DUT
Dansk Universitetspædagogisk Tidsskrift

Tema Fra data til beslutninger

Årgang 14 nr. 26 / 2019

Titel **Screening for technical flaws in multiple-choice items. A generalizability study.**

Forfattere Lotte Dyhrberg O'Neill, Sara Mathilde Radl Mortensen, Cita Nørgård, Anne Lindebo Holm Øvrehus, Ulla Glenert Friis

Sidetal 51-65

Udgivet af Dansk Universitetspædagogisk Netværk, DUN

URL › http://dun-net.dk/

Screening for technical flaws in multiple-choice items. A generalizability study.

Lotte Dyhrberg O'Neill[a,1], Sara Mathilde Radl Mortensen[b], Cita Nørgård[c], Anne Lindebo Holm Øvrehus[d], Ulla Glenert Friis[e]

[a]SDU Universitetspædagogik, Syddansk Universitet, [b]Klinisk Institut, Aalborg Universitet, [c]SDU Universitetspædagogik, Syddansk Universitet [d]Klinisk Institut, Syddansk Universitet, [e]Sundhedsvidenskabeligt Fakultetssekretariat, Syddansk Universitet

Research article, peer-reviewed

Construction errors in multiple-choice items are quite prevalent and constitute threats to test validity of multiple-choice tests. Currently very little research on the usefulness of systematic item screening by local review committees before test administration seem to exist. The aim of this study was therefore to examine validity and feasibility aspects of review committee screening for item flaws. We examined the reliability of item reviewers' independent judgments of the presence/absence of item flaws with a generalizability study design and found only moderate reliability using five reviewers. Statistical analyses of actual exam scores could be a more efficient way of identifying flaws and improving average item discrimination of tests in local contexts. The question of validity of human judgments of item flaws is important - not just for sufficiently sound quality assurance procedures of tests in local test contexts - but also for the global research on item flaws.

Introduction

Multiple-choice tests are particularly useful and effective test formats in test situations where there is a need to test knowledge (factual or applied) across a wide range of different content topics in larger groups of students (Downing & Yudkowsky, 2009). This potential for broad sampling of topics is often pivotal for 'Constructive Alignment' (Biggs & Tang, 2007), and for the validity of the exam or test in question (Swanson, Norcini, & Grosso, 1987). Another distinctly positive feature of multiple-choice tests is that excellent evidence-based guidelines for constructing good quality items exist (Case & Swanson, 2002; Haladyna, Downing, & Rodriguez, 2002; Paniagua & Swygert, 2016). However, little is known about the usefulness of this body of recommendations as a basis for systematic item screening *before* test administration by local review committees.

Quality assurance measures can be applied both before and after multiple-choice exams. *Before* the exam, item writers can themselves attempt to review their own item drafts according to guideline criteria. Such reviews involve critical reading which is qualitative in nature. Subsequently, other reviewers (e.g. colleagues or external examiners) may repeat this process of qualitative reviewing and suggest item corrections or removal (Downing, 2004, 2006; Malau-Aduli & Zimitat, 2012). *After* the exam, quality assurance typically involves quantitative analyses of test results (quantitative review), followed by yet another round of qualitative reviewing, this

[1] Kontakt: ldo@sdu.dk

time focused on discounting selected test items which performed aberrantly because of identifiable flaws missed in the initial phases (Case & Swanson, 2002). In other words: decisions based on qualitative reviews of items are fundamental to the overall quality assurance of multiple-choice tests or exams in all stages (drafting, editing and grading).

