then, the newly learned information can be assessed using the test or test format to be validated. One major advantage of this approach is that the amount of information available to test-takers is known and can be used as a strong external validity criterion. Moreover, the experimental approach to validation ensures that there are no systematic sources of variance other than the experimentally controlled levels of information that vary between conditions. The accuracy with which test scores reflect the participants' experimentally manipulated level of information therefore provides a direct and more reliable index of test validity than a mere correlation with some related test or construct.

To compare the validity of MC and DOMC test scores, we manipulated the information provided to test-takers in an online experiment. Participants were provided with a fictitious story that revealed either three, six, nine or twelve critical details, each of which solved a test question that was presented afterwards in either MC or DOMC format. To determine how well the four levels of information could be distinguished on the basis of participants' test scores, we conducted a one-dimensional linear discriminant analysis that tried to reassign test-takers to their experimental group on the basis of their test scores. The percentage of correct reassignments was then computed for both the MC and the DOMC tests in order to compare the validity of the two test formats.

Method

Design

The experiment had a 4 × 2 [level of information × test format] between-participants design. The *level of information* factor was manipulated by including 3, 6, 9 or 12 critical details in a short story, each of which solved one of the questions on a subsequent knowledge test. We refer to these four levels of information as very little, little, much, and perfect information hereinafter. The *test format* factor was varied by presenting test items in either MC or DOMC format. Each participant was randomly assigned to one of the eight cells resulting from the orthogonal crossing of the two experimental factors. A linear discriminant analysis was used to determine whether test-takers could be correctly reassigned to their level of information condition on the basis of their test scores.

Sample

An invitation e-mail was sent to members of a panel who had previously participated in online experiments for the Department of Experimental Psychology at the University of Duesseldorf. The study was started by 604 participants. Data from the 520 respondents (56% female) who finished the study and responded to all test items could be included in our analysis. The number of dropouts did not differ by test format, $\chi^2(1, N = 604) = 0.05$, $p = .822$, or information condition, $\chi^2(3, N = 604) = 1.47$, $p = .688$. The average age of participants was 38.31 years (*SD* = 12.98). The sample consisted of highly educated individuals: 69% reported having a college degree and 20% a university entrance qualification (German Abitur) as their highest educational attainment. Only 11% of the participants reported a level of education below a university entrance qualification.

Material

A fictitious short story was written for the purpose of the present study and presented to participants to induce different degrees of knowledge. Participants in each information

condition were provided with different amounts of information on critical details useful in answering the subsequent knowledge test, which consisted of twelve items constructed to measure knowledge of these critical details.

In the short story provided in the perfect information condition, 12 sentences contained a critical detail. Each of these 12 critical details solved one of the questions in the subsequent knowledge test. In the other information conditions, a randomly selected subset of only 3, 6 or 9 sentences contained critical details. If a critical detail was not disclosed, a different sentence was displayed reporting a detail that was irrelevant for the subsequent knowledge test. All experimental texts consisted of exactly 817 words to ensure that reading times did not differ between conditions, and all test items consisted of four response options – one solution and three distractors. An example of an item testing a critical detail read:

What form of sport did Luca enjoy?

a) *Handball*

b) Basketball

c) Football

d) Hockey

To provide test-takers with the critical detail necessary to answer this question, the following sentences were used: “Luca loved to do sports. He was particularly fond of *handball*.” When this critical detail was not provided in a given condition, the following sentences were displayed instead: “Luca loved to do sports. He was very proud of his sportsmanship.”

To decrease feelings of frustration that might otherwise occur due to unanswerable questions in the conditions with little knowledge, five additional dummy items were also presented and interspersed randomly among the critical items. These dummy items could be solved by all participants regardless of information condition because all texts contained the corresponding information; responses to these items were therefore not analyzed further.

Procedure

The study was conducted online and was administered using Unipark EFS Survey software (10.4, QuestBack, 2015). On the first page, participants were provided with introductory information on the study. Next, all participants were informed about both the MC and the DOMC test formats regardless of experimental condition, and no information was provided regarding whether they would later be tested in MC or DOMC format. We thus ensured that participants' encoding strategies did not vary as a function of the expected test format (cf. Finley & Benjamin, 2012). Both test formats were explained in written instructions and demonstrated using an example item. Participants were then instructed to carefully read a short story and attend to all details contained therein. The next page presented this story. The button to continue to the next page was not shown until 30 s after the presentation of the story to make sure that participants did not skip this page.

On the following 17 survey pages, the 12 critical items and 5 dummy items were administered in random order in either MC or DOMC format, depending on the experimental condition. Participants were told that some of the questions would be very difficult or even unsolvable on the basis of their knowledge. In the MC experimental condition, the question stem and all response options were shown simultaneously. When a DOMC item was presented, however, only the question stem and the first response option were shown. Test-takers then had to decide whether they accepted this option as the solution by clicking on one of two buttons labeled "true" or "false". When test-takers chose to reject a response option that was a distractor, the next response option was shown. No further response options were shown when a respondent either erroneously accepted a distractor, erroneously rejected the solution, or correctly accepted the solution. It was not possible to go back to previously presented options or items. After the experimental test was completed, test-takers were presented with a final set of five additional dummy items that could also be solved by all

participants regardless of information condition because all texts contained the information that was necessary to solve them. These additional dummy items were not analyzed, but were presented in a different test format than the experimental items to make sure that all participants encountered both test formats they were introduced to at the beginning. Thus, participants who answered the experimental items in MC format ended the testing session by answering dummy items in DOMC format, while participants who answered the experimental items in DOMC format ended the testing session by answering dummy items in MC format. The study concluded by asking participants to provide information on their age, sex, and educational attainment. In the end, participants were debriefed and informed of their test score as well as the correct solutions to all test items. Median study completion time was 12 min.

Data analysis

We used R (3.4.2, R Core Team, 2017) and the R-packages *afex* (0.18.0, Singmann, Bolker, Westfall, & Aust, 2016), *BayesFactor* (0.9.12.2, Morey & Rouder, 2015), *cocron* (1.0.1, Diedenhofen & Musch, 2016), *MASS* (7.3.47, Venables & Ripley, 2002), *papaja* (0.1.0.9492, Aust & Barth, 2016), and *propint* (0.2.12, Papenberg, 2017) in all analyses. To evaluate the effectiveness of the experimental manipulation, we conducted a 4 × 2 [information level × test format] between-subjects analysis of variance (ANOVA) on participants' test scores using generalized eta squared η_G^2 as an index of effect size. An alpha error level of .05 was applied for all significance tests.

To determine test validity, we reassigned participants to the four information levels on the basis of their test scores using a one-dimensional linear discriminant analysis (cf. Diedenhofen & Musch, 2017). The linear discriminant analysis established groups of similar test-takers by minimizing the variance of test scores within groups and maximizing the variance of test scores between groups. The reassignment was conducted twice: (a) for all

participants in the MC condition, and (b) for all participants in the DOMC condition. Each participant's reassigned information level was then compared to their experimentally induced information level, allowing us to test whether the MC or DOMC test scores were more useful in determining the level of information participants had originally been provided. We compared the proportion of correct classifications between the MC and the DOMC conditions in order to contrast the validity of the two competing test formats. We tested differences in classification accuracy using a classical χ^2-test and a Bayes factor for independence in contingency tables (Gunel & Dickey, 1974; Morey & Rouder, 2015). Bayes factors evaluate the relative evidence data provide for two competing hypotheses (Morey & Rouder, 2011). Often, a point null hypothesis is compared with an alternative hypothesis stating a range of possible effects (Kass & Raftery, 1995). The Bayes factor (BF_{10}) reflects the ratio of the evidence the data provide for preferring the alternative hypothesis that there is an effect over the null hypothesis that there is no effect. The inverse Bayes factor $BF_{01} = \frac{1}{BF_{10}}$ reflects the ratio of the evidence the data provide for retaining the null hypothesis that there is no effect over the alternative hypothesis that there is an effect. Thus, unlike classical significance tests relying on *p*-values, Bayes factors are able to provide evidence not only in favor of the alternative hypothesis, but also in favor of the null hypothesis (Dienes, 2014; Kass & Raftery, 1995). As alternative hypothesis, we used an uninformative default prior according to which all combinations of correct and incorrect classifications are equally likely a priori (see Jamil et al., 2017).

Results

Manipulation check

Figure 1 illustrates the distribution of test scores by information level and test format. The participants' level of information strongly affected their test scores, $F(3, 512) = 514.93$, $p < .001$, $\eta_G^2 = .75$, indicating that the experimental induction of knowledge was successful.

The DOMC test was found to be more difficult (M = 6.26, SD = 2.91) than the MC test (M = 7.36, SD = 2.70), $F(1, 512) = 76.70$, $p < .001$, $\eta_G^2 = .13$. The interaction of test format and information level was not significant, $F(3, 512) = 1.97$, $p = .117$, $\eta_G^2 = .01$.

– Insert Figure 1 about here –

Reliability

Reliability was determined by computing Cronbach's α, and was found to not differ between the DOMC (α = 0.72) and the MC test (α = 0.68), $\chi^2(1) = 0.77$, $p = .38$ in a significance test conducting using the software *cocron* (Diedenhofen & Musch, 2016).

Validity

Figure 1 illustrates the results of the linear discriminant analysis. A total of 74.14% of participants in the MC condition and 73.93% of participants in the DOMC condition were reassigned correctly, $\chi^2(1, N = 520) = 0.00$, $p = .956$, $BF_{01} = 10.40$. According to the Bayes factor, the data favor the null hypothesis in comparison to the alternative hypothesis by a factor of approximately 10, indicating that both test formats classified participants equally well.

Additional analyses

Testing time. It has repeatedly been reported that the DOMC format reduces testing time (Foster & Miller, 2009; Papenberg et al., 2017; Willing et al., 2015). To investigate whether this finding could be replicated in the present study, we computed the median time participants needed to respond to the 12 critical test items. The experimental validation design of the present investigation allowed us to assess how test-takers' information levels were related to their response times; this was not possible in previous studies in which the information available to test-takers was unknown. A 4 × 2 [information level × test format]

ANOVA on median response times showed that responses were faster when more knowledge was available, $F(3, 512) = 6.03$, $p < .001$, $\eta_G^2 = .03$ (see Figure 2). In contrast to the results of previous studies, test-takers responded faster to MC items (M = 7.46 s, SD = 2.20 s) than to DOMC items (M = 8.06 s, SD = 2.14 s), $F(1, 512) = 9.99$, $p = .002$, $\eta_G^2 = .02$. However, a significant interaction between test format and information level suggested that the difference between response times in DOMC and MC format depended on the test-takers' knowledge, $F(3, 512) = 2.68$, $p = .046$, $\eta_G^2 = .02$. Figure 2 illustrates the nature of this interaction: When test takers were given perfect information, MC items were processed faster (M = 6.63 s, SD = 2.09 s) than DOMC items (M = 7.66 s, SD = 1.81 s), $t(126.51) = -3.00$, $p = .003$, $d = -0.53$. When test takers were given much information, MC items were also processed faster (M = 7.21 s, SD = 2.12 s) than DOMC items (M = 8.17 s, SD = 1.79 s), $t(129.03) = -2.81$, $p = .006$, $d = -0.49$. However, when test takers were given little information, processing times did not differ significantly between MC items (M = 7.78 s, SD = 2.26 s) and DOMC items (M = 8.17 s, SD = 2.92 s), $t(118.53) = -1.47$, $p = .143$, $d = -0.26$. When test-takers had very little knowledge, there also was no significant difference between response times for MC (M = 8.25 s, SD = 2.05 s) and DOMC items (M = 7.95 s, SD = 1.78 s), $t(125.01) = 0.89$, $p = .376$, $d = 0.16$.

– Insert Figure 2 about here –

Correct responses by information availability. To investigate test-takers' responses more closely, we analyzed the extent to which participants made use of the information they were provided with in the short story. When a critical detail was presented to participants, they solved the corresponding test item correctly in 87.39% of all MC presentations, and in 80.30% of all DOMC presentations. To test this difference for statistical significance, we

computed a confidence interval for differences in proportions as proposed by Donner and Klar (1993, Method 2.1). This method takes the hierarchical structure of the data into account and nests the correctness of responses within test-takers. Because the 95% confidence interval of 7.09 percentage points ($= 87.39 - 80.30$) did not include 0 (95% CI [4.24, 9.94]), the difference was statistically significant. When participants were not provided with a given critical detail, they chose the correct answer in 17.31% of all MC presentations, but only 5.11% of all DOMC presentations, a difference of 12.20 percentage points (95% CI [9.46, 14.93]). Using the confidence interval approach proposed by Newcombe (2001) to determine the significance of the difference between the differences in solution rates, we found that solution rates for the two response formats differed more strongly when no information was provided than when information was provided ($12.20 - 7.09 = 5.11$, 95% CI [1.16, 9.06]). Thus, DOMC items were generally more difficult than MC items, and this effect was strongest when no information was available; uninformed DOMC test-takers hardly ever chose the correct solution.

