User experience of the Written Exam Question Quality tool to inform the writing of new written-exam questions

Expérience des utilisateurs de l'outil de qualité des questions d'examen écrit pour la rédaction de nouvelles questions d'examen écrit

Élise Vachon Lachiver,[1] Christina St-Onge[1]

[1]Faculty of Medicine and Health Sciences, Université de Sherbrooke, Quebec, Canada

Correspondence to: Christina St-Onge, Department of Medicine, Centre de pédagogie des sciences de la santé, Paul Grand'Maison de la Société des Médecins de l'Université de Sherbrooke Research Chair in Medical Education, Faculty of Medicine and Health Sciences, Université de Sherbrooke 3001 12e Avenue Nord Sherbrooke, QC J1H 5N4; phone: 819-821-8000, ext. 75047; fax: 819-820-6815; email: christina.st-onge@usherbrooke.ca

Abstract

Background: Creating new written-exam questions is a burdensome task for faculty members. While several guidelines exist, there had not been a previous attempt to streamline them in a user-friendly tool. We created the Written Exam Question Quality tool (WEQQ) and explored potential users' perception of this tool when writing their exam questions.

Methods: We conducted a descriptive study to explore how four Canadian faculty members used the WEQQ. We conducted structured interviews that were analyzed within and across participants to understand the latter's perceived usefulness and acceptability of the WEQQ. Quantitative data from a short questionnaire on creating exam questions and their psychometric properties were also collected.

Results and conclusion: Participants' perception of the WEQQ was positive, and they were favorable to its use. The WEQQ seemed to represent a user-friendly, easy way to help faculty members in creating multiple-choice or short-answer questions. Time on task remained the same when using the WEQQ. We were able to identify two user profiles, passive and active, which indicated how faculty members use the WEQQ to create exam questions. Future steps would be to further investigate if the WEQQ can increase the quality of written-exam questions and to understand how to promote an active use of the WEQQ when implementing this tool.

Résumé

Contexte : La création de nouvelles questions d'examen écrit est une tâche lourde pour les membres du corps enseignant. Bien qu'il existe plusieurs lignes directrices, aucune tentative antérieure n'a été faite pour les rationaliser dans un outil convivial. Nous avons créé l'outil de qualité des questions d'examen écrit (WEQQ - *Written Exam Question Quality tool*) et exploré la perception de cet outil par les utilisateurs potentiels lors de la rédaction de leurs questions d'examen.

Méthodes : Nous avons mené une étude descriptive pour explorer comment quatre membres de facultés canadiennes ont utilisé le WEQQ. Nous avons réalisé des entretiens structurés qui ont été analysés au sein des participants et entre eux afin de comprendre l'utilité et l'acceptabilité perçues du WEQQ. Des données quantitatives provenant d'un court questionnaire sur la création de questions d'examen et leurs propriétés psychométriques ont également été recueillies.

Résultats et conclusions : Les participants ont eu une perception positive du WEQQ et étaient favorables à son utilisation. Le WEQQ semble représenter un moyen convivial et simple d'aider les enseignants à créer des questions à choix multiples ou à réponses courtes. Le temps consacré à la tâche est resté inchangé lors de l'utilisation du WEQQ. Nous avons pu identifier deux profils d'utilisateurs, passif et actif, qui indiquent comment les enseignants utilisent le WEQQ pour rédiger leurs questions d'examen. Les prochaines étapes consisteront à étudier davantage si le WEQQ peut améliorer la qualité des questions d'examen écrit et à comprendre comment promouvoir une utilisation active du WEQQ lors de sa mise en œuvre.

Introduction

While more than 150 item-writing guidelines for creating written-exam questions exist,[1–3] until a recent work[3] there was little empirical evidence supporting the impact of these guidelines on the psychometric quality of the questions. The Written Exam Question Quality tool (WEQQ)[3] [Appendix A], a checklist of 14 quality indicators for written-exam questions, can discriminate between high- and low-quality written-exam questions. While we have used the WEQQ in research settings to document the quality of written-exam questions,[3] we have yet to document its utility and acceptability as a tool to inform the development of new questions. The purpose of this study was to document faculty members' perception of the WEQQ in terms of utility and acceptability, and look into the quality of newly created exam questions. We were interested in the unique perspective of participants' expertise or experience as a question writer when they created new questions with the WEQQ. We wanted to highlight similarities and differences between participants.

