Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 22 Number 3, May 2017 ISSN 1531-7714

Negatively-Worded Multiple Choice Questions: An Avoidable Threat to Validity

Neville Chiavaroli, *The University of Melbourne*

Despite the majority of MCQ writing guides discouraging the use of negatively-worded multiple choice questions (NWQs), they continue to be regularly used both in locally produced examinations and commercially available questions. There are several reasons why the use of NWQs may prove resistant to sound pedagogical advice. Nevertheless, systematic inspection of item-level analysis often reveals anomalous behavior of NWQs on high-stakes examinations, due to otherwise high-performing students selecting the incorrect option for those questions. Highlighting the negative term as commonly recommended does not prevent this, since both anecdotal and empirical evidence suggests that many students answer the question as if it were positively phrased. The continued use of NWQs in high-stakes examinations poses a significant threat to the validity of interpretation based on these assessments. This is a form of 'construct-irrelevant variance' within the control of the item writer, and is therefore completely avoidable.

Among the many recommendations given to question writers for writing single best answer MCQs, the advice to avoid negative questions is one of the most common. Most MCQ drafting guidelines list this as a key principle, while a host of university and organizational 'house rules' for developing examinations also repeat the recommendation. In terms of frequency of citation, one review of educational textbooks noted that 31 of the 35 authors specifically advise against negatively-worded MCQs (NWQs) (Haladyna and Downing, 1989a), while a more recent review of locally produced high stakes examinations in the field of Nursing listed NWQs as the second most common question writing flaw (Tarrant et al., 2006). Yet it has been estimated that between 10-20% of medical examinations contain NWQs (Rodriguez, 1997; Harasym et al., 1992). Why, then, does this guideline seem to be so often ignored? This paper seeks to explore and understand this situation, and to reiterate the key justifications for avoiding negative wording in single best answer MCQs in summative assessments.

The Appeal of the NWQ

Despite the common recommendations against their use, it seems that writing NWQs serves a purpose for many examiners and teachers. The reasons for this continuing practice has seldom been explored in the literature, but deserves to be considered, especially if one seeks to effect change in educational practice. The following outlines some of the reasons why NWQs may remain appealing despite recommendations that they be avoided, based in part on the author's experience of faculty development sessions on question writing workshops for university teachers.

1. Convenience

Academics often state that they find it hard to set questions with three or four incorrect but plausible distractors. The NWQ format alters the balance in the writer's favor – true statements can be used as distractors, leaving the writer to devise a single incorrect statement which will act as the 'correct' response for a NWQ. This would help explain the

prevalence of the practice amongst novice question writers and students (Chéron et al, 2016), although even experienced writers are known to draft NWQs. In some ways, it can be seen as an intuitive response to the not inconsiderable challenge of devising plausible distractors.

2. The qualified nature of the recommendation

Closer inspection of drafting guidelines will reveal that the recommendation against using NWQs is seldom expressed in an absolute manner. Rather, the guideline is commonly qualified by a statement to the effect that NWQs are acceptable under certain circumstances, or may be legitimate if used 'when necessary' – so long as the negative term is emphasized in some way. Two examples include: 'Use negatives sparingly. If negatives must be used, capitalize, underscore embolden or otherwise highlight' (McKenna and Bull, 1999) and 'Negative stems may be appropriate in some instances, but they should be used selectively' (Collins, 2006). Even the authors of a highly influential MCQ writing guide in the field of medical education (Case and Swanson, 2002)[1] leave the door ajar on NWQs:

Avoid negatively phrased items (eg, those with 'except' or 'not' in the lead-in). If you must use a negative stem, use only short (preferably single word) options.

Unfortunately, clauses such as 'when necessary' and 'if you must' have the potential to undermine the credibility and impact of the recommendation against using NWQs, perhaps even legitimizing the practice. At the same time, they leave the novice question writer unclear about the kind of considerations required to determine the appropriateness of using an NWQ in a given situation. As a result, the exception is easily applied to one's own question-writing quandary, and writers may be inadvertently aided in rationalizing their natural impulse towards framing their question negatively. Provisos such as emphasizing the negative term or using single word options do not help, as they imply that doing so will mitigate any potential problems with the use of NWQs. As we shall see, this expectation is unfounded.

[1]Note that while this paper draws mainly on literature and practices from the field of medical and health professional education, its arguments and conclusions remain applicable to other disciplinary areas.

3. Clinical fidelity claims

In the context of medical and health professional education, at least, question writers often argue that the nature of clinical practice frequently involves reasoning negatively. It would also seem to be what many writers have in mind when they concede the occasional use of NWQs. For example, Harasym et al (1992) write that: 'Negation in the stem should only be used when it is critical for a student to know what to avoid or what is not the case', while one university MCQ Writing Guide suggests that 'negative items are appropriate for objectives dealing with health or safety issues, where knowing what not to do is important' (Burton et al., 1991). In other words, the NWQ is taken to represent a genuine aspect of clinical decision-making, and its use is therefore justified on the basis of cognitive fidelity. Medical educationalists, however, usually counter by arguing that such knowledge is better assessed using precisely the terminology used in medical practice – such as identifying the relevant contraindication or risk – rather than framing the question in a structurally negative way.

A related argument is the claim made that providing true statements for the majority of options makes the MCQ a useful educational tool, since the student reads mainly correct information, as opposed to positively-worded questions which contain predominantly incorrect information as distractors. This argument has occasionally been expressed to the author during assessment workshops, as has also been reported by Tamir (1993). However, this notion would seem to run counter to what we know about the conditions required for effective learning, a major part of which includes the provision of timely and targeted feedback (Ramsden 2003). Such claims of incidental learning also seem counter-intuitive in summative examination contexts, where students' focus tends to be on maximizing their score rather than learning new information.