Discussion

In a first experimental validation of the DOMC test format, we tested whether test scores based on a sequential presentation of response options are as valid as traditional MC test scores. We found that MC and DOMC test scores reflected the level of information that was available to respondents equally well, suggesting that the sequential presentation of response options results in test scores that are no less valid than MC test scores. This result is in line with previous studies finding that reliability, item discrimination (Foster & Miller, 2009; Kingston et al., 2012) and criterion-related evidence of validity (Willing, 2013) do not differ between MC and DOMC tests. Unlike previous studies, however, we conducted an experimental validation. Methodologically, we therefore provide a particularly strong contribution to the still small body of literature on the validity of the DOMC format. Because

we experimentally induced different levels of information among participants, we were able to compare MC and DOMC test scores with a particularly reliable external validation criterion. In a linear discriminant analysis, we found that both test formats performed equally well; they both correctly determined the group membership of 74% of participants. A Bayes factor comparing classification performance between the MC and the DOMC condition confirmed this result and provided strong evidence for the null hypothesis of no difference between test formats. Thus, taken together, the present results strongly suggest that the DOMC test format is no less valid than the MC test format in assessing individual differences in knowledge.

When no information was provided at all, participants in the MC condition were not able to identify the solution more often than would be expected assuming a random guessing strategy. In fact, their performance (17% correct) was even below the 25% level that would be expected by chance, indicating that the experimental items contained well-functioning distractors that were more attractive to uninformed test-takers than the solution. Guessing performance was even lower in the DOMC test format (5%). When test-takers had not been provided with a critical detail, they were much less likely to select the correct answer in the DOMC than in the MC test format. This finding provides evidence for the notion that the DOMC test format is particularly successful in preventing successful guessing by uninformed test-takers, making it attractive as an alternative to the traditional MC format. However, further studies should explore the response processes involved in DOMC testing to achieve a better understanding of test-takers' decision-making and guessing strategies in sequential testing. The present findings suggest that the DOMC format tends to make test-takers rather conservative, as they were more reluctant to accept early response options than would be optimal from a normative point of view. This may have put respondents who were too conservative at a disadvantage. Potential individual differences in willingness to accept early

response options – which may be independent of individual differences in knowledge – should therefore be investigated more closely. If individual differences in response style affect DOMC test scores, they may pose a threat to the validity of DOMC testing.

The present study may have precluded a more favorable outcome for the DOMC test format because we successfully made sure that none of the items contained any cues to their solution. This is not necessarily representative of real MC examinations used in practice; studies by Brozo et al. (1984) and Tarrant and Ware (2008) found that between 28-75% of the MC items presented in exams contained cues to their solution that could be exploited by testwise examinees. Given that DOMC testing has been shown to reduce construct-irrelevant variance due to testwiseness (Willing et al., 2015; Papenberg, et al., 2017), DOMC tests are likely to outperform MC tests in terms of test score validity if items contain cues to their solution that can more easily be exploited when answer options are presented simultaneously rather than sequentially.

An important question is whether the technically more demanding DOMC format should be adopted even if it does not improve reliability and validity compared to the MC test format. The present results show that uninformed test-takers are far less likely to guess the solution when response options are presented sequentially. Susceptibility to guessing is a major disadvantage of the MC test format (Kubinger et al., 2010), a problem that may be overcome by using the DOMC format. Another potentially interesting feature of the DOMC test format is that it provides a convenient way to make items more difficult if desired in a given application. Test security is also likely to profit from the use of the DOMC test format because not all response options have to be revealed to test-takers (Foster & Miller, 2009).

Previous research indicated that testing time may be reduced considerably when response options are presented sequentially due to the application of stopping rules. Reduced testing times would then be another advantage of the DOMC test format (Foster & Miller,

2009; Willing et al., 2015; Papenberg, et al., 2017). However, the results of the present study indicate the opposite. Item responses were faster in the MC test, and this effect was most pronounced when test-takers had perfect information; there were no differences in response times when test-takers had little information. A possible explanation for this unexpected finding is that our items differed in length from items used in previous studies. In the present study, response options often consisted of only one word, and the maximum number of words per option was nine. In a previous study by Papenberg et al. (2017), many response options consisted of sentences containing more than 20 words. As the DOMC test format reduces the number of response options that have to be presented to test-takers, a larger number of words per option will result in a larger total savings in the number of words test-takers must process. It is therefore likely that testing time is reduced most when the DOMC test format is employed for response options that are rather verbose. The present results suggest that if response options consist of only a few words that can be processed quickly, the MC test format may result in shorter testing times than the DOMC test format. When response options are short, the ratio of the length of the question stem to the length of the response options becomes larger. When short response options are used, the additional time needed to make repeated decisions in DOMC items may outweigh the reduction in testing time caused by the application of stopping rules in DOMC tests, provided that the question stem has to be read in both MC and DOMC items. Considering that the present study is the first to find reduced testing times for an MC test compared to a DOMC test, future research should investigate potential moderators of the time efficiency of the MC and the DOMC test formats more closely.

To conclude, the existing body of research on the DOMC test format is still rather small and needs to be expanded. The present study provides the first experimental demonstration that DOMC tests are no less reliable and valid than MC tests. However, further validation

studies should be conducted to test whether the present results can be generalized to other tests and domains of knowledge, as well as to other populations of test-takers. A general limitation that should also be addressed in future studies is that test-takers' decision processes during DOMC testing are not yet well understood, despite the fact that understanding response processes has been identified as a key variable in test validation (Borsboom, Mellenbergh, & Heerden, 2004). Future research should therefore strive to widen our understanding of test-takers' answering processes in DOMC testing.

References

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

Aust, F., & Barth, M. (2016). *Papaja: Create APA manuscripts with rmarkdown*. Retrieved from https://github.com/crsh/papaja

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

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 & Learning Assistance. ERIC database (ED240928). Retrieved from http://eric.ed.gov/?id=ED240928

Diedenhofen, B., & Musch, J. (2017). Empirical option weights improve the validity of a multiple-choice knowledge test. *European Journal of Psychological Assessment, 33*, 336-344. doi: 10.1027/1015-5759/a000295

Diedenhofen, B., & Musch, J. (2016). cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. *International Journal of Internet Science, 11*, 51-60.

Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 1–17.

Donner, A., & Klar, N. (1993). Confidence interval construction for effect measures arising from cluster randomization trials. *Journal of Clinical Epidemiology*, *46*, 123–131.

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: Erlbaum.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences, 108*, 7716–7720.

Erdfelder, E., & Musch, J. (2006). Experimental methods of psychological assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 205–220). Washington, D.C.: American Psychological Association.

Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: Evidence from the test-expectancy paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 632–652.

Foster, D., & Miller, H. (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). *Test-wiseness as secondary cue response*. (Doctoral dissertation) No. 64-7643. Stanford University, Ann Arbor, MI: University Microfilms.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, *37*, 827–838.

Gunel, E., & Dickey, J. (1974). Bayes factors for independence in contingency tables. *Biometrika*, *61*, 545–557.

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

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*, 309–333.

Hughes, C. A., Salvia, J., & Bott, D. (1991). The nature and extent of test-wiseness cues in seventh- and tenth-grade classroom tests. *Assessment for Effective Intervention*, *16*, 153–163.

Jamil, T., Ly, A., Morey, R. D., Love, J., Marsman, M., & Wagenmakers, E.-J. (2017). Default "Gunel and Dickey" bayes factors for contingency tables. *Behavior Research Methods*, *49*, 638–652.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

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

Kubinger, K. D. (2009). Psychologische Computerdiagnostik [Computerized diagnostics in psychology]. *Zeitschrift Für Psychiatrie, Psychologie und Psychotherapie*, *57*, 23–32.

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment*, *18*, 111–115.

Lee, H., & Winke, P. (2013). The differences among three-, four-, and five-option-item formats in the context of a high-stakes english-language listening test. *Language Testing*, *30*, 99–123.

Metfessel, N. S., & Sax, G. (1958). Systematic biases in the keying of correct responses on certain standardized tests. *Educational and Psychological Measurement*, *18*, 787–790.

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

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419.

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Newcombe, R. G. (2001). Estimating the difference between differences: Measurement of additive scale interaction for proportions. *Statistics in Medicine*, *20*, 2885–2893.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. London: Sage.

Papenberg, M. (2017). *Propint: Testing for interactions in proportions*. Retrieved from https://github.com/m-Py/propint

Papenberg, M., Willing, S., & Musch, J. (2017). Sequentially presented response options prevent the use of testwiseness cues in multiple-choice testing. *Psychological Test and Assessment Modelling*, *59*, 245–266.

Poizner, S. B., Nicewander, W. A., & Gettys, C. F. (1978). Alternative response and scoring methods for multiple-choice items: An empirical study of probabilistic and ordinal response modes. *Applied Psychological Measurement*, *2*, 83–96.

Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement, 14*, 15–22.

QuestBack. (2015). *Unipark EFS survey*. Retrieved from http://www.unipark.de

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rogers, W. T., & Bateson, D. J. (1991). The influence of test-wiseness on performance of high school seniors on school leaving examinations. *Applied Measurement in Education*, *4*, 159–183.

Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement*, *14*, 15–22.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2016). a*fex: Analysis of factorial experiments*. Retrieved from https://CRAN.R-project.org/package=afex

Srp, G. (1994). *Syllogismen.* Test: Software und Manual. Frankfurt/M, Germany: Swets Test Service.

Tamir, P. (1991). Multiple choice items: How to gain the most out of them. *Biochemical Education*, *19*, 188–192.

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

Thoma, G.-B., & Köller, O. (2018). Test-wiseness: Ein unterschätztes Konstrukt? [Test-wiseness: an underestimated construct?]. *Zeitschrift für Bildungsforschung, 8*, 1–18.

Tomkowicz, J., & Rogers, W. T. (2005). The use of one-, two-, and three-parameter and nominal item response scoring in place of number-right scoring in the presence of test-wiseness. *Alberta Journal of Educational Research*, *51*, 200–215.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Willing, S. (2013). *Discrete-option multiple-choice: Evaluating the psychometric properties of a new method of knowledge assessment*. (Doctoral dissertation, Heinrich-Heine University, Duesseldorf, Germany). Retrieved from http://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=27633.

Willing, S., Ostapczuk, M., & Musch, J. (2015). Do sequentially-presented answer options prevent the use of testwiseness cues on continuing medical education tests? *Advances in Health Sciences Education*, *20*, 247–263.

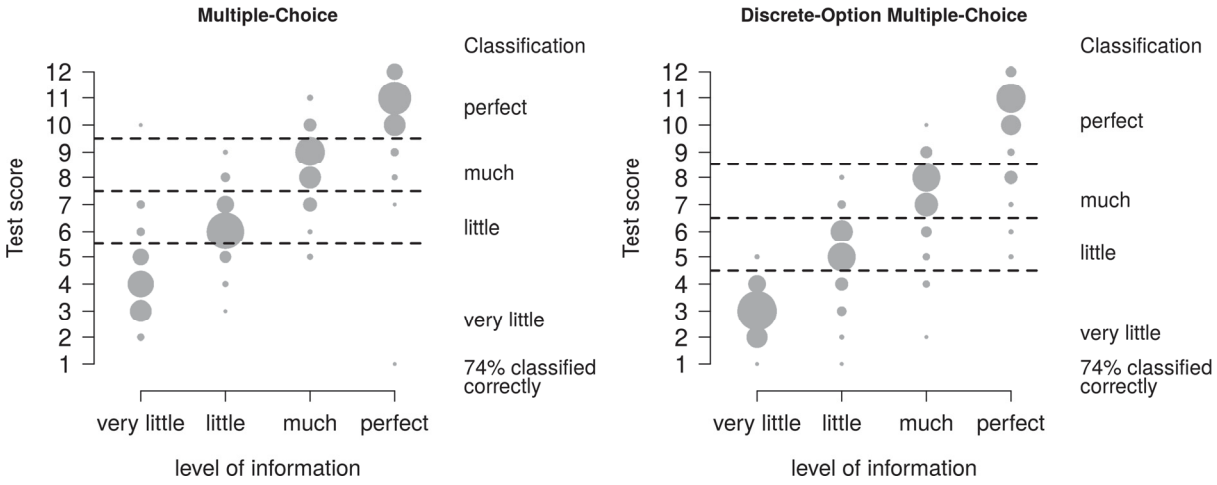

Figure 1. Distribution of test scores by information level and test format. The size of the circles represents the number of participants achieving the respective test score.

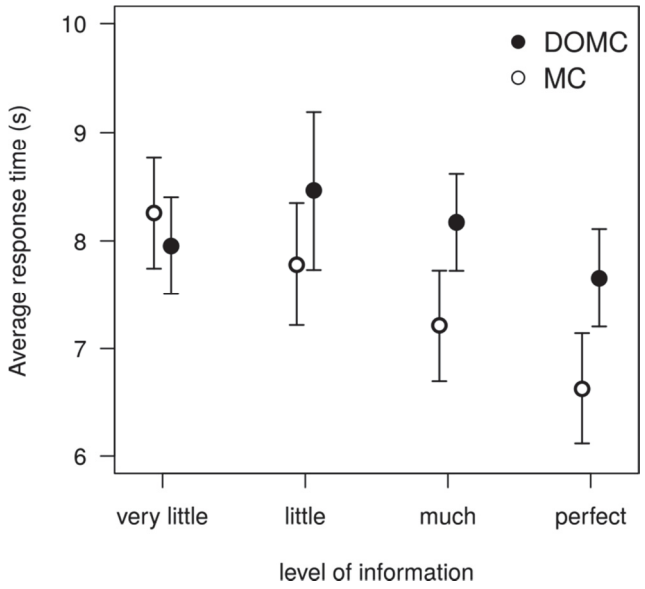

Figure 2. Average median response time for critical test items as a function of test format and level of information. Error bars indicate 95% confidence intervals.

