

Objective versus subjective methods to assess discipline-specific knowledge: a case for Extended Matching Questions (EMQs)

Slattery, R.M.^a

^aDepartment of Immunology, Monash University, Australia

Abstract

Background: Extended matching questions (EMQs) were introduced as an objective assessment tool into third year immunology undergraduate units at Monash University, Australia.

Aim: The performance of students examined objectively by multiple choice questions (MCQs) was compared to their performance assessed by EMQs; there was a high correlation coefficient between the two methods. EMQs were then introduced and the correlation of student performance between related units was measured as a function of percentage objective assessment. The correlation of student performance between units increased proportionally with objective assessment. Student performance in tasks assessed objectively and subjectively was then compared. The findings indicate marker bias contributes to the poor correlation between marks awarded objectively and subjectively.

Conclusion: EMQs are a valid method to objectively assess students and their increased inclusion in the assessment process increases the consistency of student marks. The subjective assessment of science communication skills introduces marker bias, indicating a need to identify, validate and implement, more objective methods for their assessment.

Keywords: *Extended matching question (EMQ); Objective assessment (OA); SA (SA); Marker bias; Discipline-specific assessment; Science communication assessment*

1. Introduction

1.1 Automation of assessment

There are two important reasons academics may be motivated to increase the proportion of assessments that utilize an automated process. The first has a pedagogical ideal, the second is purely pragmatic. The use of essay style questions to probe student knowledge is reported to have been used in China as far back as the 7th century, moving to the west in the mid 19th century (Lederman, 1988). Subjective tools to assess written and oral communication have since then been used extensively. However, a limitation of subjective assessment (SA) is the bias imposed by individual assessors (Langan et al., 2005; Malouff & Thorsteinsson, 2016). The presence of ‘marking bias’ necessitates either the use of multiple assessors for each student, or standardization of marks between assessors, to minimize the problem. Conversely, automated assessment processes require questions that can be marked objectively. The second reason for automation is a fiscal one: many universities worldwide are under increasing pressure to teach a greater number of students without commensurate resources (Blackmur, 2007). Automated processes can alleviate time pressures faced by academics.

1.2 Format of examination questions

Multiple Choice Questions (MCQs) are well established as an automated assessment tool (Moss, 2001), but have several drawbacks that has led some educators to limit their application (Tetteh & Sarpong, 2015). One problem is that they require the posing of numerous spurious answers that are plausible, but wrong, thus creating a culture of mistrust between student and educator. The alternative is to write MCQs as negative questions, requiring the student select one incorrect answer. Whether written in the positive or negative format, the erroneous selection of an incorrect answer risks corruption of the student’s knowledge. This is especially true with the increased adrenaline release associated with the examination experience (Schwabe et al., 2012).

An alternative to MCQs is the Extended Matching Question (EMQ) format, in which a context statement precedes multiple alternative correct answers, followed by related questions. Students inform an examination card with the letters pertaining to the most correct answers, which then undergo automated processing. EMQs have several important pedagogical advantages over MCQs (Beullens et al., 2002; Cramer et al., 2016). Firstly, the context statement serves to focus and settle the student under examination. Secondly, all alternative answers provided to the student are indeed correct, thus reminding the students of what they do know, and creating a positive examination environment. The third, and arguably most powerful, advantage of utilizing EMQs is the potential to write questions that probe the student’s differential knowledge, whereby they are required to distinguish

between two closely related alternatives. EMQs are therefore particularly well suited to health sciences education, where differential reasoning is inherent to successful learning (Beullens et al., 2002).

1.3 Introduction and approach to validate EMQs as an assessment tool

The content of the immunology units into which the EMQs were introduced comprises numerous immunological mechanisms underlying disease processes, requiring students to discern subtle features that distinguish one process from another, lending them well to this assessment approach. Historically these units relied on SA for 75% of the awarded marks, and 25% assessed by MCQs. The motivation for introducing EMQs was twofold: (i) to use a more objective assessment (OA) tool that could assess differential reasoning, and (ii) to promote time efficiency for academics during peak assessment periods.

2. Methods

EMQs were introduced over 3 years and the validity of EMQs measured over the course of the implementation, and for two years after implementation. Within each cohort student performance was measured as the number of correct answers across three examinations. Four approaches were taken to analyse the data. These included comparing the mean mark for each student's performance: 1. In MCQs and EMQs within the unit 2. In two closely related units, as a function of the percentage objective assessment 3. In subjectively and objectively assessed tasks within the unit, and 4. In oral assessment tasks from different marker groups within the unit. For each comparison made the coefficient of correlation was determined using Excel and Prism-graphpad statistical programs. Assessment of variance between marker groups was performed using Kruskal-Wallis nonparametric one-way ANOVA.

3. Results

3.1 EMQs are as valid an objective assessment tool as MCQs

Student performance was monitored within the unit undergoing transition from MCQ to EMQ based assessment. Student performance was collated across three examinations with a total of 150 questions comprising 40% MCQs and 60% EMQs. Each student was compared for their average mark in MCQs vs EMQs. A high correlation coefficient was measured in student performance assessed by these two methods (Figure 1), demonstrating EMQs are a valid OA tool. EMQs were thus introduced in subsequent years and the impact of increased percentages of OA was further monitored.