Undetected item flaws may infuse exam scores with construct-irrelevant variance, which undermines the validity of the exam (Downing, 2002, 2003, 2005; Kane, 2006). Reliable flaw detection is therefore a sine qua non for optimal test validity. Studies have shown that in practice, writing flawed items is a very common event in local educational contexts (Downing, 2002, 2005; Downing & Yudkowsky, 2009; Hansen & Dexter, 1997; Jozefowicz et al., 2002; Masters et al., 2001; Palmer & Devitt, 2007; Rodríguez-Díez, Alegre, Díez, Arbea, & Ferrer, 2016; Tarrant, Knierim, Hayes, & Ware, 2006; Tarrant & Ware, 2008; Vahalia, Subramaniam, Marks, & De Souza, 1995), and it has been suggested that item writing is 'as much art as science' (Downing, 2005; Downing & Yudkowsky, 2009; Ebel, 1951; Haladyna et al., 2002). By logic extension, it would seem a reasonable hypothesis that a qualitative flaw detection process may also be 'as much art as science', since such a process relies on human reading, evaluation and decision skills. Even though guidelines quite clearly outline a number of straightforward flaw types to look out for, screening for item flaws very often involves both interpretation and subjectivity in our experience. However, very little research on reliability aspects of qualitative item screening by committees seems to exist to support or refute this hypothesis (Engelhard Jr, Davis, & Hansche, 1999). As a consequence, the number of board members necessary for reliable item screening processes in a review committee approach does not appear to be well documented either. The human resources necessary for securing exam validity may also have important consequences for the feasibility and acceptability of a test, particularly in local, smaller educational contexts. Apart from being valid, assessments must also be reasonably feasible for test administrators in terms of time and human resources necessary (Dent, Harden, & Hunt, 2017).

Validity assumptions

The current Standards for Educational Psychological Testing describes validity in these words:

> *Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations* (American Educational Research Association, 2014).

Test scores are typically used to support claims beyond the observed performance (Kane, 2013), and in such cases either implicit or explicit score interpretations exist. According to modern Validity Theory, construct validity evidence for such score interpretations may arise from all stages of testing, i.e., from the initial development of test content to subsequent implications of decisions based on test scores (American Educational Research Association, 2014; Kane, 2006). Kane suggested that score interpretations (or 'interpretative arguments') may be categorized to relate to the following four stages: scoring, generalization, extrapolation and decision (Kane, 2006).

Proposed score interpretations related to *scoring* in a multiple-choice exam could for instance be: the recording of students' electronic responses represents students' intended answers, the answer key for the test items is appropriate, and the answer key is applied accurately and consistently. Score interpretations relating to *generalization* in a multiple-choice exam could

be: the items in the test are representative of the universe of all possible items which are congruent with the learning outcomes and the teaching and learning activities of the course, and the sample of items is large enough to control for sampling error. Score interpretations relating to *extrapolation* could for example be: the test tasks require the competencies developed in the course, we may extrapolate expertise levels from the test scores, and there are no skill irrelevant sources of variability that seriously bias the interpretation of scores as measures of students' subject knowledge. The fourth category of assumptions about scores relates to *decision* and represents going from conclusions about test takers' competences to making a decision with implications. Score interpretations relating to decision in a multiple-choice test could for example be: students with no or low level of subject knowledge are unlikely to pass the test and progress in the program.

Such a chain of validity assumptions about multiple-choice exam scores may be challenged if test items are flawed. Technical flaws in multiple-choice items tend to 'pollute' the test with competing test constructs, so that it is no longer just subject knowledge which is measured. 'Testwiseness' and 'irrelevant difficulty' have been identified as competing test constructs in items with technical flaws (Case & Swanson, 2002). Testwiseness refers to test takers ability to answer items based on logic rather than on subject expertise, whereas irrelevant difficulty is item difficulty caused by a confusing presentation of item content rather than actual subject difficulty. If competing test constructs (like testwiseness and irrelevant difficulty) are sufficiently influential in the test situation, assumptions or score interpretations such as those outlined above relating to extrapolation and decisions may be compromised. The validity argument for test scores is no stronger than the evidence for the weakest links in the chain of inferences made from scoring to decision. In other words: Technical item flaws have the capacity to be a threat to the validity of multiple-choice tests, and valid technical flaw detection processes are therefore important for the overall test validity of multiple-choice tests. *Flaw detection is also an assessment process* with its own set of validity assumptions relating to scoring, generalization, extrapolation and decision. With a modern validity framework as a backdrop (American Educational Research Association, 2014; Kane, 2006), we suggest that some important validity assumptions for a review committee's qualitative screening for technical item flaws could be: 1) the recording of reviewers' item assessments represents reviewers' intended item assessments, 2) the list of technical flaws to screen for is appropriate, 3) items are checked accurately and consistently for all technical flaw types on the list, 4) the items reviewed actually represent all the items in the proposed tests, 5) the sample of reviewers is large enough to control for reviewer bias, 6) we may extrapolate item quality from the results of the flaw detection, 7) there are no irrelevant sources of variance which bias the interpretation of the results of the flaw assessment as a measure of item quality, 8) items of lower quality are less likely to be included in the test based on the results of the flaw assessment. Assumptions 1-3 relate to scoring, 4 & 5 to generalization, 6 & 7 to extrapolation and 8 to decision.