An investigation of sequential response behavior in discrete-option multiple-choice knowledge tests

Martin Papenberg and Jochen Musch

Word count (excluding abstract and references): 11700 words

Author Note:

Martin Papenberg & Jochen Musch

Department of Experimental Psychology

University of Duesseldorf

Correspondence concerning this article should be addressed to:

Martin Papenberg

Department of Experimental Psychology

University of Duesseldorf

Universitaetsstrasse 1

Building 23.03

40225 Duesseldorf

Germany

E-mail: martin.papenberg@uni-duesseldorf.de

Abstract

To identify the solution in multiple-choice items, testwise individuals may use cues based on a comparison of all response options. Discrete-option multiple-choice (DOMC) testing is based on the sequential presentation of response options and has been shown to reduce construct-irrelevant variance due to testwiseness. Study 1, however, found evidence for individual differences in DOMC response style. Some test-takers were more reluctant to accept early DOMC response options than others, independently of their knowledge. A computer simulation based on a model of response behavior in sequential tests (MORBIST; Study 2) and a laboratory experiment (Study 3) confirmed that high acceptance reluctance is causally related to better DOMC test scores, thus demonstrating that these contain construct-irrelevant variance.

Keywords: discrete-option multiple-choice, response bias, signal detection theory, validity, individual differences, construct-irrelevant variance

An investigation of sequential response behavior in discrete-option multiple-choice knowledge tests

In multiple-choice (MC) tests, response alternatives are usually presented simultaneously. Discrete-option multiple-choice (DOMC) tests, which are based on the sequential presentation of response options, have been proposed as a potential improvement over MC tests (Foster & Miller 2009). Like MC tests, DOMC tests present a question stem and a fixed set of response options. One of these options is the solution respondents must identify; the non-correct options serve as distractors. Figure 1 illustrates the presentation of a DOMC item.

– Please insert Figure 1 about here –

Unlike in MC tests, the response options are presented sequentially and one at a time. Test-takers therefore cannot compare them with regard to their plausibility, and must assess the correctness of each response option separately. Additional response options are revealed until an item is answered either correctly or incorrectly. No further options are presented after (a) the solution has been identified as being correct, (b) the solution has been wrongly rejected or (c) a distractor has wrongfully been accepted. For each item, one point is awarded if the correct solution has been identified. Since fewer response options become known to respondents, DOMC tests preserve test security better than MC tests (Foster & Miller, 2009). However, despite the potential advantages of DOMC testing compared to traditional MC testing, a thorough evaluation of the psychometric properties of the DOMC test format is still lacking (Lindner, Strobel, & Köller, 2015).

An important potential advantage of DOMC testing is that because response options are presented one at a time, test-takers must independently assess the correctness of each response option without taking a look at the remaining response options first. In traditional MC testing, test-takers are allowed to compare all response options prior to answering, and

may therefore simply pick the most plausible option even if they are not sure of its correctness. Rendering it impossible to compare all response options prior to answering forces test-takers to make an absolute judgment instead of a relative one. It has repeatedly been shown that this results in higher difficulty for DOMC items, as they must be answered on the basis of less additional information than is available when all response options are presented (Foster & Miller, 2009; Kingston, Tiemann, Miller & Foster, 2012).

In MC items, testwise individuals may make use of superficial cues contained in one or more answer options to gather hints pointing towards the solution (Gibb, 1964; Millman, Bishop, & Ebel, 1965; Tarrant & Ware, 2008; Lindner, Strobel, & Köller, 2015; Thoma & Köller, 2018). Testwiseness is conceptually independent of test-takers' knowledge, and recognizing superficial cues may help test-takers identify the solution even in the absence of any substantive knowledge (Millman et al., 1965). Individuals who are not more knowledgeable, but more testwise and better at exploiting weaknesses in item writing may therefore obtain an unfair advantage over less testwise individuals (Lindner, Strobel, & Köller, 2015). In a comprehensive analysis of 1,220 MC items collected from 43 college exams, Brozo, Schmelzer, and Spires (1984) found that about 44% of items contained a cue that could be exploited by testwise examinees. Tarrant and Ware (2008) also found that between 28 – 75% of MC test items used in high-stakes assessments contained flaws, most of which favored testwise students. It has been argued and shown that the sequential presentation of response options successfully reduces the use and thus also the impact of construct-irrelevant testwiseness compared to MC tests (Foster & Miller, 2009; Kingston, Tiemann, Miller, & Foster, 2012; Willing, Ostapczuk, & Musch, 2015; Papenberg, Willing, & Musch, 2017).

If DOMC tests control for construct-irrelevant variance due to testwiseness better than MC tests, this should arguably increase the validity of DOMC tests. However, previous

comparisons of the MC and DOMC test formats found comparable item discrimination indices, internal consistencies, and concurrent validities (Foster & Miller, 2009; Kingston et al., 2012; Willing, 2013). While this may be considered surprising, little is actually known about the influence of construct-irrelevant response biases on MC and DOMC test scores. On the one hand, variance due to testwiseness may not necessarily contaminate the result of MC tests if good item writing practices are followed (Haladyna, 2004). On the other hand, while DOMC test results successfully suppress construct-irrelevant variance due to testwiseness (Willing et al., 2015), they may be contaminated by other sources of construct-irrelevant variance that have not yet been identified. The present study investigates this latter possibility. By examining the DOMC response process more closely, we follow recent recommendations stressing the importance of investigating the internal processes that precede the selection of answers to knowledge items (e.g. Borsboom, Mellenbergh, & Heerden, 2004; Gorin, 2007; Lindner, Strobel, & Köller, 2015; Lissitz & Samuelsen, 2007).

Based on a new causal model of how test-takers answer DOMC items, we generate hypotheses on the potential detrimental influence of construct-irrelevant sources of variance on DOMC test scores. For a more thorough evaluation of the validity of the DOMC test procedure, we propose and use MORBIST as a *model of response behavior in sequential tests*. MORBIST is based on signal detection theory, the standard model used to describe decision-making in many areas of memory research (Macmillan & Creelman, 2005; Pastore, Crawley, Berens, & Skelly, 2003). We show that MORBIST can successfully model test-takers' response behavior in DOMC tests while simultaneously accounting for the individual differences in response style that are at the center of the present investigation.

Using signal detection terminology, MORBIST considers the presentation of a correct answer a signal and the presentation of a distractor noise (cf. Stanislaw & Todorov, 1999). In signal detection theory, response decisions are made in two consecutive phases, an evaluation

phase and a decision phase. In the evaluation phase, the plausibility of the current option is assessed according to the strength of the evidence associated with this option (Pastore et al., 2003). A higher perceived strength of evidence increases the probability that a response option will be accepted as the solution. The degree to which distractors and solutions invoke different strengths of evidence is used as an index of the respondents' degree of knowledge (*d'*). In the decision phase – after an option has been evaluated – respondents have to decide whether enough evidence has been collected to justify the acceptance of the current response option. To arrive at this decision, a response criterion is used to determine whether the available evidence suffices to accept the current option. Respondents who require strong evidence to accept an option employ a conservative response criterion; respondents who require less evidence employ a more liberal response criterion (Stanislaw & Todorov, 1999).

MORBIST expects that the sequential response mode of DOMC testing influences the response criteria test-takers employ. In a DOMC item, if no previous option was the solution, the probability of a response option being the solution increases with increasing serial position. For example, the first option in a set of five response options is the solution with a probability of only $p = 1/5 = .20$. The probability is higher ($p = 1/4 = .25$) for the second option, because the second option is presented only if the first option was not the solution. Therefore, after a test-taker has successfully rejected a distractor, each further option has an increasingly higher probability of being the solution. It has been shown that respondents can and do adapt their response criteria to varying target probabilities (Buchner, Erdfelder, & Vaterrodt-Plünnecke, 1995). MORBIST therefore assumes that test-takers generally employ a more liberal response criterion with increasing option position to accommodate the fact that the number of options left is gradually decreasing.

The most important and critical assumption of MORBIST is that test-takers might systematically differ in the degree to which they are inclined to accept early response options

over and above what is warranted due to normative reasons and individual differences in knowledge. We call this potential individual response bias, which is independent of test-takers' knowledge, *acceptance reluctance* or *Festlegungszögern* (in German). High acceptance reluctance is characterized by the application of a rather conservative response criterion for early options and thus a preference for choosing later response options. In contrast, low acceptance reluctance is characterized by the application of a more liberal response criterion, which favors the acceptance of early options. In other words, MORBIST accommodates for the fact that some test-takers may have a tendency to accept early options to avoid missing the solution, whereas others might be more willing to wait longer in hope of a better fit. Such individual differences in response style need not be related to the trait under investigation, and if they occur, they may introduce bias to the measurement (Cronbach, 1950).

In a very early contribution, Wiley and Trimble (1936) noted that individual differences in response confidence across several achievement tests could be assessed with higher reliability than the measures of achievement themselves. From this observation, they concluded that "some personality trait, or traits, is operative". Thorndike (1938) also investigated the influence of a "checking threshold" on a test of maturity. In this test, respondents were free to mark statements either once or twice, depending on the degree to which they rejected the respective statement. Thorndike found that the number of item checks could be measured more consistently than the actual measure of maturity, providing evidence for consistent but construct-irrelevant individual differences in response thresholds. Ingham (1970) also explored the individual stability of the response criterion in yes/no decision tasks in the domain of acoustic sensation. Respondents were asked to indicate whether they had heard a sound or not. Ingham found that the applicable response criterion was stable both within (r = .71 – .82) and between sessions (r = .45 – .59); it could thus be shown to reliably

differ between respondents. Kantner and Lindsay (2012) recently investigated the stability of response thresholds in the domain of old/new recognition memory in a similar way. In several experiments, they observed high correlations between response thresholds across experimental sessions. They therefore concluded that "some people require more memory evidence than do others before they are willing to call an item 'old' (p. 1163)".

To thoroughly assess the DOMC testing procedure, it is important to understand how individual differences in the tendency to employ a conservative response criterion may affect item responses independently of individual differences in knowledge. To investigate this question more closely, we first report a correlational study that established that individual differences in acceptance reluctance do indeed exist (Study 1). Because individual differences in acceptance reluctance may add construct-irrelevant variance to DOMC instruments, we then report the results of a computer simulation that examined whether individual differences in the proclivity to accept early response options systematically influence DOMC test scores (Study 2). We predict and find that employing a high threshold and choosing later response options in the simulation leads to higher test scores. In Study 3, we report an experiment with real participants aimed at providing a final empirical test of whether construct-irrelevant differences in response threshold causally influence DOMC test scores. To this end, we test whether experimentally inducing a high or low readiness to accept early response options systematically leads to high or low scores in DOMC tests, respectively.

For this investigation, it was important to obtain a measure of acceptance reluctance. To create this measure, we reasoned that acceptance reluctance is arguably strongest in situations of high uncertainty. When respondents know an answer with high certainty, their response criterion is less important for the decision process (cf. Berg, 1955; Cronbach, 1946, 1950; Ingham, 1970). We therefore created an uncertain decision situation that allowed us to measure individual differences in the willingness to accept early answers. We did this by

presenting questions that had no correct answer at all alongside general knowledge items for which there was a correct answer. No-solution items have previously been used to obtain a measure of response bias in the context of measuring overconfidence in a vocabulary test (Koehler, 1974). In no-solution items, responses cannot be based on knowledge because non-existing solutions cannot be known to the respondents. If test-takers are not made aware of the presence of no-solution items, however, they can be expected to process such questions like other test items. This produces an uncertain response situation in which the decision process is unaffected by individual differences in knowledge; in SDT terms, *d'* is set to zero. We used the number of options test-takers want to be shown in no-solution items before accepting an answer as a knowledge-free individual-level measure of acceptance reluctance.

In Study 1, we explored a) whether acceptance reluctance can be measured reliably using no-solution items, b) whether acceptance reluctance is correlated with DOMC test scores, and c) whether acceptance reluctance can be used to predict test scores in DOMC tests over and above what can be predicted on the basis of test-takers' knowledge. If acceptance reluctance influences DOMC test scores independently of test-takers' knowledge levels, the validity of DOMC test scores would be called into question. We also measured several personality traits that might be associated with response thresholds to explore the nomological network of acceptance reluctance and pinpoint potential aspects of personality that might be associated with the use of a conservative response threshold. As a secondary purpose of Study 1, we further explored the general properties of the DOMC response format. This also allowed us to use the results of Study 1 as empirical benchmarks for the MORBIST computer simulation in Study 2.