In any case, the increasing emphasis in educational circles (eg Lemons & Lemons, 2013; Tractenberg, 2013) on writing questions which assess applied reasoning, rather than isolated factual recall, means that arguments about the relative merits of correct and incorrect distractors are becoming redundant. Instead, questions are increasingly being written in terms of asking students to determine the most appropriate or likely response in the context provided, so that the options are in themselves neither correct nor incorrect,

but rather more or less appropriate for the given context. This practice helps avoid the dilemma of whether to word the lead-in positively or negatively, since in the interests of plausibility all options should be theoretically appropriate under different circumstances. Although negatively-oriented questions could still be written by asking students to identify the least likely or appropriate response, the potential problems with this use of a negative orientation remain the same, as we shall soon see.

While the above explanations are not offered in any way to support of the use of NWQs, they should nevertheless caution educationalists over the language we use when discussing the practice. The rationale for avoiding NWQs is not necessarily self-evident, nor, as we have seen, is it always expressed in unambiguous terms. What is required, rather, is clear and sound educational justifications for the principle, and ideally, some evidence supporting it. I therefore now turn to those justifications and present some data as evidence in support of the principle to avoid NWQs

Justifications for Avoiding NWQs

The most common reason offered for avoiding the use of NWQs is the risk of introducing a 'double negative' – that is, the occurrence of a negative term in both the question lead-in and at least one of the options. Most commentators readily acknowledge that the mental processing required to understand and apply the particular logic of the English double negative is both complex, especially in the context of high-stakes examinations, and of little direct relevance to the knowledge or understanding being tested (eg Frary, 1995; Vahalia et al, 1995). The challenge for non-native English language speakers is even greater (Young, 2008). While it is probably the case that most double negatives occur unintentionally, the NWQ format nevertheless creates the precondition for double negatives, whether intended or not. For many assessment experts, this risk alone is sufficient justification for the avoidance of NWQs in high-stakes examinations – even when their effect may not be noticeable statistically, as Frary (1995) has previously discussed within this journal.While such items may appear to perform adequately empirically, this is probably only because brighter students who naturally tend to get higher scores are also better able to cope with the logical complexity of a double negative. The issue of the usefulness of statistical data to determine the quality of an item will be addressed shortly, but let us for now proceed under the assumption that careful construction and diligent editing can eliminate the presence of any unintended double negatives. Why else should NWQs be avoided?

Consideration for non-native English language speakers remains a major factor even when the duplication of a negative term is avoided. Simply wording a statement in the negative renders comprehension more complex, and this effect is exacerbated by testing contexts, thus adding to cognitive load and test anxiety (Abedi, 2006; Mestre, 1998; Trumbull and Solano-Flores, 2011; Young, 2008). The cognitive load of ideas expressed in negative form has been estimated as occupying 'twice as much space' in working memory as the corresponding positive form of the question (Tamir, 1993). When non-native English language speakers are involved, the demand is likely to be even greater and affect such respondents differentially. Unless the test happens to be on the understanding of English negation, negative phrasing is therefore likely to constitute a significant threat to the validity of the assessment.

Another key reason given in the literature for avoiding NWQs, or at least minimizing their use, is the concern that the negative orientation of the question may simply be missed (eg McDonald, 2013). This concern is frequently confirmed anecdotally by students during feedback discussion of MCQ results. Many examiners will attribute this to haste or carelessness on the part of the student, particularly when they have gone to the trouble of emphasizing the negative term in some way, as the educational textbooks frequently recommend. But the threat to validity of interpretation of results remains nonetheless.

More recent justifications for the avoidance of NWQs on pedagogical grounds have stemmed from the desire to improve the validity of the MCQ format in general (Haladyna and Downing, 1989b; Case and Swanson, 2002). Scholars point to the increased risk of introducing associated technical flaws, such as heterogeneous options or low cognitive levels, which NWQs appear to promote (Karegar Maher et al, 2016). Others point out that NWQs are rarely consistent with the kind of educational outcomes we expect from students. As one writing guide puts it, 'educational content tends not to be learned as a collection of non-facts or false statements, but, one would think, is likely stored as a collection of positively worded truths'

(University of Kansas, 2005). Moreover, the nature of the NWQ format means that the demand of the question rests largely on the obviousness of the incorrect statement. As Harasym and colleagues (1992) note:

(w)hen the examinee selects the false alternative, it is assumed that the student also knows the true aspects of the knowledge being tested. This assumption may not always be correct. Understanding of what is false may not necessarily indicate an understanding of what is true.

This observation, of course, could also be made of positively-worded questions; recognizing the correct (true) response does not necessarily mean that the student knows the other options are actually untrue. But given that the population of potentially incorrect answers about a topic or phenomenon is virtually infinite, it seems pointless to worry about recognition of incorrect statements instead of correct knowledge.

The potentially unlimited choice of a suitable key in NWQs has other implications. A key indicator of a well-written MCQ, according to many writing guidelines, is that the question should allow respondents to formulate a correct answer without needing to first look at the available options – a criterion commonly referred to as the 'cover options test' (Case and Swanson, 2002). Clearly, the NWQ fails this quality criterion. When faced with a negatively-worded question, a student cannot know which particular incorrect fact will be presented in the list of options as the key. This renders the question task solely one of identification, rather than generation followed by identification. This serves to further reinforce the inauthenticity of the NWQ, and, once again, its validity as an assessment format.