Methods

We conducted a descriptive qualitative study grounded in the post-positivist paradigm.[4–6] The study was approved by our institution's Research Ethics Committee - Education and Social Sciences in July 2015.

Context

We conducted our study in the context of a Canadian four-year undergraduate medical education (UGME) program. The preclinical component of this UGME program is divided into different modules that correspond to the different systems of the human body. At the end of each module, learners take a written exam comprised of Multiple Choice Questions (MCQs), Short Answer Questions (SAQs) and sometimes long-answer questions. Each year, 25-30% the exam questions must developed by the faculty members responsible for the evaluation while the rest of the exam questions come from a question bank.

Participants and recruitment

We recruited four faculty members via an email invitation sent to all the faculty members in the UGME program (*n* = 13) who created exam questions for the 2013-2014 and 2014-2015 academic years and who had to create new questions for the 2015-2016 academic years. Participants' experience with teaching and question writing are presented in the Results section.

Material and procedure

To improve clarity and flow, we present the data collection and analysis sections together so that the same subject is comprehensively covered as one unit.

Writing experience. The questionnaire included six opened-ended questions related to the participants' experience, past writing strategies, knowledge of other question-writing guidelines, and perceived potential for improvement. One of the researchers (EVL) met with the participants to present them the WEQQ and ask them to use it when writing their questions. During this meeting, participants were invited to answer the questionnaire on writing experience. The researcher (EVL) instructed participants to use the WEQQ when creating their new exam questions for the 2015-2016 year. In addition, participants had to adhere to the UGME program requirements with respect to number of new questions to include, general degree of difficulty, and types of exam questions. Descriptive analyses were computed in Excel for the questionnaire on writing experience and used to summarize and subsequently compare participants' experience (writer profile) when creating exam questions.

Perceived acceptability and utility. The interview guide contained nine questions as a starting point to document how using the WEQQ could influence participants' experience of creating new exam questions. These questions addressed perceived usefulness and acceptability of the WEQQ, as well as their inclination to promote its use in their professional environment. One member of the team conducted a structured, 30-minute individual interview with each participant two weeks after they created their new exam questions. The individual interviews were recorded for transcription and analysis purposes. The structured interviews were coded in *Dedoose* software[7] using a coding tree developed inductively by the co-authors. One author (EVL) analyzed the data using the different codes. Periodic and extensive discussions with the second author (CSTO) ensured the internal coherence of the coding. Analyses were first carried out per participant to identify the elements related to certain themes: the use, acceptability, and perception of the WEQQ. We compared results from each participant A cross-sectional analysis was carried out to highlight the contrasts and characteristics present in the four participants for the previously mentioned themes. We were interested in the unique perspective of participants' expertise or experience as a question writer when they created new questions with the WEQQ. We wanted to

highlight similarities and differences between participants. Excerpts presented in the results have been translated from French.

Exam questions and their quality. Exam questions, and their discrimination coefficients, for the 2013-2014, 2014-2015 and 2015-2016 exams were provided to the researchers by the UGME program. The WEQQ (Appendix A) was used to assess compliance with the guidelines. Participants' exam questions were evaluated blindly by EVL (without knowledge of year or participants) using the WEQQ. A compliance score (percentage) was calculated for each question per participant. For example, if the MCQ question respected seven of the eight guidelines, the compliance was 87%. Also, for each participant, we calculated a central tendency of their annual mean compliance score for the three years of data.

Exam questions and their psychometric properties. The discrimination index was calculated using the corrected point-biserial correlation coefficient ($_{pb}$).[8,9] Discrimination indices are interpreted similarly to a correlation coefficient, a value closer to one representing a strong association between the variables (score on an item and overall score on the test), and positivity and negativity indicating the direction of that relation. In this context, a high discrimination coefficient suggests that a WEQQ indicator can be used to identify quality written exam questions.