The aim of this study was to examine aspects of validity and feasibility of a review committee screening approach for the identification of technical item flaws. The objectives were to examine the generalizability of reviewers' item screening for technical flaws, and to exemplify and discuss the potential impact of these results on item quality and feasibility in a local context. These objectives examine validity assumptions 5) and 6) outlined above.

Methods

Design

The reliability study was designed as a generalizability study rooted in Generalizability Theory (GT) (Brennan, 2001). Multiple-choice items (n=160) were checked independently by raters (n=5), who checked for the presence/absence of flaws based on a list of 19 predefined constructions errors. In other words, the items (i) were the object of measurement and raters (r) were the facet of differentiation. The generalizability design used was 'items crossed with raters', also described as the 'i x r' design in GT (Brennan, 2001). The general formula for the Index of Dependability (the Φ or phi coefficient) for reviewers' item screening process is outlined in equation 1, where σ^2 is the variance.

$$\Phi = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_r^2 + \sigma_{ir,e}^2} \qquad \text{Eq. 1}$$

This coefficient (eq. 1) is not the dependability of any final consensus decisions reached about an item. It describes the extent to which independent reviewers agreed on the categorization of the item as either flawless or flawed. The formula for decision studies for alternative numbers of item reviewers is described in equation 2.

$$\Phi_{Decision} = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_r^2/n_r + \sigma_{ir,e}^2/n_r} \qquad \text{Eq. 2}$$

Participants

Raters were all educationalists with knowledge of the National Board of Medical Examiners (NBME) guidelines on item writing (Case & Swanson, 2002).

Rater 1 had 3 years of experience in teaching item construction for university teachers and a background in the natural sciences. Rater 2 had a medical background and around 13 years of experience with teaching item construction to medical item writers, quality assurance of items and statistical item analyses and evaluations. Rater 3 had a biomedical background and many years of practical experience as an educationalist but was a novice to item screening. Rater 4 was a medical doctor with a PhD in medical education and around 3 years of experience with screening items qualitatively for flaws. Rater 5 had a health science background, a master's degree and a PhD in medical education, and 3 years of experience with teaching item construction to medical item writers, qualitative screening for item flaws, and statistical item analyses and evaluations. None of the raters were experts in all the medical specialist content tested in the exams. We considered this sample of raters a random sample from a universe of educationalists in the national higher education context with a variety of prior practical and theoretical experience with item construction.

Review Criteria

Item reviewers searched for technical item flaws in items based on criteria rooted in the item writing guidelines of the National Board of Medical Examiners (NBME) (Case & Swanson, 2002). Since the reliability of judgments of item relevance has already been examined by others previously (Norcini & Grosso, 1998), we limited the flaw types that reviewers screened for to be the technical flaws described in table 1, which are related to the undue influence of 'testwiseness' and 'irrelevant difficulty'.

Table 1. Types of construction errors (n=19) the reviewers screened for.

Error type	Explanation
Unfocused lead-in	The stem of an item should be focused. It should pose a clear question, and it should be possible to arrive at an answer with the options covered
Grammatical cue	One or more distractors do not follow grammatically from the stem
Logical cue	A subset of options is collectively exhaustive
Absolute term in option	Terms such as 'always' or 'never' are in some options
Long correct answer	The correct answer is longer, more specific, or more complete than other options
Word repeats	A word or a phrase is included in the stem and in the correct answer
Convergence	The correct answer includes the most elements in common with the other options
Long/complicated/double options	Options are long, complicated, or double.
Inconsistent use of numeric data	Numeric data in options, such as intervals, are not stated consistently
Vague term in option	Vague terms such as 'rarely' or 'usually' or 'frequently' etc. are used in the options
Non-parallel language in options	The language in the options is not parallel
Illogical order of options	Options are in nonlogical order
AOTA/NOTA option used	'All of the above' or 'None of the above' is used as an option
A tricky/complicated stem	The stem is unnecessarily tricky, complicated or verbose
Inter-dependent items	The answer to an item is 'hinged' to the answer of a related item
Overlapping options	The answer choices should be independent and non-overlapping
Negations in lead-in	Negatively phrased lead-ins containing words such as 'except' or 'not' etc. should be avoided.
Options not in same domain	The choices must stem from the same content dimension or domain (e.g. all diagnoses, tests, treatments, prognoses, disposition alternatives etc.)
Implausible distractor	Distractors which are blatantly absurd or ridiculous should be avoided