Finally, a further threat to validity posed by NWQs may be inferred from the literature relating to questionnaires and attitudinal surveys. The purposeful inclusion of negatively-worded items has been standard practice in attitudinal surveys for many years, in order to minimise the potential effects of response bias (van Sonderen et al, 2013). However, evidence is mounting that this practice introduces other threats to validity (Weemes et al, 2003; Roszkowski and Soven, 2010; van Dam et al, 2012; van Sonderen et al, 2013). As one study concludes, respondents either 'process positively worded items differently than negatively-worded items or [they] do not read the negatively worded items as carefully as they do positively worded items' (Weems et al, 2003). Whichever the case, this is increasingly being recognised as an unacceptable threat to the validity of the results, and survey researchers are increasingly recommending to avoid the practice.

Yet, in spite of the above rationales, NWQs remain in use in summative assessments in both educational and credentialing contexts. It would appear that pedagogical justifications are not sufficient to guide or change educational practice. Unfortunately, as we shall see, the empirical evidence can also be inconclusive. In the remainder of this paper, I argue that educators need to consider carefully where to look for the relevant evidence. I subsequently present and discuss several NWQs with associated performance data in order to make the nature of validity threat posed by such questions more explicit and, hopefully, more compelling.

Empirical Evidence for the avoidance of NWQs

Test-level statistics would seem to be a natural starting point for evaluating the potential impact of NWQs on test scores, and several studies have attempted to do just that. Downing (2005), for instance, demonstrated that flawed questions in general (including NWQs) are generally deleterious to student performance on examinations. He calculated that the median variance in test scores contributed by flawed test questions was 20%, and could be as high as 40% on some examinations. As a result of such flaws, three out of the four examinations analyzed were more difficult and students were less likely to achieve a passing score on these examinations. Harasym and colleagues (1992) reported several studies that concluded that NWQs tended to be more difficult than positively-worded versions, due in part to the inherently greater cognitive load negatively-oriented questions require, although the authors themselves questioned the results due to the different wording contained in the positive versions. Tamir (1993) found that items assessing high cognitive levels tended to be more difficult than positively-worded versions, an effect he attributed to the fact that the associated information processing involved 'more steps and is more complex than in the positive mode'; low cognitive items, however, showed no difference in statistical properties according to orientation of the stem. In general, though, scholars tend to conclude that the

impact of NWQs on test characteristics is actually varied and unpredictable (Violato and Marini, 1989; Rodriguez, 1997; Carter and Miller, 2006). In other words, the evidence against NWQs obtained from overall test-level data is generally inconclusive, since they produce mixed or negligible results in terms of the difficulty, average discrimination and reliability of NWQs.

The ambiguous nature of such evidence of the impact of NWQs is confirmed by the following data drawn from a first year medical course at the author's institution. While NWQs are no longer used in the course, there was a period in the recent past when such questions were not only tolerated, but used freely, in one case making up approximately half of the examination. Fortuitously, from both a pedagogical and research perspective, the NWQs in this case were administered in blocks (presumably in order to minimize the potential disorienting effect of the alternating direction of the stem). Such a design, it turns out, provides a useful opportunity in the form of a 'natural experiment'. Figure 1 and Table 1 display data obtained from the examination under discussion.

To judge from the above data alone, one might conclude that there was little to be concerned about regarding NWQs. There is minimal difference in the statistical characteristics of the positively and negatively worded questions, and the internal consistency, overall facility, standard deviation and average discrimination index are very similar, as confirmed by the relatively high correlation between the two groups of items. But this should not surprise educators or psychometricians – as the relative difficulty and discriminative power of NWQs (or indeed any question format) depends greatly on the *overall* quality of the questions. The only clues in the above data that the NWQs may be less robust psychometrically are to be found in the slightly lower reliability (although the fewer questions is a factor) and, more tellingly, in the greater proportion of questions with discrimination indices below 0.20 (the

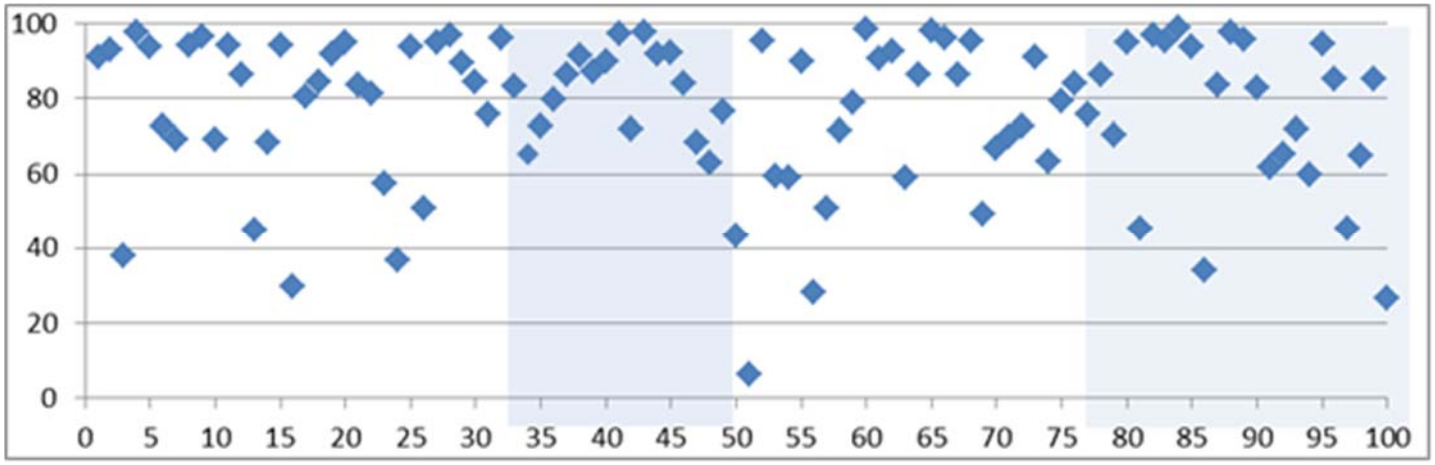


Figure 1. Distribution of item facilities (vertical axis, %) for end of year examination with positively and negatively worded questions (first year medicine, 100 MCQs, 2 hours, N=292)