Descriptive analyses were done for the item discrimination coefficients. Simple ANOVAs were done, per participants, to test for mean differences in discrimination coefficients per year. The significance level was set at p = 0.05. Quantitative analyses were done using SPSS version 24.0.[10]

Results

Sociodemographic data and results obtained from the questionnaire on experience are presented in Table 1.

Compliance with guidelines and quality of questions

The percentage of guideline compliance did not change overtime (see Table 2). We observed slight variations in mean discrimination over time, with small increases for P2 and P4, where tendency for the discrimination coefficients to improve was greater between the second and third year of creating new exam questions (which is the year the guidelines were used). For P1 and P3, the discrimination coefficients decreased slightly over time (see Table 2). However, there were no statistical differences for question discrimination over time (P1: $F_{(2, 73)}$ = 0.677, p = .511; P2: $F_{(2, 126)}$ = 2.238, p = .111; P3: $F_{(2, 115)}$ = 0.421, p = .657; P4: $F_{(2, 66)}$ = 1.412, p = .251)

Table 1. Writing experience and socio-demographics data of our participants

	Participant 1	Participant 2	Participant 3	Participant 4
Gender	Man	Man	Man	Woman
Years of experience	6	4	9	5
Resources used	None	UGME guide	None	UGME guide
Training	None	Item writing workshop	None	UGME training for writers
Perception of the creation of new exam questions	Complex task that allows knowledge refresh	A stimulating challenge	Very difficult	Stimulating when it's in my area of expertise. Sometimes more difficult when it further connects from my daily practice.
Perception of the WEQQ	No surprises, guidelines are normal and correct	Concise and clear. A simple and easy to use guide	Could have a negative effect of using guidelines too systematically	Very clear, reduces the risk of error
Perceived impact of WEQQ on writing time	Decreased time to create new exam questions	Decreased time to create new exam questions	Decreased time to create new exam questions	Increased time to create new exam questions (done more conscientiously)
Encline to share the WEQQ with colleagues	Yes	Yes	Yes, but with some caveats	Yes

Table 2. Description of quantitative data according to writing year for each participant.

Unit	Years new questions were written	Number of new questions	Mean compliance with guidelines (SD)	Mean discrimination	P value for discrimination
P1	2013-2014	26	87%	0,225	
	2014-2015	25	86%	0,196	0.511
	2015-2016*	27	85%	0,193	
P2	2013-2014	44	88%	0,134	
	2014-2015	38	87%	0,157	0.111
	2015-2016*	47	84%	0,187	
P3	2013-2014	22	89%	0,162	
	2014-2015	59	88%	0,147	0.657
	2015-2016*	37	87%	0,136	
P4	2013-2014	21	87%	0,092	
	2014-2015	26	88%	0,093	0.251
	2015-2016*	26	87%	0,125	

* indicates the year the WEQQ was used to create new exam questions

Perception of the WEQQ

Participants seemed to appreciate the WEQQ format (a simple checklist), saying that "…it gives something that is a little structuring" (P4). They perceived the guidelines as being simple, concise, and clear, making the WEQQ useful. All participants mentioned that they will surely use the tool in the next cycle of creating new exam questions. Three participants (P1, P2, and P3) stated that their writing time was not increased compared to the previous years, stating 'it did not decrease the time on task, but it also did not increase it'(P2), which they all appreciated. For P4, the added time when using the WEQQ is due to that task being done much more conscientiously. After a few uses, this participant thinks it will take less time to create new exam questions.

P3 indicated being hesitant to use the guidelines too systematically. According to this participant, the WEQQ guidelines are more suggestions; they did not "consider them as strict rules to follow…"(P3). This participant was also somewhat reticent to share the WEQQ with colleagues, specifying that we must always use our judgment when applying guidelines. Similarly, two participants (P3 and P4) raised concerns about using the WEQQ. Given their teaching and evaluation contexts, certain guidelines could not always be respected. For example, in the WEQQ, we encourage writing questions with only one correct answer, while the program allows "questions that ask to list several factors" (P3). In this regard, P3 mentioned that it was therefore very important to always exercise judgment when writing and not to use systematically the guidelines. They suggested this could be added as an introduction to the WEQQ.