Study 1

Method

Sample. Participants were 236 members of an online research panel who had previously consented to receive invitations to online studies conducted by the University of Duesseldorf. Only the data of the 196 participants (56% female) who finished the test were used for analysis. The participants' average age was 34.23 years ($SD = 10.98$). All participants indicated German as their native language. The majority of the sample consisted of highly educated individuals; 46% percent reported a bachelor, master or equivalent degree as their highest educational attainment, while another 46% reported a university entrance qualification (German Abitur). Only 8% reported not holding one of the above two educational qualifications.

Materials.

No-solution items. No-solution items posed a question to which no solution exists. Because none of the available response options solved the no-solution items, these items could not be answered correctly. However, the number of options participants decided to look at in no-solution items before eventually accepting an option was used as a knowledge-free measure of acceptance reluctance. Test-takers who accepted a later response option were considered to exhibit high acceptance reluctance. To measure acceptance reluctance unobtrusively, it was important to make sure that it was not obvious that some of the questions did not have any solution. For this reason, the six no-solution items were interspersed randomly among 18 DOMC items taken from an existing knowledge test for which solutions were available (BOWIT; Hossiep & Schulte, 2008). Care was taken to make all of the five response options in no-solution items appear equally plausible. To this end, no-solution items such as the following were used:

Who discovered the Reynald Islands?

a. Juan Fernández Ladrillero

b. Hernando de Alarcón

c. Ruy López de Villalobos

d. Sebastiàn Vizcaíno

e. Garcia Jofre de Loaísa

The Reynald Islands do not exist, although this could not be known to the respondents because there are an immense number of islands in the world. The Indonesian archipelago alone, for example, consists of more than 16,000 islands (Wikipedia, 2017). Therefore, test-takers were not able to recognize that no-solution items did not contain a solution on the basis of their knowledge. As a result, the number of options they rejected before eventually accepting one could be used as a measure of their acceptance reluctance. The German version and an English translation of the six no-solution items are available on https://osf.io/637x8/.

Knowledge items. The 18 DOMC knowledge items served two purposes. First, they were used to investigate the relationship between acceptance reluctance and DOMC test scores. Second, they camouflaged the no-solution items interspersed among the DOMC items. The DOMC items were selected from the BOWIT (Bochumer Wissenstest = Bochum Test of General Knowledge; Hossiep & Schulte, 2008), a well-validated German-language general knowledge test that covers similar content as the six newly-created no-solution items. The original BOWIT items contain five options, with option 5 always being "none of the above is true." This option, which never was the solution in the items we selected, is not suitable for DOMC testing because of the sequential presentation of response options. It was therefore removed from all items and replaced with an additional plausible option to make sure that both the 18 BOWIT and the six no-solution items consisted of five response options each. The final total test thus consisted of 18 + 6 = 24 items.

To obtain an additional, external measure of participants' general knowledge unaffected by the method-specific variance we expected to contaminate DOMC scores, we selected another 10 BOWIT items and presented them in traditional MC format after participants had answered the DOMC items. The option "none of the above is true", which was not the solution to any of these items, was again removed to make the format of the MC items compatible with that of the DOMC items.

Personality scales. To better understand the psychological nature of acceptance reluctance and establish a nomological web of its correlates, we measured several personality variables and response biases in order to investigate their potential relationship with acceptance reluctance. We wanted to explore whether acceptance reluctance is best conceived of as a specific trait limited to DOMC testing, or should rather be considered a more general trait overlapping with other personality traits.

Risk-taking propensity. Decisions in DOMC tests are made in an uncertain situation of incomplete information in which test-takers permanently face the risk of either prematurely accepting a distractor or missing the solution. More risk-seeking respondents might be willing to wait longer for a better response option, whereas risk-averse participants might prefer to accept an early plausible option. Risk-taking propensity can be operationalized in different ways; we decided to include both a behavioral and a self-report measure of risk-taking propensity. To this end, we presented the single question advocated by Dohmen et al. (2011) as a short and valid measure of risk-taking propensity ("How do you see yourself: are you generally a person who is fully prepared to take risks, or do you try to avoid taking risks?"). Participants were asked to indicate their answer to this question on a 7-point Likert scale ranging from 1 ("not at all willing to take risks") to 7 ("very willing to take risks"). A short version of the balloon analogue risk task (BART) was used as an additional behavioral measure of risk-taking propensity (Lejuez et al., 2002). In the BART, respondents

are asked to pump up a balloon. Each click on a button pumps up the balloon and generates money in a virtual bank. Participants accumulate more money as the balloon gets larger. However, all money is lost if the balloon explodes, which can happen after any pump. To avoid the risk of explosion, respondents can decide at any point to finish a run by cashing in the money they have already accumulated instead of pumping up the balloon further. This ensures that the money for this run can no longer be lost due to an exploding balloon. In the original BART, the risk propensity score was computed as the average number of pumps in all runs in which the balloon did not explode. Traditionally, trials in which the balloon exploded were not considered in determining the risk propensity score. This procedure, however, leads to artificially low scores for very risk-prone participants (Pleskac, Wallsten, Wang, & Lejuez, 2008). Therefore, for the present investigation, we created a short online version of the BART consisting of six runs using an improved BART procedure. Following a proposal by Pleskac et al. (2008), we did not record the number of actual pumps, but the number of intended pumps irrespective of whether they led to an explosion. We were thus able to use all runs in computing the risk propensity score, with no runs lost due to exploding balloons (Pleskac et al., 2008). To record the intended pumps independent of explosion status, we recorded and visualized the intended number of pumps by letting participants increase the air pressure in an intermediate air reservoir, which was only later released into the balloon. The air pressure in the intermediate reservoir could be increased up to 30 times by clicking a button. Participants were told that the balloon would explode after 15 clicks on average. Afterwards, when the reservoir was released, the balloon was set to explode after a random number of air pressure increases. Each increase in air pressure generated units of a virtual currency, and participants were instructed to attempt to maximize their earnings. The average number of reservoir pumps across the six trials was used as the participant's risk propensity score, and could range from 1 to 30.

Need for cognition is the tendency to engage in and enjoy effortful cognitive activity (Cacioppo, Petty, Feinstein, & Jarvis, 1996). We surmised that persons high in need for cognition might be reluctant to accept early response options because they would want to see more options before making a decision. To obtain a short measure of need for cognition, we selected the six best discriminating items (three of positive and three of negative polarity) from the German Need for Cognition Scale published by Bless, Wänke, Bohner, Fellhauer, and Schwarz (1994). One example item read: "I prefer my life to be filled with puzzles that I must solve." Negatively polarized items were inverted, and need for cognition scores were computed as the average of the six items, which were presented on a 7-point Likert scale.

Perfectionism. A perfectionist predisposition might stimulate test-takers to collect as much information as possible before making a decision. Therefore, perfectionist test-takers might wish to see more response options before committing to one. To assess individual differences in the tendency to strive for perfection, we used the striving for perfection scale, which consists of five items on a 6-point Likert scale (Stoeber, Otto, Pescheck, Becker, & Stoll, 2007). One example item read: "I strive to be as perfect as possible". The score on the striving for perfectionism scale was computed as the average of the five item scores.

Overclaiming is the tendency to exaggerate one's knowledge. To assess this tendency, Paulhus, Harms, Bruce, and Lysy (2003) developed an overclaiming test that asks participants to indicate how familiar they are with some rare words. Unbeknownst to the respondents, some of the words in the overclaiming test are fictitious and therefore cannot possibly be known. The tendency to overclaim is operationalized as the degree to which participants nevertheless claim knowledge of these non-existent words. We surmised that overclaiming participants, who readily claim to be knowledgeable on an unknown matter, might also accept an earlier response option in a DOMC test. To assess overclaiming, we employed a German variant of the overclaiming test (VOC-T) consisting of 12 real words

(e.g., *Platine*) and three fictitious words (e.g., *Enklivie*; Ziegler, Kemper, & Rammstedt, 2013). Participants indicated their familiarity with each word on a 7-point scale, with higher values indicating greater familiarity. To compute an overclaiming index, hit and false alarm rates were calculated. The hit rate was the proportion of the twelve real words with which the participants indicated familiarity, and the false alarm rate the proportion of the three fictitious words with which test-takers indicated familiarity. Following the procedure detailed in Paulhus and Harms (2004), the calculation of hit and false alarm rates was repeated for each possible cutoff point on the 7-point scale. For example, when the cutoff was set between the third and fourth rating scale categories, a rating greater than three for a real word was recorded as a hit, and a rating greater than three for a fictitious word was recorded as a false alarm. The resulting six pairs of hit and false alarm rates were then averaged to obtain a global hit and false alarm rate for each participant. The sum of the global hit and false alarm rate corresponds to an overall yes rate, and was used as an index of overclaiming (cf. Paulhus & Harms, 2004).

Self-ratings of knowledge and intelligence. The thresholds employed in DOMC tests may be associated with knowledge or intelligence. For exploratory reasons, we therefore collected self-ratings of knowledge and intelligence using single items that asked respondents to provide an estimate of the percentage of the population they expected to have more general knowledge or be more intelligent than themselves. A higher percentage was considered indicative of lower self-rated knowledge or intelligence.

Procedure. All study instruments were implemented as a web survey using the software Unipark EFS Survey (QuestBack, 2017). Participants started the study by clicking on a link in an email invitation. The first page welcomed the participants and obtained their informed consent to participate. On the next page, the short version of the BART was introduced. Two test runs familiarized participants with the task, followed by six runs for

which responses were recorded and analyzed. Afterwards, the following measures were presented in the following order: (1) need for cognition, (2) striving for perfectionism, (3) overclaiming, (4) DOMC knowledge and no-solution items, and (5) MC knowledge. Prior to taking the DOMC test, participants were informed of the rules of the DOMC test format. The study concluded with questions pertaining to demographic variables and with self-ratings of general knowledge, intelligence, and risk-taking propensity. In the end, participants were debriefed and thanked for their participation, and received feedback on their performance in the knowledge tests.

Results

We used the statistics program R (3.4.3, R Core Team, 2016) and the R packages *afex* (0.19.1, Singmann, Bolker, Westfall, & Aust, 2016), *cocor* (1.1.3, Diedenhofen & Musch, 2015), *papaja* (0.1.0.9492, Aust & Barth, 2016), and *psychometric* (2.2, Fletcher, 2010) in our analyses. As recommended by Lakens (2013), we used Cohen's $d_z = \frac{t}{\sqrt{n}}$ as a measure of effect size when comparing two dependent means (cf. Rosenthal, 1991). When comparing two independent means, we used the standard effect size *d* (Cohen, 1988). According to Cohen (1988), a value of 0.2 can be interpreted as a small effect, a value of 0.5 a medium effect, and a value of > 0.8 a large effect for both of these measures. Analysis of variance (ANOVA) effect sizes were computed using partial eta-squared η_p^2, with $\eta_p^2 \geq 0.01$ implying a small effect, $\eta_p^2 \geq 0.06$ a moderate effect, and $\eta_p^2 \geq 0.14$ a large effect according to Cohen (1988).

Evaluating the index of acceptance reluctance. The average number of response options presented in the six no-solution DOMC items was computed as an index of acceptance reluctance. On average, participants were presented with 3.12 (*SD* = 0.82) of the five response options in DOMC no-solution items. Whereas participants most frequently accepted the third of the five options, some participants saw either a larger or a smaller

number of options before accepting one. Most importantly, the number of response options test-takers saw was consistent across the six DOMC no-solution items, as indicated by a Cronbach's α of .63 (Cronbach, 1951). Thus, there were reliable individual differences in how rapidly test-takers accepted a distractor as the solution in no-solution items.

To validate the proposed index of acceptance reluctance, we used the response criterion *c* participants employed in DOMC knowledge items. We computed *c* separately for each option position to investigate whether test-takers adapted their response criterion over the course of each DOMC item. To find out whether individual differences in acceptance reluctance measured on the basis of the no-solution items were related to the response criterion test-takers employed in real DOMC knowledge items, we also analyzed the response criterion *c* as the dependent variable in a covariance analysis (ANCOVA). In this analysis, the position of the response option was used as a repeatedly measured independent variable, while acceptance reluctance served as the covariate. To compute the response criterion *c*, we determined test-takers' false alarm and the hit rates for the first, second, third and fourth response options in BOWIT DOMC knowledge items. It was not possible to compute *c* for the fifth and last response option because this option is only shown when it is the solution; thus, a false alarm or correct rejection can never occur. The relative false alarm and hit rates for each position were transformed into *c* using formula (7) from Stanislaw and Todorov (1999):

$$c = -0.5 * (z(\mathrm{H}) + z(\mathrm{FA})) \quad (1)$$

In formula (1), z is the inverse of the cumulative normal distribution function. A loglinear correction was applied to compensate for hit or false alarm rates of 100% or 0% (Stanislaw & Todorov, 1999). According to Formula (1), $c > 0$ indicates a conservative criterion and $c < 0$ a liberal criterion. The response criterion is neutral ($c = 0$) if participants do not lean towards either rejecting or accepting a response option.