Table 1. Test data for end of year examination with positively and negatively worded questions (first year medicine 2007, 100 MCQs, 2 hours, N=292)

	N	Internal Consistency	Average Facility	SD	Average DI	% of Qs with DI <0.20	Correlation
PWQ	59	0.78	76.8%	5.82	0.27	19%	0.76
NWQ	41	0.73	77.4%	4.48	0.27	29%	
Entire test	100	0.86	77.0%	9.66	0.27	23%	

conventional rule of thumb for acceptability of the index[2]). Nevertheless, the key point here is that such aggregated statistics are in general not helpful for the issue under consideration. NWQs are *not* inherently easier or more difficult than positively-worded questions; they are as difficult or as discriminating as the clarity of the lead-in, the obviousness of the key, the plausibility of the distractors, and the overall substantive content allow. As Tarrant and colleagues have rightly noted in relation to NWQs:

> … to ensure there is no ambiguity in the question, item-writers often make the correct answer (the incorrect option) so obviously incorrect that students can easily spot the answer and the question becomes too easy to adequately discriminate between the most and least able students in the test (Tarrant et al, 2006).

Such considerations are fundamental aspects of the 'art' of item writing (Ebel, 1951). Given Frary's observation noted earlier on the ambiguity of test-level data for double negative questions, we should therefore not be surprised to find that test-level data does not provide clear or consistent evidence of the psychometric inferiority of NWQs. This is looking for the evidence in the wrong place.

A parallel may be drawn with the fundamental difference in educational assessment between reliability and validity. A test that measures a coherent domain consistently will likely show high reliability, but this will not necessarily mean it is valid for the intended purpose. Similarly, any potential problem with NWQs will not necessarily be observed systematically at test level. Incomplete evidence, or the wrong kind of evidence, can be misleading. What is required in evaluating the potential impact of NWQs is information relating to how they perform intrinsically in terms of the information they provide about the students' knowledge and understanding of a particular topic, as sampled by the question. For this, we need to drill down further to look at the psychometric properties of the individual question.

Item-level data

When analyzing individual questions, the discrimination index (DI) and distractor analysis are particularly valuable. The DI reflects the association between performance on an individual question and performance on the test overall. It is commonly calculated as a correlation coefficient for each question option, and figures normally range between 0.5 and -0.5. However, as noted above, a positive value of above 0.20 has been the conventional threshold of acceptability, indicating a relatively strong association between the selection of a particular option and successful overall performance on the test (Chiavaroli and Familari, 2011). A negative value on the other hand indicates an inverse relationship between item and test performance, while a value around zero indicates no particular association between selection of an option and performance on the test. Questions which are in some way anomalous or inconsistent with the majority of questions on the test (or sub-test, depending on the level of analysis) will have low or negative DIs on the designated key. While the DI represents the coefficient value for the key, 'distractor analysis' enables consideration of the above-mentioned relationships for all options in an MCQ. Often, when there is a low or negative DI for the key, one or more of the distractors will have positive coefficient values, indicating an (unexpected) association between choosing an *incorrect* option, and otherwise generally successful performance on the test. This is clearly counter to expectations within a relatively coherent domain of assessment.

The following question, which was used in a final summative examination for first-year medical students in two different years, will illustrate these considerations.

EXAMPLE 1

In some situations exercise can lead to skeletal muscle injury. Which one of the following statements is **NOT CORRECT**?

a. Lengthening (pliometric or eccentric) contractions are most likely to cause muscle damage.

b. Elevation of intramuscular calcium is prevented following muscle damage to reduce the extent of damage.

c. Elevated levels of muscle specific enzymes appear in the plasma following muscle damage.

d. The area of muscle damage can be repaired following activation of satellite cells.

[2] Similar results are to be found in another exam from the same period in which NWQs made up 26% of the entire paper. The relevant data is provided in Appendix 1.

e. Muscle damage can decrease the maximum force output of the damaged muscle.

Table 2. Item Analysis for Example 1

Occasion 1: 296 first year medical students (full cohort); final exam of 100 questions (2006)

Option	Percentage of students	Discrimination Index*
A	2%	-0.08
B (key)	77%	0.07
C	10%	0.03
D	7%	-0.03
E	4%	-0.11

Occasion 2: 84 first year medical students (full cohort); final exam of 100 questions (2010)

Option	Percentage of students	Discrimination Index*
A	Nil	N/A
B (key)	77%	0.06
C	8%	0.03
D	5%	-0.09
E	10%	-0.05

**Exam data calculated using Quest software (Adams and Khoo, 1998)*

Apart from the very similar data generated across two different cohorts, this question is a useful example of a relatively easy question (77% facility on both occasions) which has a low DI, representing an anomalous pattern of responses on both occasions. The DI informs us that the group who selected the keyed response varied in terms of their overall performance. In other words, many students who otherwise performed well on the exam failed to respond correctly to this question (most of whom apparently opted for option C). The examination review panel, including the question writer and other subject matter experts, confirmed option B, an incorrect statement, as the key (and, incidentally, option C as a true statement and therefore an incorrect response to the question). Thus, a major concern with this question, from a psychometric perspective, is to understand why option many high-performing students were unable to identify option B as the key, when so many of their lesser performing peers were able to do so.