Items reviewed

The items reviewed originated from two multiple-choice exam papers administered in a medical master's degree program at a Danish University. The items were in the One-Best-Answer format (Case & Swanson, 2002), with three answer options (A, B, or C) per item. There is quite solid research-based evidence showing that test items seldom contain more than three useful options anyway (Rodriguez, 2005), and that local test developers may be better off using 3

options as opposed to 4 or 5 options (Haladyna & Downing, 1993), as long as the test contains more than 35-40 items (Downing & Yudkowsky, 2009). The two tests from which the items for this study were drawn contained 80 items each. The overall purpose of the two exams was to test clinical knowledge across a number of medical specialties. A variety of regional medical doctors representing different fields of medical expertise (medical specialties) constructed test items for these exams. Item writers had undergone courses in item construction based on the item writing guidelines of the National Board of Medical Examiners (NBME) in the US (Case & Swanson, 2002). The guidelines and instructions delivered to item writers included thorough explanations and exemplification of all of the technical flaws listed in table 1.

Data collection

Two randomly selected exam papers were drawn from the pool of all existing multiple-choice exam papers developed in the context (n_{items}=160). Each exam paper was independently reviewed by the raters who searched for 19 specific types of item flaws (table 1). Each rater independently filled out a prepared evaluation sheet, indicating which type or types of errors were detected in each item. We subsequently dichotomized the data for the purpose of dependability analyses, so that raters' interpretation of the presence of one or more errors in an item was coded 1, while their interpretation of an absence of errors for an item was coded 0.

Analysis

GENOVA for PC was used to estimate variance components and calculate dependability/phi coefficients for test situations with alternative numbers of reviewers as outlined in equation 2 (Brennan, 2003).

In order to *exemplify and illustrate the potential impact* of the dependability results on item quality and feasibility in the context, we compared three different screening approaches and the effects they would have had on selected test parameters and use of time. The three screening approaches compared were: No screening at all, a qualitative screen, and a quantitative screen. The qualitative screen example was based on the results of our 5 reviewers' evaluations and the inclusion criteria for a test item was that none of the reviewers had found any flaws in the item. In the quantitative screen example, only items with item-rest correlations of 0.15 or above were included in a test, as this has been recommended as minimum levels of acceptable item discrimination by experts (Haladyna, 2012). These three examples of approaches to quality assurance were compared on the following parameters: The number of items that ended up in the test (n items included), the mean item difficulty (DIF) of included items, the mean item discrimination (DI) of included items, the proportion of non-functioning distractors of included items, and the total number of screening hours spent.

The item difficulty index (DIF) was the percentage of examinees answering an item correctly (Case & Swanson, 2002). We used the item-rest correlations as the item discrimination index (DI). Non-functioning distractors were defined as the incorrect answer options (distractors) which less than 5% of examinees had chosen (Haladyna & Downing, 1993; Malau-Aduli & Zimitat, 2012). The number of non-functioning distractors were counted and converted to a proportion of the total number of distractor options in the included test items. Differences in DIF, DI, and non-functioning distractors between the three screening examples were analyzed either with t-tests or Fisher's Exact Tests. In addition, flawless items (items judged flawless by all 5 raters simultaneously) were compared to flawed items with respect to DIF, DI and non-

functioning distractors. All item analyses were performed with IC STATA 15. The alpha command in STATA was used to obtain DI values for items.

The calculations of DIF, DI and non-functioning distractors were based on real students' (n=128 and n=166) exam performances in the two authentic exams in which the items had been administered. Individual student cases in the exam data were identified by unique numbers, and researchers in this study did not have access to any keys which could break students' anonymity. This project was exempt from ethics review by the regional ethics committee as surveys, database studies, and quality assurance studies do not require their permission. Permission from the Danish Data Protection Agency was not required either, because the data is not considered sensitive data. Data was stored according to current laws on data protection.

Results

Item reviewers found that 19-50% (or 31-80) of the 160 items contained at least one flaw, with an average item flaw rate across all five raters of 39% (or 62/160 items). Only 21% (or 34/160) of the items reviewed were categorized as flawless by all five raters simultaneously. In the remaining items (n=126), reviewers found between 1-7 different types of flaws with a median of 2 flaws per item. The most common flaw detected by reviewers was an unfocused lead-in (table 2).

