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 × 3 (testing format [DOMC, MC] × 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]), $t(60) = 2.61$,

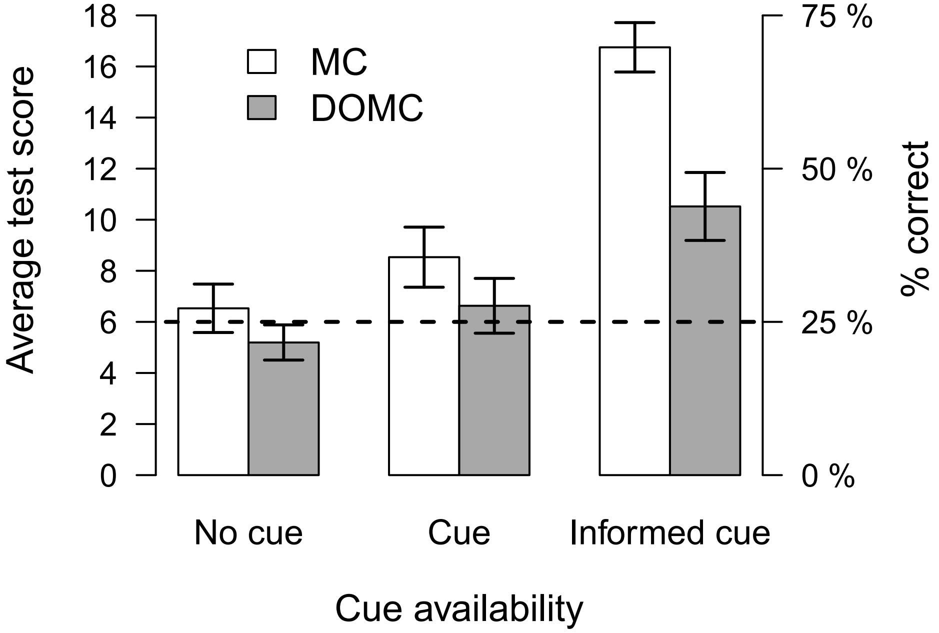

Figure 1: Test scores and their 95 % confidence intervals are shown as a function of (1) the two testing formats multiple-choice (MC) and discrete-option multiple-choice (DOMC), and (2) the availability of testwiseness cues. The dashed line indicates the chance level of 25 %, which is the expected test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

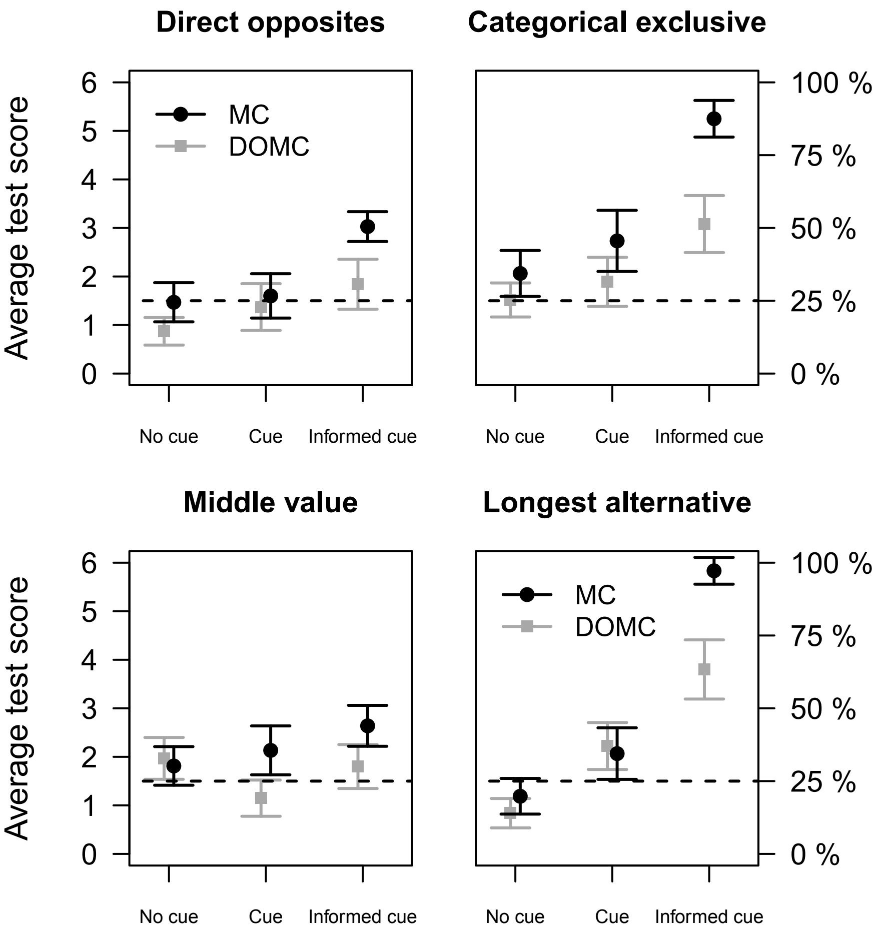

Figure 2: Average test scores and their 95 % confidence intervals are shown by cue type, test format, and cue availability. Each testwiseness cue was included in six testwiseness items. The dashed line indicates the chance level of 25 %.

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_G^2
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 testwiseness, 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