Adjusting the degrees of freedom using a Greenhouse-Geisser correction, we found a strong effect of response option position on the response criterion *c*, $F(2.82, 547.40) = 54.92$, $p < .001$, $\eta_p^2 = .22$. Test-takers adapted their response criterion and generally became more liberal with increasing option position (see Figure 2). We also found that test-takers' response criterion was related to acceptance reluctance, $F(1, 194) = 9.78$, $p = .002$, $\eta_p^2 = .05$. Participants who were more reluctant to accept distractors in no-solution items generally also employed a more conservative response criterion in DOMC items (see Figure 2). Consequently, acceptance reluctance was also associated with the acceptance of fewer distractors in the DOMC knowledge test, $r = -.36$, $p < .001$.

– Please insert Figure 2 about here –

Next, we explored the relationship between acceptance reluctance and test scores on the BOWIT knowledge test. Generally, there was a positive correlation between test scores and acceptance reluctance. This was true both for the DOMC knowledge test, $r = .28$, $p < .001$, and for the MC knowledge test, $r = .17$, $p = .017$. We compared these two Pearson correlations using the R-package *cocor* (Diedenhofen & Musch, 2015), and found acceptance reluctance to be more strongly associated with DOMC than with MC test scores, $z = 2.00$, $p < .05$.

We then conducted a hierarchical regression to test whether acceptance reluctance was related to DOMC test scores beyond what could be explained on the basis of participants' level of general knowledge. To this end, we compared a regression model that included MC test scores as the only predictor of DOMC test scores with a model that added acceptance reluctance as an additional predictor. MC test scores explained 50% of the variance in DOMC test scores, $R^2 = .50$, $F(1,194) = 192.80$, $p < .001$. Adding acceptance reluctance

as an additional predictor improved this prediction by a small but significant margin, $\Delta R^2 = .02$, $F(1, 193) = 10.17$, $p = .002$.

It is possible that individual differences in knowledge contribute to the observed correlation between DOMC test scores and acceptance reluctance. More knowledgeable participants might be more reluctant to endorse incorrect options because they are used to being able to tell with confidence whether an answer is true or false. We therefore computed a partial correlation between DOMC test scores and acceptance reluctance that controlled for the overall level of knowledge as measured by the MC test. This partial correlation was significant, $r = .22$, $p = .002$, providing additional evidence for the notion that acceptance reluctance influences DOMC test scores independently of participants' knowledge.

Personality correlates of acceptance reluctance. Table 1 details the correlations between acceptance reluctance and all other measures we collected. Acceptance reluctance was not related to any of the personality variables we assessed. In particular, it was unrelated to risk-taking propensity (regardless of whether this was measured via self-ratings or via the balloon analogue risk task), need for cognition, perfectionism, overclaiming, self-rated knowledge and self-rated intelligence.

– Please insert Table 1 about here –

Additional analyses. To better understand the processes involved in DOMC testing, we conducted further explorations of test-takers' responses to DOMC knowledge items. We found that false alarms occurred more frequently than misses across all 18 DOMC items, indicating that test-takers prematurely accepted a distractor ($M = 4.74$, $SD = 2.59$) more often than they wrongfully rejected a correct solution, ($M = 3.95$, $SD = 2.47$), $t(195) = 3.48$, $p < .001$, $d_z = 0.25$. A potential explanation for this preponderance of false alarms is the

higher base rate of distractors. Depending on the position of the solution, there are between one and four opportunities to accept a distractor in a DOMC test item, but only one opportunity to miss the solution. Moreover, we found that when the solution was presented as the first option, DOMC knowledge items were solved correctly in 59% of all cases. When the solution was presented at Position 2, 3, 4 or 5, the proportion of correct responses was 57%, 56%, 49% and 39%, respectively. Thus, DOMC items were more difficult if the solution occurred in a later position. Across all DOMC item presentations, test-takers were presented with 2.48 of the 5 response options on average (SD = 1.31). In 30% of all cases, only one option was presented; 2, 3, 4 or 5 options were presented in 26%, 20%, 14% and 10% of all item presentations, respectively.

Discussion

Study 1 provided first evidence for the notion that responses in DOMC tests are affected by individual differences in acceptance reluctance. To arrive at this conclusion, we tested whether the number of option rejections in no-solution items varied systematically between test-takers. The use of no-solution items ensured that response decisions were not influenced by test-takers' knowledge. Surprisingly, readiness to reject distractors was quite consistent, as indicated by a Cronbach's α of .63. These systematic differences among test-takers on no-solution items indicated that a factor other than knowledge influences DOMC test-takers' responses.

In the present investigation, we conceptualize acceptance reluctance as an individual's stable preference for later response options in DOMC items. According to MORBIST, this preference is expected to shape the response criterion test-takers employ in DOMC items. In particular, MORBIST assumes that test-takers adapt their response criterion to the position of DOMC response options. In accordance with MORBIST's predictions, we found in a repeated-measures ANCOVA that test-takers' response criterion was related to both

acceptance reluctance and the position of response options. We also observed a significant correlation of $r = .28$ ($p < .001$) between DOMC test scores and acceptance reluctance, indicating an advantage for patient test-takers who did not accept options prematurely. Importantly, the partial correlation ($r = .22$, $p = .002$) between DOMC test scores and acceptance reluctance that controlled for the overall level of knowledge as measured by an independent MC test was also significant, providing evidence for the notion that acceptance reluctance influences DOMC test scores independently of participants' knowledge.

In additional analyses, we found that the premature acceptance of a distractor caused an incorrect response more often than the wrongful rejection of the solution. We therefore assume that the main advantage of high acceptance reluctance lies in avoiding the premature acceptance of distractors. We also found that the difficulty of DOMC items depended on the position of the solution. DOMC items were particularly difficult when the solution was presented late, as the fourth or fifth option. This is most likely because the probability of committing at least one error accumulates across option positions. When the solution is shown as the first option, test-takers have only one opportunity to commit an error, namely by missing the solution. If the solution is shown as the fourth option, however, there are three opportunities to commit a false alarm before the solution can even be presented.

Although we found evidence that acceptance reluctance as a response style was related to response decisions in a DOMC test, we did not find any relationships between acceptance reluctance and the personality variables we assessed. None of them predicted the individual differences in acceptance reluctance we observed, suggesting that acceptance reluctance might be a separate trait that can and should be distinguished from other personality variables. This would be in line with most previous research, which tended to find little evidence for systematic relationships between response biases and personality traits (McGee, 1962; Rorer, 1965; but see Berg, 1955). With regard to participants' risk-taking propensity as

measured in the BART, for example, it is possible that risk aversion has diverging effects depending on whether test-takers primarily try to avoid false alarms or avoid missing the correct solution. Such individual differences would make it impossible to predict the direction of the correlation between risk-taking propensity and acceptance reluctance.

Study 2

The results of Study 1 suggest that DOMC test-takers exhibit systematic and reliable differences in their willingness to accept early response options. In a signal detection framework, MORBIST can account for this finding by assuming that test-takers differ in the response criteria they employ at different DOMC answer positions. While it is rational to use a more lenient response criterion for late answer options, MORBIST assumes that some test-takers are also willing to accept an early response option. A critical question is whether such individual differences in response style cause differences in test scores that are unrelated to ability. The results of the correlational Study 1 suggest that high acceptance reluctance is related to a higher probability of answering DOMC items correctly. If high acceptance reluctance increases DOMC test scores, this would imply that DOMC test scores contain variance due to a construct-irrelevant response style. In Study 2, we therefore conducted a computer simulation to investigate whether MORBIST can account for the observation that high acceptance reluctance leads to higher DOMC test scores. To validate MORBIST, we also investigated whether the model can account for some known characteristics of DOMC testing. To this end, we tested whether MORBIST a) correctly predicts that DOMC items are more difficult than traditional MC items; b) can account for the fact that in DOMC tests, late answer options are less likely to be seen by the test-taker; c) can account for the fact that the difficulty of DOMC items increases with the increasing serial position of the solution; and d) can account for the finding that there typically are more false alarms than misses in a DOMC test. All MORBIST simulations were conducted using the R programming language (R Core

Team, 2016). The simulation code is freely available and can be accessed via https://osf.io/637x8/.

Basic assumptions

In the simulation, MORBIST modeled the sequential processing of DOMC response options. Each DOMC item was represented by an answer vector indicating the correctness of all response options. Thus, for example, (0, 1, 0, 0) represented a DOMC item in which the solution was presented as the second of four answer options. The position of the solution was determined randomly for each item. For each response option, a random value was generated representing the level of evidence evoked in a hypothetical observer. We employed the Gaussian model of signal detection, thus assuming that evidence strength is distributed normally, with different means for solutions and for distractors (Macmillan & Creelman, 2005). In our simulation, the signal-detection parameter *d'* represented test-takers' ability to differentiate between solutions and distractors. Therefore, the mean of the distribution of the distractors was set to 0 and the mean of the distribution of the solution was set to *d'* for all test-takers. The standard deviation was set to 1 for both the solution distribution and the distractor distribution.

For each option presented, the response process was simulated by drawing a random number from the applicable distribution of evidence strengths (for either solutions or distractors). Next, this random number was used to determine whether the strength of evidence associated with the current option surpassed the response criterion *c,* which represented the level of evidence MORBIST required to accept an option as being correct. This response criterion was taken from the vector of response criteria MORBIST created for each simulated test-taker. To this end, each test-taker was modeled with a vector of response thresholds $c = (c_1, ..., c_n)$ for a DOMC item comprising n response options. An option was accepted as the solution whenever its evidence strength was higher than its associated

response criterion. DOMC scoring was applied by awarding one point for each item in which the correct response option was accepted. The total DOMC test score was computed as the sum of all DOMC item scores.

In MC tests, all answer options are available simultaneously. Therefore, no fixed response criterion has to be applied when answering MC items; instead, the option that yields the highest strength of evidence can always be chosen as the solution (cf. Green & Moses, 1966; Norman & Wickelgren, 1969). Thus, whenever the solution invoked the highest strength of evidence in the simulation, one point was awarded for correctly solving this item. A total MC test score was then computed as the sum of all MC item scores.

First simulation

In the first simulation, 1,000 test-takers were simulated to process 100 DOMC items, and another 1,000 test-takers were simulated to process 100 MC items. Each item consisted of five answer options. The ability parameter *d'* was simulated for each test-taker by drawing a random value from a normal distribution ($M = 1.5, SD = 0.5$). In the DOMC condition, the response criterion vector was set to $c = (0.47, 0.32, 0.13, 0, -\infty)$ for each test-taker. With the exception of the last option – which, if presented, must be the solution and should therefore be accepted – this vector corresponded to the empirically-determined average response criteria test-takers employed in Study 1 (cf. Figure 2). Response decisions were recorded as determined in the simulation. For the DOMC condition, we also recorded the serial position of the solution, the number of options shown, and the number of false alarms and misses.

Results. The simulation successfully reproduced what has repeatedly been observed in empirical studies: the DOMC items were more difficult ($M = 0.54$, $SD = 0.14$) than the MC items ($M = 0.65$, $SD = 0.15$), $t(1998) = 17.28$, $p < .001$, $d = 0.77$. On average, 2.48 (SD = 1.30) answer options had to be shown before a DOMC item was answered either correctly or incorrectly. Item presentation ended after the first, second, third, fourth or fifth answer

option in 30%, 26%, 21%, 15%, and 9% of all item presentations, respectively. This pattern is very similar to the empirical results we observed in Study 1, where the respective numbers were 30%, 26%, 20%, 14%, and 10%.

The difficulty of the simulated DOMC items varied as a function of the serial position of the solution. If the first option was the solution, items were solved correctly in 60% of all cases. If the solution was the second, third, fourth or fifth option, items were solved in 58%, 55%, 48% and 47% of all cases, respectively. This pattern is very similar to the empirical results observed in Study 1, where the respective numbers were 59%, 57%, 56%, 49% and 39%. In the simulation, false alarms occurred more often ($M = 0.26$, $SD = 0.10$) than misses ($M = 0.20$, $SD = 0.06$), $t(999) = 21.43$, $p < .001$, $d_z = 0.68$,similar to what was observed empirically in Study 1. This observation does not directly contribute to an independent validation of MORBIST, however, because it depends on the values chosen for the criterion vector, which we determined using the responses empirically observed in Study 1. The fact that employing these criteria produced a simulated response behavior that closely mirrored the empirically observed behavior nevertheless attests to the validity of MORBIST.

Second simulation

In the second simulation, we systematically varied the response criterion vector to simulate high and low acceptance reluctance, respectively. In the low acceptance reluctance condition, we used a neutral and rather liberal response criterion that remained constant for all but the last option. To this end, a criterion vector with thresholds $c = (0,0,0,0,-\infty)$ was employed to make sure that the response criterion was midway between the distribution of the solution and the distributions of the distractors, and that the last option was always accepted if no prior option had been chosen. In the high acceptance reluctance condition, we used a response criterion that was more conservative than in the low acceptance reluctance condition for early options, and gradually became more liberal for options presented later,

$c = (0.6, 0.4, 0.2, 0, -\infty)$. The ability parameters *d'* were again drawn as random values from a normal distribution ($M = 1.5, SD = 0.5$). In total, 2,000 test-takers – 1,000 per acceptance reluctance condition – each processing 100 DOMC items were simulated.