Table 2. Types of construction errors detected in the items (n=160) by reviewers.

Error type	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	SUM
Unfocused lead-in	5	34	15	25	12	91
Grammatical cue	1	0	5	2	0	8
Logical cue	8	5	38	0	1	52
Absolute term in option	1	0	3	0	0	4
Long correct answer	5	12	9	4	3	33
Word repeats	8	1	8	6	0	23
Convergence	13	13	6	4	2	38
Long/complicated/ double options	7	2	1	3	4	17
Inconsistent use of numeric data	1	0	1	1	0	3
Vague term in option	1	3	2	2	0	8
Non-parallel language in options	3	0	1	23	0	27
Illogical order of options	3	1	2	0	0	6
AOTA/NOTA option used	2	0	0	0	0	2
A tricky/complicated stem	5	2	1	2	1	11
Inter-dependent items	0	0	2	0	0	2
Overlapping options	2	1	4	9	1	17
Negations in lead-in	3	1	1	5	2	12
Options not in same domain	8	4	11	1	6	30
Implausible distractor	11	5	7	5	3	31

The estimated variance component values (σ^2) were: 0.055 (SE=0.010) for items, 0.012 (SE=0.008) for raters, and 0.173 (SE=0.010) for items by raters.

Using five reviewers for item screening yielded a dependability coefficient of 0.60 (table 3).

Table 3. Dependability coefficients (Φ) for alternative numbers of item reviewers.

N reviewers used	Φ
1	0.23
2	0.37
3	0.47
4	0.54
5	0.60
14	0.81
31	0.90

For very high levels of dependability (Φ>0.90) of a qualitative item screening process, more than 31 reviewers should be used according to these estimates (table 3).

The mean item discrimination (DI) of the 34 items considered flawless by all five reviewers simultaneously was 0.18 compared to 0.14 for the 126 items categorized as flawed by one or more reviewers, and this difference was statistically significant (p=0.017). In contrast, there was no statistically significant difference in item difficulty or in the proportion of non-functioning distractors between items judged flawed and flawless respectively by the five raters.

The qualitative screening approach example could potentially have resulted in discussions of a very large proportion (79%) of the test items reviewed if consensus deliberations among the five reviewers been pursued (table 4).

Table 4. Different screening approach examples and their effects on selected test parameters.

Screening method	Test inclusion criteria	N Items Included (%)	Mean DI (SD)	Mean DIF (SD)	% non-functioning distractors (n/N)	Total screening hours spent
No screening	None	160 (100)	0.15[a] (0.10)	80[a] (18)	53%[a] (168/320)	0
Qualitative review (n_r=5)	Flawless items	34 (21)	0.18[a] (0.11)	84[a] (20)	57%[a] (39/68)	30
Quantitative review (n_r=1)	Items with DI ≥0.15	75 (47)	0.24[b] (0.06)	0.82[a] (16)	48%[a] (72/150)	2

DI= discrimination index, DIF= item difficulty, n_r=number of reviewers used. Values with different superscripts in the same column are significantly different (p <0.05).

Discussion

Qualitative screening of multiple-choice items for flaws required many reviewers for higher levels of dependability in a local test context. Post-exam statistical screening appeared to be a more feasible way of improving the ability of the tests to discriminate between examinees.

Flaw rates and types

Previous studies have reported that as many as 20-75% of test items produced displayed at least one flaw (Downing, 2002, 2005; Haladyna & Downing, 1993; Hansen & Dexter, 1997; Jozefowicz et al., 2002; Masters et al., 2001; Palmer & Devitt, 2007; Rodríguez-Díez et al., 2016; Tarrant et al., 2006; Tarrant & Ware, 2008), which seems to suggest that the average flaw rates in our sample of 39% were somewhere in the middle range in comparison. We found 'unfocused lead-ins' to be the most common flaw type in our setting (table 2). Other studies have found similar results, but it is generally difficult to compare the precise composition of flaw types between different studies directly because flaw lists are not completely congruent between studies (Downing 2005; Tarrant and Ware 2008).