A clue to this anomaly is provided by the performance of the highest-achieving student in the 2010 cohort on this examination (see Appendix 2 for a graphical representation of the student's performance). The student in question achieved 99% of questions correct out of 100; the only question the student failed to get correct was the above question. This is very surprising given that the majority of the student's peers were able to answer correctly. In other words, based on the student's performance on the test overall, this question should have posed minimal challenge for this student. In the absence of plausible alternative explanations based on the content of the question, which the examination panel was unable to provide, the most likely explanation for the lapse on this question would appear to be accident rather than ignorance. Taken with the question performance data on two cohorts, the item-level data strongly suggests that the negative orientation of the question was overlooked by several high-performing students, including the highest scorer.[3]

The following is another example of an NWQ with an anomalous pattern of responses. In this case, the question formed part of an examination for optometry candidates as part of a credentialing examination.

EXAMPLE 2

Tinted lenses for outdoor use are LEAST likely to benefit a person with which of the following ocular conditions?

A. Holmes-Adie pupil
B. Retinitis pigmentosa
C. Keratoconus
D. Hemianopia

Table 3. Statistics for Example 2
(Credentialing examination, 72 candidates, 235 questions)

Option	Percentage of students	Discrimination Index*
A	6%	0.10
B	17%	0.27
C	8%	0.00
D (key)	69%	-0.27

**Exam data calculated using Quest software (Adams and Khoo, 1998)*

In this example, the negative DI (Table 3) for the correct response reveals that, as a group, the candidates

[3]Of course, many lower achieving students can be assumed to have also missed the negative term, but the discrimination index does not help identify those students, as an incorrect response in their case appears less anomalous.

who selected the correct answer (D) performed worse overall than the remaining 31% of candidates who selected one of the other responses. This was perplexing data for the examination committee, particularly given the high negative value for the DI, which the majority had correctly selected as the key. Discussion with the question writer and subject matter experts confirmed option D (Hemianopia) as the correct answer, that is, the least likely to benefit from tinted lenses; they also confirmed that this was considered a relatively easy question for the candidature. Yet it appeared to stump the best-performing candidates. The explanation for this outcome seemed to lie in the nature of option B (Retinitis pigmentosa). The experts noted that tinted lenses are in fact *especially* indicated for this condition (Eperjesi et al., 2002). In the absence of alternative content-based explanations, the most plausible explanation for the above pattern of responses is, once again, that in the process of working through each option, several high-achieving candidates overlooked or forgot the negative orientation of the stem. In so doing, they appear to have been drawn into selecting the *most appropriate* response, that is, the positive version of the question.

In many ways, this is perfectly understandable. In a clinical context, the most natural question to ask, and one which respondents would be expected to be asked most often, would be to consider and justify when tinted lenses *would* be appropriate, not when they wouldn't. The situation represented in this NWQ is therefore highly inauthentic – a practicing optometrist with tinted lenses in hand wondering which patient would *least* benefit from their use. Case and Swanson (2002) refer unfavorably to such questions as 'waiting room items' for this very reason. In answering this question, it appears that the best performing students, who presumably are well on the way to becoming effective practitioners, have responded to the question as they have been taught, not as they were asked - 'reflexly' connecting the management (tinted lenses) with the appropriate condition ('retinitis pigmentosa'). Emphasizing the negative term has not removed this risk. In this situation, it appears to have been an advantage to *not* have the high level of knowledge or reasoning which would instinctively draw one to the natural association between treatment modality and condition, since such learning may have interfered with the logic imposed by the negative structure of the question. This is particularly worrying in the context of a credentialing examination, where high stakes pass/fail decisions should be based on what a candidate knows and understands, not on whether they accidentally miss a negative term. Given the stakes involved, and bearing in mind that for some candidates such decisions can indeed come down to a single item, this risk can only be described as representing a significant threat to the validity of any decision based upon them.

Further examples of NWQs with similarly anomalous data are provided in Appendix 3.

Implications

The above examples suggest that one of the major risks of NWQs – missing the negative term – is a real one and appears to be a plausible explanation when high-performing respondents fail to answer correctly relatively easy NWQs. The author contends that if examiners look closely enough at their data, they will find many NWQs which behave in a similarly anomalous way. The arguments and examples offered in this paper obviously do not prove that this is always the reason for mis-performing NWQs, but rather demonstrate that the problem exists. A fuller idea of the extent of the problem may be gained from systematic inspection of item analysis data of local examinations. In the author's own school's post-administration review panels, for instance, we have noticed that on many occasions, NWQs, despite (now) occupying a very small proportion of test space when used, tend to be over-represented amongst the problematic questions. In most cases, there is no apparent content-based reason for the problematic response pattern.

Such anomalous performance data is frequently described in assessment circles as 'construct-irrelevant variance' (or CIV), which has been defined as 'systematic error (rather than random error) introduced into the assessment data by variables unrelated to the construct being measured' (Downing and Haladyna, 2004). In most cases, being able to consistently spot negative terms in an examination is unrelated to the ability to understand a particular content area. The presence of question flaws with the potential to introduce CIV into a student's test performance – even when it is not certain whether the irrelevant variance has occurred – is probably sufficient to constitute a threat to the validity of any conclusions we may wish to draw from a test which includes NWQs. When the kind

of evidence presented and discussed above emerges, then the threat to validity becomes even harder to ignore. While the consequences of compromised validity for a student scoring 99% on an examination may be minimal, for lower-achievers such consequences can be far-reaching and much more detrimental.