Results. As surmised and consistent with the correlational findings in Study 1, a high simulated acceptance reluctance led to higher test scores ($M = 54.84$, $SD = 12.34$) than a low simulated acceptance reluctance ($M = 51.06$, $SD = 12.30$), $t(1998) = 6.86$, $p < .001$, $d = 0.31$.

Discussion

In the first simulation, MORBIST successfully reproduced several known properties of the DOMC test format. First, DOMC test scores were lower than MC test scores. This was expected because performance in a yes/no decision task is usually poorer than performance in a forced-choice task (Jang, Wixted, & Huber, 2009). In the simulation, the odds of solving an MC item were higher than the odds of solving a DOMC item (odds ratio = 0.65 / 0.54 = 1.20). This odds ratio closely mirrors previous comparisons of MC and DOMC test scores. In two empirical data sets, Kingston et al. (2012) reported higher odds of solving an MC item than a DOMC item (odds ratios = .72 / .62 = 1.16 and .73 / .60 = 1.22, respectively).

MORBIST also successfully reproduced three empirical results obtained in Study 1. For DOMC items, (1) difficulty increased when the solution was presented as one of the later options; (2) false alarms occurred more often than misses; and (3) item presentation most frequently stopped after the presentation of only one option, and least frequently after the presentation of all five options. Frequencies for the presentation of two, three, or four response options lay in between these two extremes. This completely parallel pattern of results provided an encouraging first validation of MORBIST, and lent support to the notion that MORBIST adequately captures the process of responding to a DOMC item. Importantly, the second simulation showed that MORBIST also predicts increased DOMC test scores for

high acceptance reluctance. In this simulation, high acceptance reluctance was operationalized using an adaptive criterion that started conservatively and gradually became more liberal for later response options. This led to increased test scores even though acceptance reluctance was independent of knowledge in the simulation, and thus was not confounded with knowledge as in the correlational Study 1. Taken together, all simulation results were consistent with the empirical results obtained in Study 1, and supported the notion that acceptance reluctance adds construct-irrelevant variance to DOMC test scores.

Study 3

In Study 1, we established the occurrence of reliable individual differences in acceptance reluctance using a DOMC test with no-solution items. In Study 2, we used MORBIST to simulate how acceptance reluctance contributes to performance on DOMC tests. The simulations showed that individual differences in acceptance reluctance systematically influence DOMC test scores; high acceptance reluctance leads to higher scores. However, the MORBIST simulation was based on assumptions that may not be met in real DOMC testing situations, and could therefore only lend plausibility to the notion that DOMC test scores are contaminated by construct-irrelevant variance due to individual differences in acceptance reluctance. In Study 3, we tested this notion experimentally with human participants in a more ecologically-valid manner that did not depend on the correctness of the assumptions underlying MORBIST.

We experimentally manipulated acceptance reluctance within test-takers to investigate its causal influence on achievement. We thus avoided the confound between knowledge and acceptance reluctance that occurred in Study 1, where acceptance reluctance was correlated with both DOMC and MC test scores. In Study 1, more knowledgeable participants were also more reluctant in accepting answers, even though acceptance reluctance was measured using separate no-solution items. More knowledgeable participants might have been more reluctant

to endorse incorrect options because they are used to being able to tell with confidence whether an answer is true or false. Using an experimental approach, Study 3 aimed to establish an unambiguous causal link between acceptance reluctance and DOMC test scores that did not depend on individual differences in knowledge. To this end, participants completed a DOMC test in two different conditions, a high and a low acceptance reluctance condition, which were realized within participants using a payoff manipulation. High acceptance reluctance was induced by offering a higher payoff for later choices, provided that they were correct; low acceptance reluctance was induced by offering a higher payoff for early choices. We expected that test scores in the high acceptance reluctance condition would be higher than in the low acceptance reluctance condition even though the counterbalanced within-participants design made sure that there were no systematic individual differences in knowledge between participants exhibiting high and low acceptance reluctance. Observing this predicted pattern would provide strong support to the notion that high acceptance reluctance is causally linked to better DOMC test performance independently of test-takers' knowledge.

Method

We asked participants to learn vocabulary in an unfamiliar language. We experimentally induced high or low levels of acceptance reluctance and investigated their influence on scores in a subsequent DOMC vocabulary test.

Materials. To avoid floor and ceiling effects in DOMC test scores, we experimentally controlled for participants' knowledge levels. To this end, participants were shown a list of words written in the fictive language Dothraki together with their German translation (e.g., "Sieger" - "najahak"). We expected participants to not be familiar with these words because there are no native speakers of Dothraki except for some fictional horsemen in the TV series "Game of Thrones". In total, 64 words were selected from a Dothraki dictionary created by

the linguist who devised the Dothraki language for the series (Peterson, 2014). To construct the DOMC test, four distractors were created for each word (e.g., “jahanak”, “ajanak”, “njaka” and “kanajak”), resulting in five response options for each item. Each DOMC item stem had the form “The dothraki word for (*German word*) is …”. Response options were presented sequentially and in random order beneath the continuously shown item stem.

It was not possible to use the same items in both acceptance reluctance conditions because the payoff matrix was varied within participants. The 64 items were therefore split into two sets of equal size that were presented as two separate tests, one in the high acceptance reluctance and one in the low acceptance reluctance condition. The assignment of the two item sets to the two payoff conditions and the order of the payoff conditions were counterbalanced across participants. The complete list of all experimental stimuli – including the distractors and English and German translations – can be retrieved from https://osf.io/637x8/.

Design. Low and high acceptance reluctance were established using a payoff manipulation that was conducted within participants. Thus, each test-taker completed two DOMC test halves that employed different payoff matrices. One half of the items were presented along with a payoff matrix that incentivized the selection of early options to induce low acceptance reluctance. In this phase of the experiment, participants received a payoff of 25, 16, 9, 4, and 1 points if they accepted a solution presented as the first, second, third, fourth, or last option, respectively. The other half of the test items were presented along with a payoff that incentivized the selection of late options to induce high acceptance reluctance. In this phase of the experiment, participants received a payoff of 1, 4, 9, 16, and 25 points if they accepted a solution presented as the first, second, third, fourth, or last option, respectively. Additionally, one point was awarded for each correct rejection of a distractor regardless of payoff condition. This made sure that participants received points for all correct

decisions, but not for any incorrect decisions. Regardless of payoff condition, the optimal decision was always to select "true" whenever a solution was shown, and "false" whenever a distractor was shown. This strategy, however, was only available in trials in which test-takers actually knew the solution. In trials in which test-takers were not sure of the solution, the rational decision was to either increase or decrease acceptance reluctance over the course of an item according to the applicable payoff matrix.

During the test, the payoffs associated with each response option were always displayed on the "true" and "false" buttons, respectively. The payoff associated with the "true" button was updated according to the payoff matrix whenever a new response option was presented, while the "false" button always indicated that one point could be gained by rejecting an option, provided that this option was actually wrong. By employing one of the above payoff schemes on the "true" button, it was thus possible to induce a bias toward either early or late response options. The dependent variable was the DOMC test score, which could range from 0 to 32 in each test half representing the high and the low acceptance reluctance conditions, respectively.

Sample. The experiment was conducted in a computer laboratory at the University of Duesseldorf. Participants were seated in front of a computer screen presenting all experimental stimuli, and entered their responses using a mouse. Data were collected from a total of 223 participants. The data from the 32 participants who indicated a native language other than German were discarded to make sure that all participants were capable of understanding the detailed instructions explaining the payoff matrices. The final sample therefore consisted of $n = 191$ participants (78% female). Participants' age ranged from 17 to 53 years ($M = 21.30$, $SD = 4.35$). Participants were incentivized by the opportunity take part in a lottery in which three prizes (100, 50, and 30 Euros) were raffled off among the 30% of participants who earned the most points. This lottery was conducted to encourage

participants to follow the instructions closely. The sample consisted of 167 psychology students from the University of Duesseldorf who received partial course credit for their participation, 20 students from various other faculties, and 4 volunteers not studying at the University of Duesseldorf. Using the software G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009), we determined that this sample size was sufficient to detect small effects ($d_z = .20$) of acceptance reluctance on DOMC test scores with a statistical power of 87% in a one-tailed *t*-test conducted in accordance with the one-directional nature of the hypothesis (Cohen, 1988).

Procedure

The experiment was presented in a web browser using Unipark EFS Survey (QuestBack, 2017). The first page presented introductory information on the study and obtained participants' informed consent. On the next page, participants were introduced to the DOMC test format in written instructions that provided a detailed explanation of the payoff manipulation. Participants were instructed to maximize their payoff over the course of the study by considering the number of points offered as a function of the position whenever they were not sure about their answer. To achieve the desired experimental manipulation of acceptance reluctance, we pointed out that it would be rational to select earlier options when early options were associated with a higher payoff, and to select later options when late options were associated with a higher payoff. Participants were also informed that they would only receive points when they made a correct decision, and that they would only be allowed to take part in the lottery for three monetary prizes if they did well on the task. This made sure that participants primarily tried to answer correctly, but took the payoff matrix into account whenever they were not sure about the correct answer. To illustrate the response format and the varying payoffs, test-takers next worked on six example items, half of which employed a high and half a low acceptance reluctance payoff matrix. In each example item,

participants were asked to identify the correct English translation of a German word from among five answer alternatives in the sequential DOMC test format.

The subsequent learning phase presented 32 Dothraki words for five seconds each, along with their German translations. Following this learning phase, the payoff matrix used in the first half of the test phase was explained briefly again to ensure that test-takers would remember and consider the payoffs when making their response decisions. Next, the 32 Dothraki words previously encountered during the learning phase were presented again in random order in DOMC test format. The order in which the response options were presented in each item was determined randomly for each participant. Following this testing phase, another learning phase and another testing phase were conducted in which the remaining set of Dothraki words was taught and tested using the other payoff matrix. Finally, some demographic data were collected. Test-takers were then debriefed and thanked, and were provided with feedback regarding the correctness of their responses and the total number of points they had earned.

Results

Manipulation check. To check the effectiveness of the manipulation, we conducted a 2×4 within-subjects ANOVA analyzing the influence of the payoff matrix [low vs. high acceptance reluctance] and the position of response options [1, 2, 3, 4] on the test-takers' decision threshold *c*, which was estimated using the standard signal detection formula (1) previously used in Study 1. As expected, this ANOVA revealed a strong effect of position; participants generally decided more liberally on later response options, $F(2.83, 538.04) = 85.89$, $p < .001$, $\eta_p^2 = .31$ (see Figure 3). There was also a strong effect of the payoff manipulation, $F(1,190) = 110.76$, $p < .001$, $\eta_p^2 = .37$. Participants responded more conservatively in the high acceptance reluctance condition employing higher payoffs for later options (average $c = 0.32$, $SD = 0.45$) than in the low acceptance reluctance condition

employing higher payoffs for early options (average $c = 0.11$, $SD = 0.41$). The experimental induction of two different levels of acceptance reluctance was thus found to be successful. As illustrated in Figure 3, there was also a significant interaction between payoff matrix and position, $F(2.90, 551.85) = 14.06$, $p < .001$, $\eta_p^2 = .07$, indicating that as expected, the difference in the threshold employed in the high and low acceptance reluctance conditions was largest for early options and became smaller for later response options. Due to their more conservative response criterion, test-takers were presented with more response options per item in the high acceptance reluctance condition ($M = 2.60$, $SD = 0.31$) than in the low acceptance reluctance condition ($M = 2.39$, $SD = 0.35$), $t(190) = 7.18$, $p < .001$. This difference in the number of response options shown was equivalent to an effect size of $d_z = 0.52$.

– Please insert Figure 3 about here –

Hypothesis test. A paired *t*-test showed that – as predicted by MORBIST – DOMC test scores were significantly higher in the high acceptance reluctance condition ($M = 18.60$, $SD = 4.55$) than in the low acceptance reluctance condition ($M = 18.07$, $SD = 4.96$), $t(190) = 2.02$, $p = .022$, one-tailed, $d_z = 0.15$. We were thus able to confirm the central prediction of a positive causal effect of acceptance reluctance on DOMC test scores.