Dependability of qualitative technical flaw detection

We found estimated variance components and corresponding standard errors of 0.055 (SE=0.010) for items, 0.012 (SE=0.008) for raters, and 0.173 (SE=0.010) for items by raters. In comparison, when Norcini and Grosso examined the generalizability of 37 reviewers' judgments of item relevance, they found estimated variance components and corresponding standard errors of 0.135 (SE=0.063) for items, 0.122 (SE=0.073) for raters, and 0.641 (SE=0.074)

for items by raters (Norcini & Grosso, 1998). They concluded that their standard errors were small relative to the size of the variance components and indicative of reasonably good estimation, so a similar judgment of our results seems fair.

The resultant moderate dependability coefficient (Φ=0.60) for an item review using five reviewers confirmed our initial assumption that flaw detection (like item writing) is challenging. It also means that the evidence for validity assumption 5 described above was weak to modest. For very high levels of dependability in technical flaw detection, we would have needed the independent judgements of at least 31 reviewers with comparable backgrounds instead of just 5 (table 3), which would most likely not be a feasible solution in the context. Norcini and Grosso found comparable results when they examined the dependability of judging item relevance with a design similar to ours (Norcini & Grosso, 1998). Thirty-seven medical practitioners independently rated one-best-answer items for relevance for professional practice in general internal medicine on a five-point scale. Results of their variance components analysis indicate that for ratings of item relevance with high (Φ>0.90) levels of dependability, 51 raters should assess the items (Norcini & Grosso, 1998). In contrast to our results and the results of Norcini and Grosso (1998), we found three studies reporting higher levels of reliability in flaw detection. One study examined whether 39 experienced reviewers on an item review committee could accurately identify test items constructed or selected to exhibit 16 different cultural or technical flaws accurately. The results indicated that the mean accuracy rates amongst reviewers were high following directly on from a 60-minute tailored training session (Engelhard Jr et al., 1999). However, the authors also mentioned that 'the specific training provided' in the tailored training session immediately before the review and the 'obviousness of flaws' could have accounted for the substantive results reported. Two studies reported reliability coefficients of 0.89 when using three expert NBME item reviewers. The expert reviewers independently categorized items into just five broader quality categories relating to: the type of knowledge tested in the item (applied or factual knowledge), the item format (one-best-answer versus true-false), and the presence/absence of technical flaws (Jozefowicz et al., 2002; Wallach, Crespo, Holtzman, Galbraith, & Swanson, 2006). The superiority of the expert reviewers used is a plausible explanation for the high levels of reliability presented in these studies. On the other hand, keeping the quality criteria fewer and broader could also have secured higher levels of reliability. Recognizing whether an item contains a vignette or not and whether it is of the one-best-answer format or not would appear to be easier to spot than technical flaws, because the latter comes in so many shapes and sizes (table 1). It is therefore also possible that our flaw focus (technical flaws only) may have contributed to a relatively larger influence of cognitive overload and subsequent errors in the flaw assessment. Our results cannot be interpreted to mean that awareness of technical item flaws is not warranted, as existing evidence indicates that training item writers may improve the quality of test items developed in local contexts for in-house exams (Jozefowicz et al., 2002; Naeem, van der Vleuten, & Alfaris, 2012; Wallach et al., 2006). Likewise, one study reported significant improvements in the quality of multiple-choice exams after the introduction of peer review workshops for item writers in which the focus was on three issues relating to the relevance of items (Malau-Aduli and Zimitat 2012). While longer lists of technical flaws to avoid may be very useful in the general training of item writers, and perhaps as check lists for individual item writers and their peer reviewers reviewing only a limited number of items, our results seem to question their use in systematic judgements of whole tests (many items) by a local review committee.

Apart from the reflections our results may trigger in local educational test contexts about the efficiency of quality assurance processes, the dependability of human judgments of item flaws is also important in the global research on the effects of item flaws on student performances. In research studies where only a few item reviewers have been used and no reliability/generalizability coefficient of the flaw assessment in question has been reported, a concern about the correct categorization of items as flawed/not-flawed is warranted. As we have shown in this study, this categorization is not necessarily straightforward even amongst experienced reviewers.

Impact on validity

The potential impact of the dependability results reported above on test validity aspects and feasibility in a local context is exemplified in table 4.