It remains to be studied exactly why and when students miss the negative orientation of NWQs. This would require in-depth, qualitative analysis of student thinking processes around NWQs. Evidence discussed above points to the high-level cognitive challenge of retaining the negative orientation in mind while endeavoring to work out the substantive challenge posed by higher reasoning questions. In such circumstances, students would need to be particularly vigilant against their natural – and normally educationally-advantageous – instinct to draw the relevant link between the two. But the usefulness of such research is in fact questionable, when the format itself seems to be inherently invalid, for the reasons presented in this paper.

Conclusion

Earlier we noted Frary's observation that data on double negative questions is frequently inconclusive, due to the capacity of brighter students to readily resolve the construct-irrelevant demands of such questions. This paper has attempted to explore and illustrate the opposite effect – where brighter students may actually be *more* likely to miss the negative orientation of a question. From a validity perspective, it is sufficient to know that this can and does happen; and that when it does, we cannot know whether the student failed the question due to ignorance or accident.

The examples and arguments presented in this paper therefore reveal only the tip of the iceberg – only when the higher performing students miss the negative are the risks inherent in the use of NWQs exposed. For the remainder of the cohort, any evidence that they have been similarly affected remains submerged beneath apparently sound empirical data. This is a further and less obvious problem with NWQs – their detrimental impact frequently remains impervious to the otherwise illuminating effect of item analysis.

What we should recognize, and what this paper has attempted to demonstrate further, is that emphasizing the negative term is no guarantee that it will not be overlooked (if you'll forgive the double negative in this non-testing context). The desire to emphasize the negative term is understandable and even admirable. It is the pedagogical equivalent of helpful signs, like 'mind the gap', alerting the public to potential hazards. But highlighting negative terms also exposes the inherent inauthenticity and pedagogical risk of such questions – test developers have the opportunity, and responsibility, to avoid creating the hazard in the first place.

References

Abedi, J. (2006). Language issues in item development. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Adams, R.J. & Khoo, S.T. (1996). *Quest: The Interactive Test Analysis System* (computer software). Version 2.1. Melbourne, Australia: ACER Press

Burton, S.J., Sudweeks, R.R., Merrill, P.F., & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty*. Brigham Young University Testing Services and the Department of Instructional Science, Provo, UT. URL: https://testing.byu.edu/handbooks/betteritems.pdf [accessed 23 December 2016]

Carter, J.T. & Miller, S.J. (2006). *Performance of positively and negatively worded questions on a classroom exam*. Paper presented at the annual meeting of the American Association of Colleges of Pharmacy. San Diego, CA,

Case, S.M. & Swanson, D.B. (2002). *Constructing Written Test Questions for the Basic and Clinical Sciences*. 3rd Ed (rev.). Philadelphia, PA: National Board of Medical Examiners.

Chéron, M., Ademi, M., Kraft, F., & Löffler-Stastka, H. (2016). Case-based learning and multiple choice questioning methods favored by students. *BMC Medical Education,* 16(41), 1-7.

Chiavaroli, N. & Familari, M. (2011). When majority doesn't rule: The use of discrimination indices to improve the quality of MCQs, *Bioscience Education,* 17(1):1-7; URL: http://dx.doi.org/10.3108/beej.17.8 [accessed 23 December 2016].

Collins, J. (2006). Education techniques for lifelong learning: Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radiographics*, 26(2), 543–551.

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: Theory and Practice,* 10(2), 133–143.

Downing. S.M. & Haladyna, T.M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education,* 38(3), 327–333.

Ebel, R.L. (1951). Writing the test item. In E. F. Linquist (Ed.), *Educational Measurement* (1st edn). Washington, DC: American Council on Education.

Eperjesi, F., Fowler, C.W., & Evans, B.J. (2002). Do tinted lenses or filters improve visual performance in low vision? A review of the literature. *Ophthalmic and Physiological Optics,* 22(1), 68–77.

Frary, R.B. (1995). More multiple-choice item writing do's and don'ts. *Practical Assessment, Research & Evaluation*, 4(11). URL: http://PAREonline.net/getvn.asp?v=4&n=11 [accessed 23 December 2016].

Haladyna, T.M. & Downing, S.M. (1989a). A taxonomy of multiple-choice item writing rules. *Applied Measurement in Education,* 2(1), 37–50.

Haladyna, T.M. & Downing, S.M. (1989b). Validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education,* 2(1), 51–78.

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

Harasym, P.H., Price, P.G., Brant, R., Violato, C., & Lorscheider, F.L. (1992). Evaluation of negation in stems of multiple-choice items. *Evaluation & the Health Professions*, 15(2), 198–220.

Karegar Maher, M.H., Barzegar, M., & Gasempour, M. (2016). The relationship between negative stem and taxonomy of multiple-choice questions in residency pre-board and board exams. *Research & Development in Medical Education,* 5(1), 32-35.

Lemons, P.P. & Lemons, J.D. (2013). Questions for assessing higher-order cognitive skills: It's not just Bloom's. *CBE-Life Sciences Education,* 12(1), 47-58.

McDonald, M.E. (2013). *The Nurse Educator's Guide to Assessing Learning Outcomes.* (3rd ed.). Burlington, MA: Jones and Bartlett.

McKenna, C. & Bull, J. (1999). *Designing effective objective test questions: An introductory workshop*. Computer Assisted Assessment (CAA) Centre, Loughborough University. Leicestershire, UK. URL: http://caacentre.lboro.ac.uk/dldocs/otghdout.pdf [accessed 23 December 2016].

Mestre, J.P. (1988). The role of language comprehension in mathematics and problem solving. In R. R. Cocking and J. P. Mestre (Eds.), *Linguistic and Cultural Influences on Learning Mathematics*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ramsden, P. (2003). *Learning to Teach in Higher Education.* (2nd ed.) London and New York: RoutledgeFalmer.