Validation of MORBIST. To determine whether MORBIST can adequately describe the empirical results and reproduce the DOMC test scores observed in Study 3, we used MORBIST to simulate DOMC test scores based on the abilities *d'* and response criteria *c* we observed for the 191 participants in Study 3. To compute *d'* as an index of each participant's ability to distinguish between correct and incorrect options, we categorized his or her answers to every single presentation of a solution or a distractor across all 64 items. This resulted in a

2x2 table for each participant containing the number of hits (accepted solutions), misses (rejected solutions), false alarms (accepted distractors) and correct rejections (rejected distractors). On the basis of this table, an ability index *d'* was computed for using formula (1) from Stanislaw and Todorov (1999):

$$d' = z(\mathrm{H}) - z(\mathrm{FA}) \qquad (2)$$

In addition, we constructed eight 2x2 tables for each participant separating hits, false alarms, correct rejections, and misses by the position of the response options [1, 2, 3, 4] and the level of acceptance reluctance [low, high]. On the basis of these tables, eight response criteria were computed for each participant using the procedure detailed in Study 1. As in Study 2, we always assumed a criterion of $-\infty$ for the fifth option. We then used MORBIST to first simulate responses for all 191 test-takers using the response criteria observed for the 32 items presented in the high acceptance reluctance phase. Next, we simulated the responses of the same 191 test-takers using the response criteria observed for the 32 items in the low acceptance reluctance phase. This process was repeated 10,000 times, with descriptive statistics as well as the *t*-value associated with each paired comparison of low and high acceptance reluctance test scores recorded for each run. The simulation code is freely available and can be accessed via https://osf.io/637x8/. Across all simulations, high acceptance reluctance led to 18.70 correct answers on average (SD = 4.91), while low acceptance reluctance led to 18.31 correct answers (SD = 5.00). The higher test scores empirically observed in the high acceptance reluctance condition in Study 3 could thus successfully be reproduced by MORBIST. On average, the simulated test scores between the high and the low acceptance reluctance condition were correlated with r = .69; this closely reproduced the empirical correlation of .71 observed in Study 3.

Finally, we compared the likelihood of the data observed in Study 3 under the null hypothesis of no influence of individual differences in acceptance reluctance, and the

alternative MORBIST prediction of a systematic positive influence of acceptance reluctance on DOMC test scores. This allowed us to compute a Bayes factor that quantified the relative evidence for the null hypothesis and the prediction made by MORBIST, and to base our statistical inference directly on substantive psychological theory, as recommended by Vanpaemel (2010). To compute the likelihood of the data – i.e., the observed test score difference between high and low acceptance reluctance – under MORBIST, the probability density of the observed *t*-value ($t = 2.02$) was estimated on the basis of the distribution of the *t*-values computed in the simulation. This was done using a logsplines estimator (Kooperberg, 2016; cf. Morey, Rouder, Pratte, & Speckman, 2011). The likelihood of the observed data under the null hypothesis was determined as the probability density of the central *t* distribution with $n - 1$ degrees of freedom at $t = 2.02$. The ratio of these two likelihoods—the Bayes factor—was 6.25. Hence, the data were approximately 6 times more likely under MORBIST – which predicts an effect of acceptance reluctance on test scores – than under the null model assuming no effect of acceptance reluctance. This finding can be interpreted as substantial evidence supporting MORBIST (Kass & Raftery, 1995).

Figure 4 illustrates MORBIST's predictions and the resulting Bayes factor. The relative evidence the data provide for the two competing models are illustrated by the two black circles showing the likelihood ratios under the two models (cf. Rouder, Morey & Wagenmakers, 2016). The vertical line in Figure 4 shows the size of the effect of acceptance reluctance on test scores observed for the human participants in Study 3. For this figure, t-values were converted to the effect size index $d_z = \frac{t}{\sqrt{n}}$ for better interpretability. The figure shows that MORBIST predicts a relatively small effect size of acceptance reluctance on test scores, given the participants' observed response criteria.

– Please insert Figure 4 about here –

Additional analyses. In exploring the test results, we noticed that test-takers scored better in the second test ($M = 18.71$, $SD = 4.92$) than in the first ($M = 17.96$, $SD = 4.58$), $t(190) = 2.89$, $p = .004$, $d_z = 0.21$, regardless of the order in which the two payoff conditions were presented. This result indicates that a learning effect occurred over the course of the experiment, leading to increased test scores in the later test. This learning effect is most likely attributable to the test-takers' increasing familiarity with the DOMC test format or the Dothraki language and was controlled for by counterbalancing the order of the two payoff conditions. Therefore, it does not affect the main result that acceptance reluctance leads to better scores.

Discussion

Applying signal detection theory to the analysis of responses to DOMC tests, we developed MORBIST, a model of response behavior in sequential tests. According to MORBIST, response decisions are the joint product of test-takers' knowledge and their response criterion. The response criterion is assumed to vary over the presentation of a DOMC item. Generally, test-takers should become increasingly more liberal from the first to the last response option to compensate for the fact that later response options have a higher probability of being the solution because fewer responses are left. Additionally, MORBIST assumes that there are individual differences in acceptance reluctance, that is, in the tendency to reject early response options. On the basis of these assumptions and using a signal detection model, MORBIST predicts that higher acceptance reluctance is causally related to higher scores, implying that DOMC test scores are contaminated with unwanted construct-irrelevant variance. Thus, MORBIST not only provides a model for the internal response process during DOMC testing, but also suggests a direct link between these processes and test validity (cf. Borsboom et al., 2004).

Manipulating acceptance reluctance within participants using a payoff manipulation, we found that high acceptance reluctance leads to better test scores than low acceptance reluctance even when no individual differences in knowledge are involved. This experimental finding is consistent with the correlational finding in Study 1 that high acceptance reluctance – operationalized as the number of options test-takers rejected in no-solution items – is associated with higher test scores. Moreover, it shows that this correlation cannot be solely accounted for by individual differences in knowledge.

Although the effect size of the influence of acceptance reluctance on DOMC test scores was relatively small in the experimental setting of Study 3, it should not be concluded that in more applied settings, acceptance reluctance has only minor effects on DOMC test scores. This is because Study 3 was not designed to determine the extent to which individual differences in acceptance reluctance distort DOMC test scores, but rather to establish their causal connection. Under such circumstances, MORBIST predicted a small effect size of acceptance reluctance on DOMC test scores (see Figure 4). Moreover, while the observed effect of acceptance reluctance on test scores was relatively small on the aggregate level (d_z = 0.15), it is important to note that individuals employing a lower response threshold are systematically put at a disadvantage by the DOMC test format. DOMC tests are therefore unfairly biased against test-takers with low acceptance reluctance.

Study 3 also showed that in principle, test-takers' response criteria in DOMC testing can be influenced by employing appropriate incentives. This demonstrates that test-takers are capable of strategically altering their response behavior in DOMC tests. It is therefore remarkable that our exploratory analyses in Study 3 showed that test-takers tended to perform better in their second DOMC test. This finding suggests that test-takers employ better response strategies as they become increasingly aquainted with the DOMC test format, and is reminiscent of a debate surrounding the multiple choice (MC) question format some decades

ago. In an early study employing the MC test format, Swineford (1941) reported that boys were more willing than girls to gamble when they were unsure of an answer, and suggested that such differences in willingness to guess may systematically bias MC test scores. Several decades later, however, Ben-Shakhar and Sinai (1991) only found minor gender differences in guessing tendencies, which therefore also accounted for only a small fraction of the observed gender difference in multiple-choice tests. This implies that the influence of construct-irrelevant variables such as individual differences in the propensity to take risks or willingness to accept early answers may decrease over time as test-takers become more acquainted with a new test format. It is therefore conceivable that the variance in DOMC test scores that is attributable to individual differences in acceptance reluctance gradually diminishes with repeated DOMC testing. This possibility deserves further investigation in future studies of the DOMC test format.

On a theoretical level, our investigations suggest that MORBIST offers a useful framework for describing response decisions in DOMC tests. In the simulations conducted in Study 2, MORBIST reproduced several known properties of the DOMC test format, including its higher difficulty compared to MC tests and the influence of the position of the solution on the difficulty of DOMC items, as well as the reduced number of options that usually have to be shown. Moreover, using a Bayesian framework, Study 3 found that the likelihood of the observed difference in test scores between low and high acceptance reluctance was six times higher under MORBIST than under the null hypothesis, providing substantial evidence that acceptance reluctance systematically influences DOMC test scores. This result encourages further investigations of acceptance reluctance as a “cognitive trait” (Kantner & Lindsay, 2012).

In Study 1, acceptance reluctance was correlated across several no-solution items, thus establishing the existence of reliable individual differences in acceptance reluctance in the

first place. While acceptance reluctance had never before been investigated as a cognitive trait in previous studies of DOMC testing, our findings resemble other demonstrations of individual differences in response styles for other response formats such as true-false tests (Cronbach, 1946; Swineford, 1938; Thorndike, 1938). Response styles have also received increased attention in personality research in recent years (e.g., Kantner & Lindsay, 2012; Wetzel, Lüdtke, Zettler, & Böhnke, 2016; Ziegler, 2015). Other research domains that may profit from a closer look at individual differences in acceptance reluctance include consumer product choice (e.g., Bearden & Connolly, 2007) and eyewitness performance in police lineups (e.g., Mickes, Flowe, & Wixted, 2012; Meisters, Diedenhofen, & Musch, 2018).

To summarize, the proposed model of response behavior in sequential tests (MORBIST) was found to provide a promising foundation for investigating the response processes involved in answering DOMC questions. Simulations conducted on the basis of MORBIST– as well as additional correlational and experimental investigations – showed that knowledge-independent individual differences in acceptance reluctance systematically affect DOMC test scores, which therefore include some construct-irrelevant variance. To make an informed decision on the usefulness of the DOMC test format in practical settings, this problem has to be weighed against the potential positive aspects of DOMC testing, which include better control of testwiseness, increased test security, and reduced testing times compared to conventional MC tests.

References

Aust, F., & Barth, M. (2016). *papaja: Create APA manuscripts with RMarkdown*. Retrieved from https://github.com/crsh/papaja

Bearden, J. N., & Connolly, T. (2007). Multi-attribute sequential search. *Organizational Behavior and Human Decision Processes*, *103*, 147–158.

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28*, 23–35.

Berg, I. A. (1955). Response bias and personality: The deviation hypothesis. *Journal of Psychology, 40*, 61–72.

Bless, H., Wänke, M., Bohner, G., Fellhauer, R. F., & Schwarz, N. (1994). Need for Cognition: Eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben [Need for cognition: A scale measuring engagement and happiness in cognitive tasks]. *Zeitschrift Für Sozialpsychologie*, *25*, 147–154.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

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 & Learning Assistance. ERIC database (ED240928). Retrieved from http://eric.ed.gov/?id=ED240928

Buchner, A., Erdfelder, E., & Vaterrodt-Plünnecke, B. (1995). Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General*, *124*, 137–160.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*, 197–253.

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

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*, 475–494.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*, 3–31.

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

Diedenhofen, B., & Musch, J. (2015). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLoS ONE*, *10*, e0121945.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, *9*, 522–550.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.

Fletcher, T. D. (2010). *psychometric: Applied Psychometric Theory*. Retrieved from https://CRAN.R-project.org/package=psychometric

Foster, D., & Miller, H. (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). *Test-wiseness as secondary cue response*. (Doctoral dissertation) No. 64-7643. Stanford University, Ann Arbor, MI: University Microfilms.

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher*, *36*, 456–462.

Green, D. M., & Moses, F. L. (1966). On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin*, *66*, 228–234.

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

Hossiep, R., & Schulte, M. (2008). *BOWIT: Bochumer Wissenstest*. Göttingen: Hogrefe.

Ingham, J. G. (1970). Individual differences in signal detection. *Acta Psychologica*, *34*, 39–50.

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*, 291–306.

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition*, *40*, 1163–1177.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

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

Koehler, R. A. (1974). Overconfidence on probabilistic tests. *Journal of Educational Measurement*, *11*, 101–108.

Kooperberg, C. (2016). *logspline: Logspline Density Estimation Routines*. Retrieved from https://CRAN.R-project.org/package=logspline

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 1–12.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., … Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*, 75–84.

Lindner, M. A., Strobel, B., & Köller, O. (2015). Multiple-Choice-Prüfungen an Hochschulen? Ein Literaturüberblick und Plädoyer für mehr praxisorientierte Forschung [Are Multiple-Choice Exams Useful for Universities? A Literature Review and Argument for a More Practice Oriented Research]. *Zeitschrift Für Pädagogische Psychologie*, *29*, 133–149.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*, 437–448.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

McGee, R. K. (1962). The relationship between response style and personality variables: The measurement of response acquiescence. *The Journal of Abnormal and Social Psychology*, *64*, 229–233.

Meisters, J., Diedenhofen, B., & Musch, J. (2018). Eyewitness identification in simultaneous and sequential lineups: an investigation of position effects using receiver operating characteristics. *Memory. Advance online publication*.

Mickes, L., Flowe, H. D., & Wixted, J. T. (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 test-wiseness. *Educational and Psychological Measurement*, *25*, 707–726.

Morey, R. D., Rouder, J. N., Pratte, M. S., & Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, *55*, 368–378.

Norman, D. A., & Wickelgren, W. A. (1969). Strength theory of decision rules and latency in retrieval from short-term memory. *Journal of Mathematical Psychology*, *6*, 192–208.

Papenberg, M., Willing, S., & Musch, J. (2017). Sequentially Presented Response Options Prevent the Use of Testwiseness Cues in Multiple-Choice Testing. *Psychological Test and Assessment Modelling*, *59*, 245–266.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A'and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review*, *10*, 556–569.

Paulhus, D. L., & Harms, P. D. (2004). Measuring cognitive ability with the overclaiming technique. *Intelligence*, *32*, 297–314.