In the three examples of approaches illustrated here, there were no significant differences in mean item difficulty (DIF) between the three screening approach examples compared (table 4). However, others have reported that flawed items tended to be more difficult for test takers and to increase fail-rates (Downing, 2002, 2005). In one such study, items with and without flaws intended to test basic science were compared. Results showed that flawed items were 7% or nearly ½ SD more difficult for students than items without flaws, and the flawed items failed nearly one fourth more students than the flawless items (Downing, 2002). Similar results were also found later in a larger study of four basic science examinations administered to year-one and year-two medical students. Across examinations, only 47% of students passed the flawed items while 53% students passed the flawless items (Downing, 2005). Studies in other settings have since corroborated these results, finding flawed items to be 7.4-12.3% more difficult for students than flawless items (Caldwell & Pate, 2013; Pate & Caldwell, 2014). In addition, one study also found high-achieving students were more likely than borderline students to be penalized by flawed items (Tarrant & Ware, 2008). The results of these studies are evidence of flawed items' potentially negative effects on test validity, and they appear to be in line with the existence of 'irrelevant difficulty' as a competing test construct.

The evidence for the assumption that we could extrapolate item quality from our reviewers' flaw detection (validity assumption 6) was not particularly convincing. Although we found the mean item discrimination was significantly higher (0.18) in the items judged flawless by all reviewers simultaneously compared to the rest of the items (0.14), there was no significant difference in average item discrimination between the 'no screening' approach and the 'qualitative review' approach (table 4). The lack of dependability of the 'qualitative review' example presented here was associated with too much noise and too little useful signal. In contrast, using a 'quantitative review' approach post-exam (based on minimum recommendations of item discrimination) would have been a more efficient way of improving the average item discrimination in the test context (table 4). This ability of a test to discriminate between high and low-ability test takers has been described as 'a fundamental principle of all educational measurement and a basic validity principle' (Downing & Yudkowsky, 2009), and curbing construct-irrelevant variance arising from poorly crafted items is considered one of the important ways of improving item discrimination and overall test validity (Downing, 2002).

We found no statistically significant difference in the proportion of non-functioning distractors in items judged flawed versus flawless by the five reviewers, and no statistically significant differences in the proportions of non-functioning options in the three screening approach examples illustrated in table 4. In contrast, others have reported a significant decrease in the

number of non-functioning item options after the introduction of peer review workshops for item writers where the focus was on item relevance (Malau-Aduli & Zimitat, 2012). Item relevance was not evaluated by our reviewers (table 1).

Impact on feasibility

Multiple-choice tests are often hailed for their efficiency in the scoring process compared to other written test formats, but resources must instead be spent on the training of item writers and on writing items in sufficient numbers. Reviewing items qualitatively will further detract from the feasibility of this test format in local contexts. The five raters in this study spent approximately 6 hours each on a review of the 160 items, and others have reported similar rates for other experienced reviewers (Engelhard Jr et al., 1999). Based on the dependability results described above, we would have needed to spend at least 186 hours screening qualitatively and have had access to 31 experienced reviewers for high levels of dependability, or perhaps even more if we had also included judgments of item relevance, as indicated by the results of a previous study (Norcini & Grosso, 1998). To maintain a core of that many reviewers on an exam committee at course level is most likely an unacceptable challenge in many local educational contexts. In addition, subsequent consensus work about the final fate of each of the many items (n=126 in our case) perceived to be flawed by at least one reviewer would further detract from the feasibility. The question is whether it would be practically possible in very large exam committees. As this example illustrates, insufficient dependability of flaw detection has the power to undermine both exam validity as well as the feasibility of the quality assurance process.

In contrast, using the 'quantitative review' screening approach instead, which is based on students' actual exam performance as opposed to reviewers' judgments, would in our experience probably require around 2-3 hours of work for one person (table 4).

Limitations

The global evidence on the reliability of qualitative human judgments of item flaws appears to be quite sparse. In this study, we concentrated solely on the detection of a range of technical flaws that may favor testwise students or infuse the test with construct-irrelevant difficulty, and thereby threaten test validity. Other fundamental issues, such as whether the rules of the item format are respected, correct keying, and whether the amassed content and taxonomic level of test items is congruent with the course learning goals, the teaching and learning activities in the course as well as with subsequent professional practice etc. are of course also extremely important for the overall test validity.