Rodriguez, M.C. (1997). *The art and science of item-writing: A meta-analysis of multiple-choice item format effects*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. URL: https://pdfs.semanticscholar.org/f39c/c381e1c3e8a3db6f5e9c884abbfbecf4b057.pdf [accessed 23 December 2016]

Roszkowski, M.J. & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, 35(1), 113–130.

Tamir, P. (1993). Positive and negative multiple choice items: How different are they? *Studies in Educational Evaluation,* 19(3), 311–325.

Tarrant, M., Knierim, A., Hayesm S.K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today* 26(8), 662–671.

Tractenberg, R.E., Gushta, M.M., Mulroney, S.E., & Weissinger, P.A. (2013). Multiple choice questions can be designed or revised to challenge learners' critical thinking. *Advances in Health Sciences Education, 18*(5), 945-961.

Trumbull, E. & Solano-Flores, G. (2011). The role of language in assessment. In M. del Rosario Basterra, E. Trumbull, and G. Solano-Flores (Eds.) *Cultural Validity in Assessment: Addressing Linguistic and Cultural Diversity.* New York, NY: Taylor and Francis.

University of Kansas. (2005). *Special Connections: Multiple-Choice Questions. Designing Multiple-Choice Questions.* The University of Kansas, Lawrence, KS. URL: http://www.specialconnections.ku.edu [accessed 23 December 2016].

Vahalia, K.V., Subramaniam, K., Marks, S.C. Jr, & De Souza, E.J. (1995). The use of multiple-choice tests in anatomy: Common pitfalls and how to avoid them. *Clinical Anatomy*, 8(1), 61–65.

van Dam, N.T., Hobkirk, A.L., Danoff-Burg, S., & Earleywine, M. (2012). Mind your words: Positive and negative items create method effects on the Five Facet Mindfulness Questionnaire. *Assessment,* 19(2),198-204.

van Sonderen, E., Sanderman, R., & Coyne, J.C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PLoS ONE* 8(7): e68967. doi:10.1371/journal.pone.0068967

Violato, C. & Marini, A.E. (1989). Effects of stem orientation and completeness of multiple-choice items on item difficulty and discrimination. *Educational and Psychological Measurement,* 49(1), 287-295.

Weems, G. H., Onwuegbuzie, A. J., Schreiber, J. B., & Eggers, S. J. (2003). Characteristics of respondents who respond differently to positively and negatively worded items on rating scales. *Assessment & Evaluation in Higher Education, 28*(6), 587-607.

Young, J.W. (2008). Ensuring Valid Content Tests for English Language Learners. *R&D Connections*, No. 8. Princeton, NJ: Educational Testing Service.

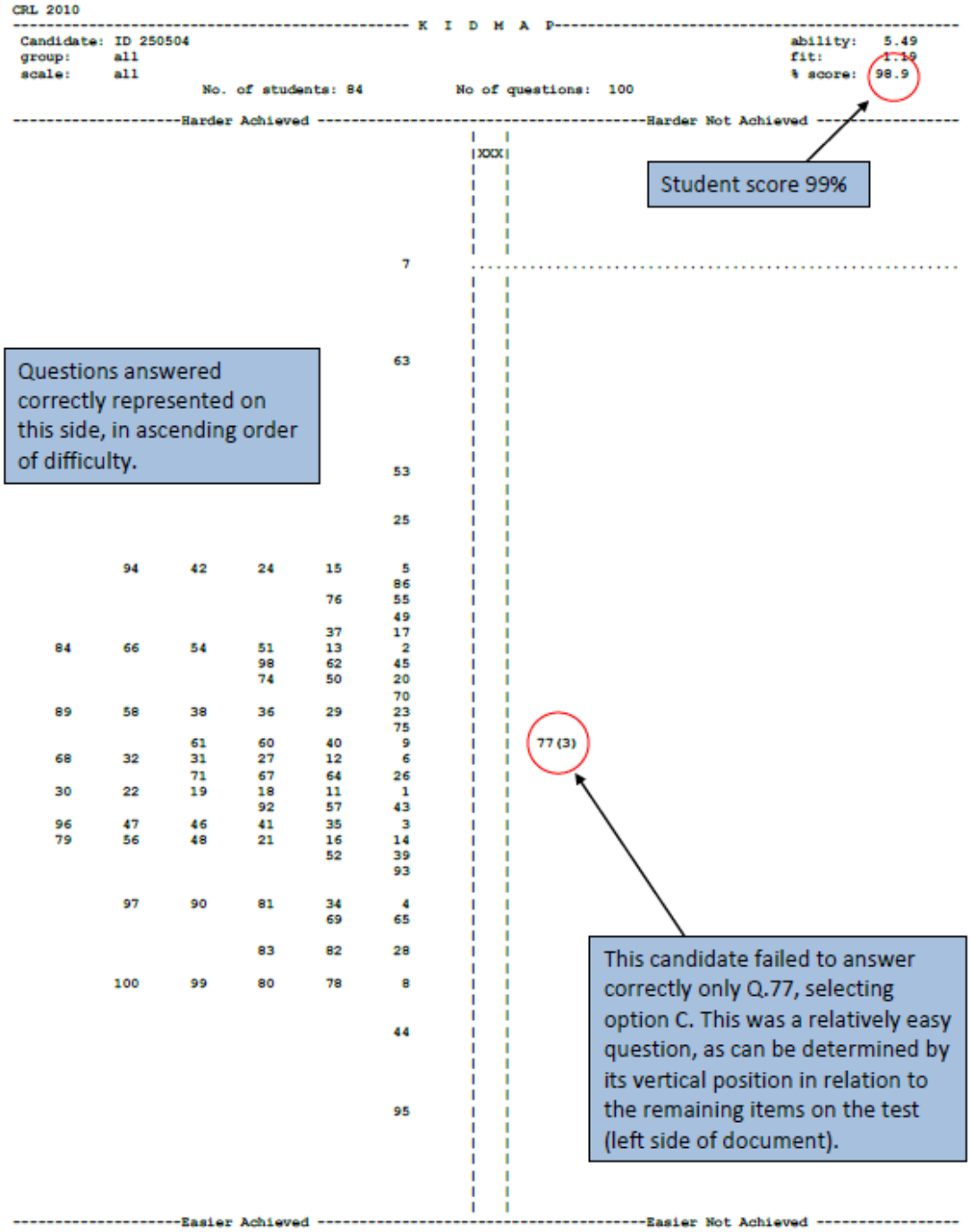
Appendix 1: Test data for summative examination with positively and negatively worded questions (First year medicine 2006, 100 MCQs, 2 hours, N=296)