Identification and description of writer profiles

When comparing participant characteristics (Table 1), we were able to identify two potential profiles among our writers: active (P2 and P4) or passive (P1 and P3) users based on common characteristics (within a profile) and differing characteristics (between profiles). The active users differed from the passive users in that they were more junior writers, already actively using guidelines (provided by the UGME program), and took part in item-writing workshops. P2 and P4 used the guidelines more actively because *"it [the guidelines] could help to improve learning assessment, because, with the way the tool is made, it allows you to be very specific in what you want to assess"* (P2). The passive users had been writing questions for longer, stating they were already implicitly using the guidelines and that they saw them "as a reference document" (P1 and P3), In addition, P1 indicated that "*if [he] did not have this grid, [he] would not have used a grid, [he] would have just used common sense.*"

Discussion

The purpose of our study was to explore and document user perspective about the use of evidence-informed guidelines for writing exam questions. We observed little improvement in quality of the items when participants used the WEQQ, however, we could identify two types of users. Passive users were less enthusiastic about using the guidelines and made less effort to adhere to them. Active users tended to be more motivated to use the guidelines. While using the guidelines this way may have increased the time to create their questions, active users appreciated the learning opportunity. They even considered that it reduces the time needed to create new exam questions. As such, these preliminary findings support the idea of engaging and convincing potential users of the relevance of a new tool.

The tool seemed more useful for those participants who were motivated and who appreciated just-in-time support.

The use of these evidence-informed guidelines seemed to offer a just-in-time faculty development opportunity. Participants accessed the guidelines at the time they were creating new written-exam questions as opposed to attending a workshop and forgetting some, if not all, of the content when it came time to using it. This could be a significant facilitator[11] at the individual level, but also has organizational downstream consequences for the programs, reducing the need to organize formal faculty development sessions.

One of the limitations of this study is that the participants already had sound strategies (reflected in the compliance rate) and had some interest in writing questions, meaning they were already developing good-quality questions. This sampling bias could have explained some of the results, such as the ceiling effect. The context of one medicine program and its relatively high-stakes assessments can likely explain these results. The timing of the interview (two weeks after) could equally pose a concern with recall related to the use of the guidelines. Recommending a "think-aloud" method with participants could be helpful to gain an understanding of how they apply the guidelines from the WEQQ "in the moment." Other contexts could allow us to test out the entire tool, since some guidelines were not aligned with the recommendations from our UGME program. Also, the students' point of view was omitted from our study, even though it could have been interesting to ask for it, especially from an implementation science perspective. We acknowledge that being the author of the proposed tool, being the one to present it to the participants, and being the one to interview them may have caused an acquiescence bias that might also have influenced the data. We have tried to mitigate this bias by asking the participants to speak freely and looking at other data sources such as the quality of questions.

Conclusion

We found that the perception of the usefulness of the WEQQ varied according to user profile. The tool seems more beneficial for active users than passive users. Participants perceived the tool as a way to reduce writing time and organize the task of writing new questions. The quality of questions, however, did not significantly improve for experienced question writers. Future studies should examine the effect -on the quality of questions- of implementing the WEQQ in other programs—and for more novice writers.
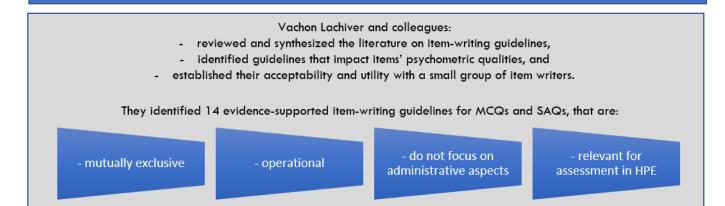
References

1. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ*. 2002;15(3):309–33. https://doi.org/10.1207/S15324818AME1503_5

2. Paniagua M, Swygert KA, Downing SM. Written tests: writing high-quality constructed-response and selected-response items. *Assess Health Prof Educ*. 2019;109–26.