Future research

Influential assessment literature often originates from large-scale professional test agencies in the English-speaking world, but the resulting *guidelines may represent infeasible and non-transferrable paragons of perfection in smaller-scale educational settings*. We believe there is a need to critically and openly examine if and how multiple-choice tests can be valid and feasible in local/small scale test situations around the globe. It would be very helpful if guidelines on how to optimize quality assurance processes in smaller-scale/lower resource contexts were available. Finally, systematic reviews of the effects on test validity of item writer training and of violating proposed item writing principles - in local educational contexts across the globe - also seem to be missing.

Conclusion

Collectively, we found only weaker evidence for validity assumptions relating to generalization and extrapolation in this study of qualitative item screening by a review committee. Furthermore, review committee quality assessments seemed to have the capacity to detract from the feasibility of multiple-choice exams.

Acknowledgements

The authors would like to thank the item reviewers who participated in this study.

Declaration of Interest

The authors declare no conflict of interest.

References

American Educational Research Association, A. P. A., National Council of Measurement in Education. (2014). *Standards for educational and psychological testing.* Whashington (DC): American Educational Research Association.

Biggs, J., & Tang, C. (2007). Teaching for Quality Learning at University 3rd edition Open university Press: Maidenhead.

Brennan, R. L. (2001). In *Generalizability Theory*. New York: Springer.

Brennan, R. L. (2003). GENOVA for PC. Retrieved from https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs#GENOVA

Caldwell, D. J., & Pate, A. N. (2013). Effects of question formats on student and item performance. *American Journal of Pharmaceutical Education*, 77(4), 71.

Case, S., & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences* (3rd ed.). Philadelphia (PA): National Board of Medical Examiners.

Dent, J., Harden, R. M., & Hunt, D. (2017). *A practical guide for medical teachers*: Elsevier Health Sciences.

Downing, S. M. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education*, 7(3), 235-241.

Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical Education*, 37(9), 830-837.

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education*, 38(9), 1006-1012.

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

Downing, S. M. (2006). Twelve steps for effective test development. *Handbook of test development*, 3-25.

Downing, S. M., & Yudkowsky, R. (2009). *Assessment in health professions education*: Routledge.

Ebel, R. L. (1951). Writing the test item. *Educational measurement*, 185-249.

Engelhard Jr, G., Davis, M., & Hansche, L. (1999). Evaluating the accuracy of judgments obtained from item review committees. *Applied Measurement in Education*, 12(2), 199-210.

Haladyna, T. M. (2012). *Developing and validating multiple-choice test items*: Routledge.

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

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

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business*, 73(2), 94-97.

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house médical school examinations. *Academic Medicine*, 77(2), 156-161.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17-64): ACE/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73.

Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education*, 37(8), 919-931. doi:10.1080/02602938.2011.586991

Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *Journal of Nursing Education*, 40(1), 25-32.

Naeem, N., van der Vleuten, C., & Alfaris, E. A. (2012). Faculty development on item writing substantially improves item quality. *Advances in Health Sciences Education*, 17(3), 369-376.

Norcini, J., & Grosso, L. (1998). The generalizability of ratings of item relevance. *Applied Measurement in Education*, 11(4), 301-309.

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. *BMC Medical Education*, 7(1), 49.

Paniagua, M. A., & Swygert, K. A. (2016). *Constructing Written Test Questions For the Basic and Clinical Sciences* (4th ed.). Philadelphia, PA: National Board of Medical Examiners.

Pate, A., & Caldwell, D. J. (2014). Effects of multiple-choice item-writing guideline utilization on item and student performance. *Currents in Pharmacy Teaching and Learning*, 6(1), 130-134.

Rodríguez-Díez, M. C., Alegre, M., Díez, N., Arbea, L., & Ferrer, M. (2016). Technical flaws in multiple-choice questions in the access exam to medical specialties ("examen MIR") in Spain (2009–2013). *BMC Medical Education*, 16(1), 47.

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

Swanson, D. B., Norcini, J. J., & Grosso, L. J. (1987). Assessment of clinical competence: written and computer-based simulations. *Assessment & Evaluation in Higher Education*, 12(3), 220-246. doi:10.1080/0260293870120307

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

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.

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

Wallach, P. M., Crespo, L. M., Holtzman, K. Z., Galbraith, R. M., & Swanson, D. B. (2006). Use of a Committee Review Process to Improve the Quality of Course Examinations. *Advances in Health Sciences Education*, 11(1), 61-68. doi:10.1007/s10459-004-7515-8