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology*, *84*, 890–904.

Peterson, D. J. (2014). *Dothraki: A conversational language course based on the hit original HBO series Game of Thrones*. New York, NY: Living Language.

Pleskac, T. J., Wallsten, T. S., Wang, P., & Lejuez, C. (2008). Development of an automatic response mode to improve the clinical utility of sequential risk-taking tasks. *Experimental and Clinical Psychopharmacology*, *16*, 555–564.

QuestBack. (2017). *Unipark EFS Survey*. Retrieved from http://www.unipark.de

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin*, *63*, 129–156.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: SAGE Publications, Incorporated.

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra: Psychology*, *2*, 1–12.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2016). *afex: Analysis of Factorial Experiments*. Retrieved from https://CRAN.R-project.org/package=afex

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*, 137–149.

Stoeber, J., Otto, K., Pescheck, E., Becker, C., & Stoll, O. (2007). Perfectionism and competitive anxiety in athletes: Differentiating striving for perfection and negative reactions to imperfection. *Personality and Individual Differences*, *42*, 959–969.

Swineford, F. (1938). The measurement of a personality trait. *Journal of Educational Psychology*, *29*, 295–300.

Swineford, F. (1941). Analysis of a personality trait. *Journal of Educational Psychology, 32*, 438–444.

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

Thoma, G.-B., & Köller, O. (2018). Test-wiseness: Ein unterschätztes Konstrukt? [Test-wiseness: an underestimated construct?]. *Zeitschrift für Bildungsforschung, 8*, 1–18.

Thorndike, R. L. (1938). Critical note on the Pressey Interest-Attitudes Test. *Journal of Applied Psychology*, *22*, 657–658.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment*, *23*, 279–291.

Wikipedia. (2017). List of islands of Indonesia — Wikipedia, The Free Encyclopedia. Retrieved from https://en.wikipedia.org/wiki/List_of_islands_of_Indonesia

Wiley, L. N., & Trimble, O. C. (1936). The ordinary objective test as a possible criterion of certain personality traits. *School and Society*, *43*, 446–448.

Willing, S. (2013). *Discrete-Option Multiple-Choice: Evaluating the Psychometric Properties of a New Method of Knowledge Assessment*. (Doctoral dissertation, Heinrich-Heine University, Duesseldorf, Germany). Retrieved from http://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=27633.

Willing, S., Ostapczuk, M., & Musch, J. (2015). Do sequentially-presented answer options prevent the use of testwiseness cues on continuing medical education tests? *Advances in Health Sciences Education*, *20*, 247–263.

Ziegler, M. (2015). "F*** You, I Won't Do What You Told Me!" – Response Biases as Threats to Psychological Assessment. *European Journal of Psychological Assessment*, *31*, 153–158. doi:10.1027/1015-5759/a000292

Ziegler, M., Kemper, C., & Rammstedt, B. (2013). The vocabulary and overclaiming test (VOC-T). *Journal of Individual Differences*, *34*, 32–40.

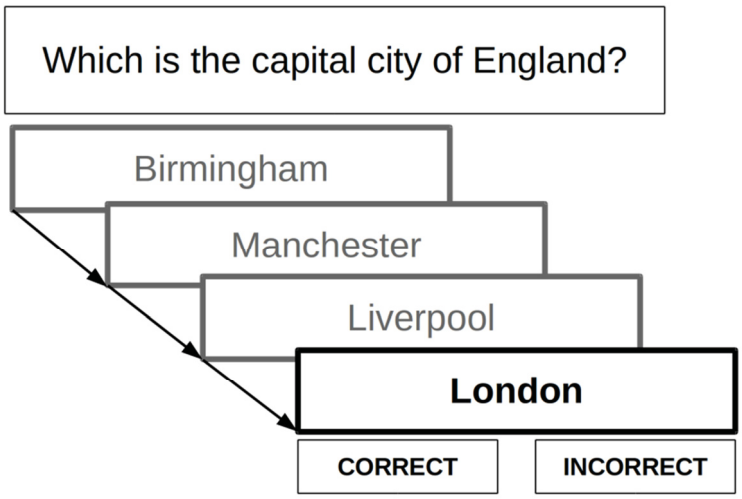

Figure 1. Illustrative example of a DOMC item in which the solution ("London") is presented as the fourth answer option, after the presentation of three distractor options ("Birmingham", "Manchester", and "Liverpool") in sequential order. The acceptance of one of these three distractor options would be recorded as a false alarm, while not accepting the correct solution would be recorded as a miss. One point is awarded if all distractors presented prior to the solution are rejected and the solution is rightfully accepted.

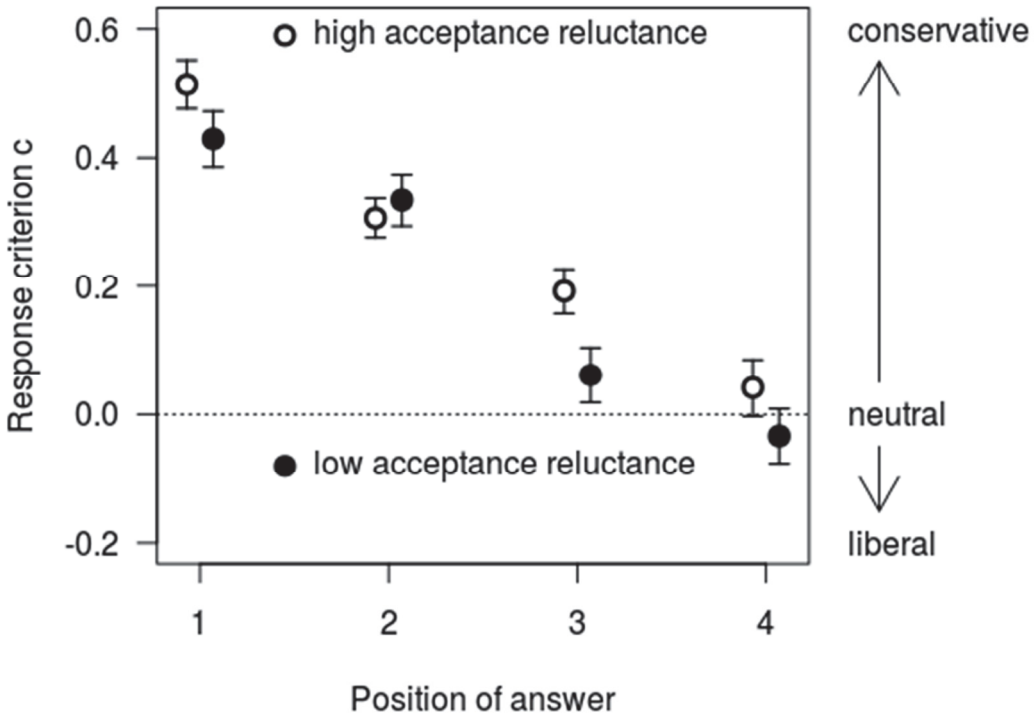

Figure 2. The average response criterion and its associated standard error by option position and acceptance reluctance, as determined by a median split of the number of options requested in no-solution items.

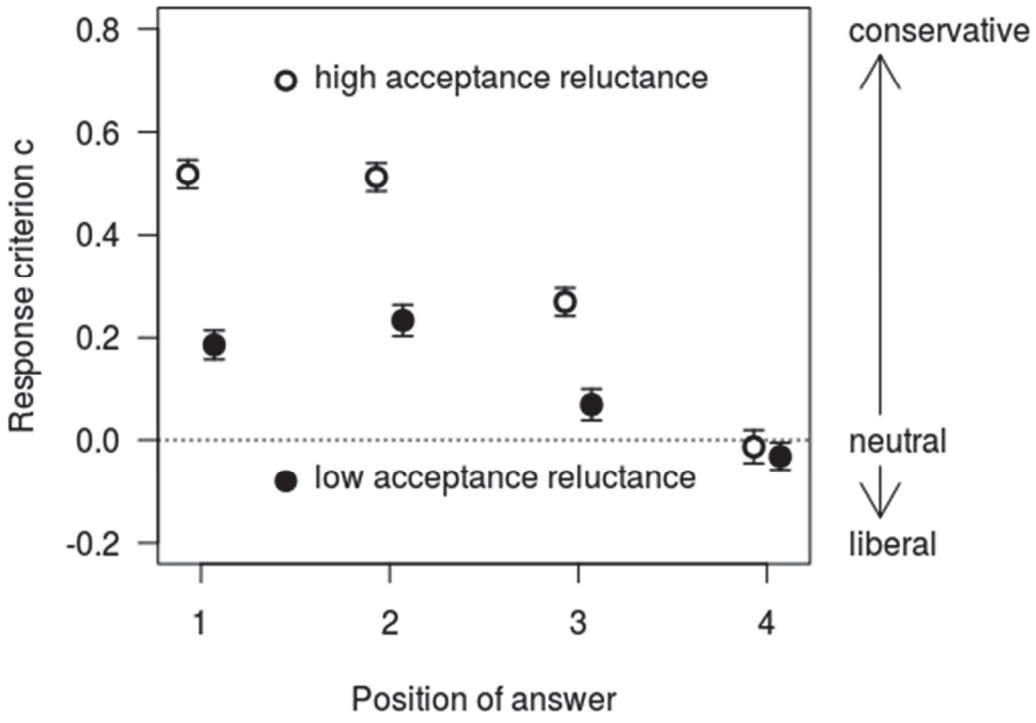

Figure 3. Test-takers responded more conservatively when high acceptance reluctance was awarded with a larger payoff. This effect was most pronounced for early options. Error bars indicate standard errors.

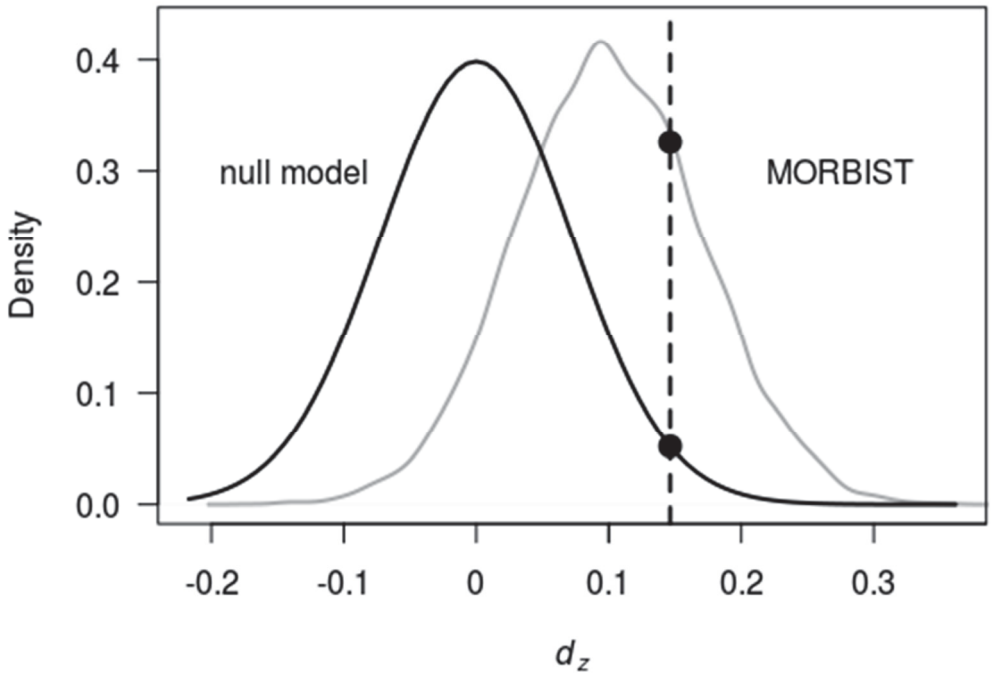

Figure 4. Effect sizes as predicted under the null hypothesis and the MORBIST simulation. The dotted line indicates the observed effect size in Study 3.

Table 1. An overview of the correlates of acceptance reluctance.

	1	2	3	4	5	6	7	8	9
1 Acceptance reluctance	.63	—	—	—	—	—	—	—	—
2 Knowledge (DOMC)	.28**	.77	—	—	—	—	—	—	—
3 Knowledge (MC)	.17*	.71**	.72	—	—	—	—	—	—
4 Balloon analogue risk task	.03	-.01	.04	.81	—	—	—	—	—
5 Need for cognition	.03	.19**	.17*	.07	.76	—	—	—	—
6 Perfectionism	-.03	.10	.04	.01	.21**	.91	—	—	—
7 Overclaiming	.01	.31**	.30**	.01	.28**	.01	—	—	—
8 Self-rated knowledge	-.04	-.36**	-.35**	-.05	-.15*	-.11	-.24**	—	—
9 Self-rated intelligence	-.03	-.17*	-.21**	.06	-.14*	-.20**	-.14*	.79**	—
10 Self-rated risk taking	-.05	.10	.07	.06	.13	-.10	.29**	-.09	-.02

Note. For indices consisting of more than one test unit, the diagonal reports Cronbach's α. * $p < .05$; ** $p < .01$.