3. Vachon Lachiver É, St-Onge C, Cloutier J, Farand P. La rédaction de questions à choix multiple et de questions à réponse ouverte et courte pour les examens écrits dans les programmes de formation en santé : une étude docimologique des consignes de rédaction. *Pédagogie Médicale*. 2017;18(2):55–64. https://doi.org/10.1051/pmed/2018002

4. Denzin NK, Lincoln YS, Denzin NK, Lincoln YS. Handbook of qualitative research. *Handb Qual Res*. 1994; Available from: https://search.ebscohost.com/login.aspx?direct=true&db=snh&AN=COMP0998985617&site=ehost-live

5. Denzin NK, Lincoln YS (Eds). The Sage handbook of qualitative research. Thousand Oaks, CA: SAGE; 2011.

6. Young ME, Ryan A. Postpositivism in health professions education scholarship. *Acad Med J Assoc Am Med Coll*. 2020;95(5):695–9. https://doi.org/10.1097/ACM.0000000000003089

7. Dedoose Version D Version 7.0.2, web application for managing, analyzing, and presenting qualitative and mixed method research data (2016). Los Angeles, CA: SocioCultural Research Consultants, LLC. www.dedoose.com. 2016

8. Carroll JB. Correcting point-biserial and biserial correlation coefficients for chance success. *Educ Psychol Meas*. 1987;47(2):359–60. https://doi.org/10.1177/0013164487472007

9. DeMars CE. *Classical test theory and item response theory*. Wiley Handb Psychom Test Multidiscip Ref Surv Scale Test Dev. 2018;49–73.

10. IBM Corp. Released 2016. *IBM SPSS Statistics for Windows, Version 24.0.* Armonk, NY: IBM Corp.; 2016.

11. Thomas A, Bussières A. Knowledge Translation and Implementation Science in Health Professions Education: Time for clarity? *Acad Med J Assoc Am Med Coll*. 2016;91(12):e20. https://doi.org/10.1097/ACM.0000000000001396

Appendix A. Written Exam Question Quality (WEQQ):14 evidence-informed guidelines to create MCQs and SAQs

Written Exam Question Quality (WRQQ) evidence-informed guidelines to create MCQs and SAQs

Vachon Lachiver and colleagues:
- reviewed and synthesized the literature on item-writing guidelines,
- identified guidelines that impact items' psychometric qualities, and
- established their acceptability and utility with a small group of item writers.

They identified 14 evidence-supported item-writing guidelines for MCQs and SAQs, that are:

- mutually exclusive

- operational

- do not focus on administrative aspects

- relevant for assessment in HPE

8 MCQ specific Question-writing guidelines

1. Stems must be unambiguous and clearly state the problem.
2. Avoid distractors that can clue test-wise examinees; for example avoid absurd options, formal prompts, or semantic clues (overly specific or overly general).
3. Avoid answer options: "All of the Above" or "None of the Above".
4. Answer options should be logically independent of one another (options not be overlapping).
5. Answer options should all be grammatically consistent with stem.
6. Answer options should include only one correct answer.
7. Avoid overspecific knowledge when developing the item.
8. Avoid questions based on opinions.

6 SAQ specific Question-writing guidelines

1. Questions should use appropriate vocabulary for the learners level.
2. Avoid a negative wording in the question (e.g. Which is not the correct answer?).
3. Write the question in a way that there is only one correct answer.
4. Avoid answers that exceed a short sentence.
5. Indicate the degree of accuracy expected, when relevant.
6. Indicate whether the insertion of irrelevant elements will be penalized.

1- Vachon Lachiver E, St-Onge C, Cloutier J, Farand P. La rédaction de questions à choix multiple et de questions à réponse ouverte et courte pour les examens écrits dans les programmes de formation en santé : une étude docimologique des consignes de rédaction. Pédagogie Médicale. 2017; 18(2):55-64.