(Data produced using Quest ver 2.1 (Adams and Khoo, 1998)

	N	Internal Consistency	Average Facility	SD	Average DI	% of Qs with DI < 0.20	Correlation
PWQ	74	0.85	72.0%	8.4	0.28	17.6%	0.71
NWQ	26	0.58	57.1%	3.3	0.24	30.8%	
Entire test	100	0.87	68.1%	11.0	0.27	21.0%	

Appendix 2: Pattern of responses to a first year summative medical examination by the highest-performing student (First year medicine, full cohort of 84 students, 2 hours, 100 questions)

(Data produced using Quest ver 2.1 (Adams and Khoo, 1998)



Appendix 3: Further NWQs with anomalous performance data (from First year medicine summative examinations)

(Data produced using Quest ver 2.1 (Adams and Khoo, 1998)

Q44 Which of the following statements regarding cardiac myocyte contraction is **NOT CORRECT**?

A. The binding of calcium to tropomyosin is important in the process of contraction.
B. Myocyte excitation causes extracellular calcium to move into the cytosol.
C. Extracellular calcium is the major source accounting for the increase in cytosolic calcium that triggers contraction.
D. Hormones such as adrenaline will increase intracellular calcium.
E. Drugs such as digoxin will increase intracellular calcium.

```
Item   44: item 44                          Infit MNSQ = 1.25
                                             Disc = 0.01

Categories        1 [0]     2 [0]     3 [1]     4 [0]     5 [0]    missing

Count               131        15       133         8         2          0
Percent (%)        45.3       5.2      46.0       2.8       0.7
Pt-Biserial        0.18     -0.27      0.01     -0.14     -0.12
Mean Ability       1.58      0.64      1.45      0.83      0.46         NA
StDev Ability      0.70      0.65      0.72      0.61      0.25         NA
```

Comment: Low discrimination for key; many high achieving students appear to have selected A instead.

Q60 Which of the following statements concerning rheumatic endocarditis is **NOT CORRECT**?

A. The vegetations that form on the heart valves are sterile.
B. The disease is caused by Streptococcal infection, usually of throat or skin.
C. The disease is a significant cause of morbidity in areas of Australia.
D. Acute rheumatic myocarditis usually resolves with no sequelae.
E. The vegetations lead to valve destruction with perforation.

```
Item   60: item 60                          Infit MNSQ = 1.02
                                           Disc = 0.10

Categories        1 [0]     2 [0]     3 [0]     4 [0]     5 [1]    missing

Count               89        21        33       133        20          0
Percent (%)       30.1       7.1      11.1      44.9       6.8
Pt-Biserial      -0.03     -0.07     -0.16      0.12      0.10
Mean Ability      0.94      0.83      0.68      1.08      1.31         NA
StDev Ability     0.66      0.79      0.70      0.70      0.84         NA
```

Comment: Low discrimination for key, though still positive; many high achieving students appear to have selected D instead. Note that the final phrase of option D makes this a double negative question, which may have impacted on the psychometric quality of the question.

Q90 Which **one** of the following changes in the cardiovascular system is **NOT** a physiological change of ageing?

A. Increase in both systolic and diastolic blood pressure levels.
B. Decrease in the resting heart rate.
C. Decrease in the maximal heart rate able to be achieved.
D. Increase in the risk of coronary artery disease.
E. An increase in the left ventricular mass.

```
Item   91: item 91                              Infit MNSQ = 1.20
                                                      Disc = -.04

Categories        1 [0]     2 [1]     3 [0]     4 [0]     5 [0]    missing

Count              165        63         3        67        27         0
Percent (%)       50.8      19.4       0.9      20.6       8.3
Pt-Biserial       0.01     -0.04     -0.06      0.17     -0.20
Mean Ability      1.65      1.59      1.25      1.88      1.15        NA
StDev Ability     0.77      0.76      0.94      0.63      0.66        NA
```

Comment: Negative discrimination for key; the highest achieving students tended to select D. Note that the wording of option B makes this a double negative question. This may have further impacted on the psychometric quality of the question.

Acknowledgment:

The author would like to acknowledge colleagues in the Melbourne Medical School and the Optometry Council of Australia and New Zealand for permission to use the questions discussed in this paper, and for their helpful discussions of the content of these questions. However, any errors of interpretation or explanation of the question content remain the author's responsibility.

Citation:

Chiavaroli, Neville (2017). Negatively-Worded Multiple Choice Questions: An Avoidable Threat to Validity. *Practical Assessment, Research & Evaluation*, 22(3). Available online: http://pareonline.net/getvn.asp?v=22&n=3

Corresponding Author

Neville Chiavaroli
Department of Medical Education, Melbourne Medical School
The University of Melbourne

email: n.chiavaroli [at] unimelb.edu.au