3.2 Increasing objective assessment did not impact on unit performance

We compared the mean student performance of each year's unit cohort as a function of percentage objective assessment. There was no significant difference in the mean mark of students in any of the years during which the percentage OA was increased from 25% to 72.5% (Mean unit mark ranging between 67.6 ± 9.6 and 69.8 ± 12).

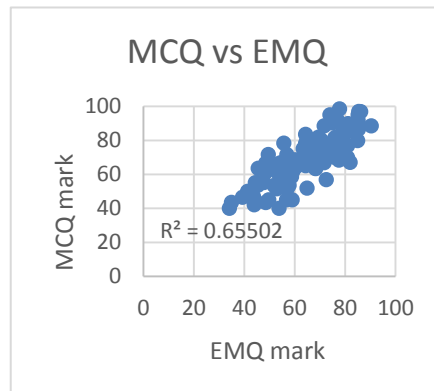


Figure 1. Comparison of performance in MCQ vs EMQ.

3.3 Objective assessment improved the correlation of student performance

Next, it was important to ascertain whether increasing the percentage OA perturbed student ranking. The performance of students co-enrolled in related units was compared across 3 years of transition from 25%-59% combined percentage OA, and 2 subsequent years in which the combined percentage OA was maintained at 59% (Figure 2). The results show that in 2012, prior to the introduction of EMQs, and when each of the units utilized only 25% OA in the form of MCQs, there was a poor correlation of student performance between the two units. In the subsequent year, the first semester unit remained unchanged (25% OA), and EMQs were introduced to the second semester unit, increasing the percentage OA to 65%. This led to an improved correlation of student performance between the units. In 2014 and subsequent years, after increase of the percentage OA to 45% and 72.5% in semester 1 and 2 units respectively, the correlation of student performance increased further. Indeed, the increased OA from 25% to 65%, almost doubled the coefficient of correlation between student performance in the two units (Figure 2 insert). This indicates that OA improved consistency of student performance in the two units, thus validating the reproducibility of OA as superior to that of SA.

3.4 Objectively and subjectively awarded marks were poorly correlated

The stronger correlation of student performance between units provided by OA raises questions about the accuracy of SA to measure student performance. To address this issue, we compared student performance measured by OA versus SA within the same unit.

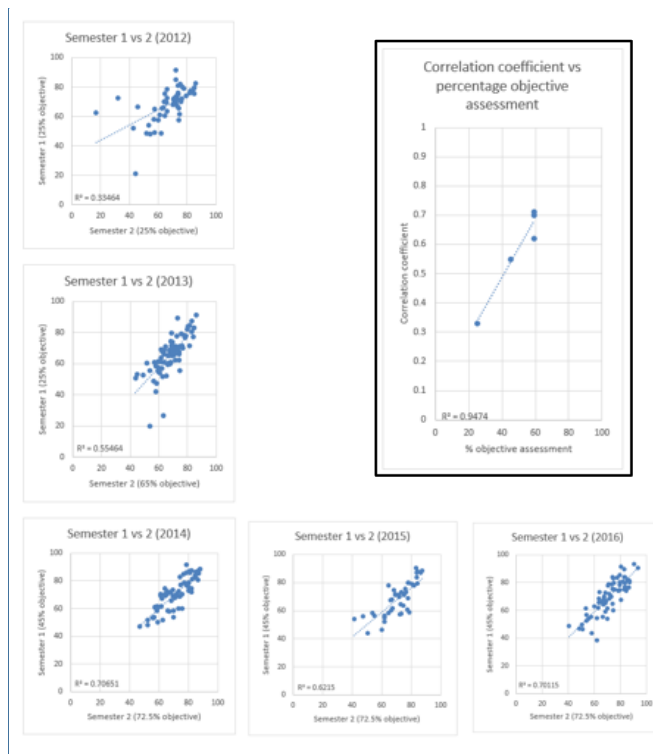


Figure 2. Comparison within the student cohort of performance in related units with increasing percentage objective assessment. Insert: Correlation of student marks as a function of percentage objective assessment.

3.4.1 Written and oral assessments versus EMQ

Prior to undertaking the written assessment task, students were required to read, interpret, and summarize in tabulated form, scientific data from two manuscripts. After receiving written and verbal feedback from a mentor, students wrote an essay discussing the data. Essays were subjectively assessed by markers against a rubric that defined the apportioning of marks for specific attributes present in the essay. Student performance in the written task was compared with that in OA (EMQs). There was a poor correlation between performance in the written assessment and performance measured objectively, despite clear marking guidelines and a marking rubric (figure 3a upper panel).

The oral assessment task was also the culmination of a formative learning process that required students to engage in critical problem solving in a small group setting, facilitated

by the tutor. Students were then required to orally present their solutions to the assessing tutor. Each student presented orally seven times throughout the semester. Oral assessment marks were compared with performance in OA (EMQs). Despite the fact that tutors were given assessment guidelines and advised to use the full spread of marks, there was a poor correlation between student performance in the oral assessment, measured subjectively, versus performance measured objectively by EMQs (Figure 3a lower panel).

The poor correlation between OA and SA is striking. It may be that student aptitude and confidence in the skills required to perform well in these tasks is poorly correlated with their ability to perform well in EMQ-based examinations. Another potential explanation for the poor correlation is the subjectivity in the allocation of marks for student performance in these tasks.

3.4.2 Marker bias in subjective assessment

To determine whether marker bias was consistently associated with specific individuals we compared the marks awarded by individual markers with marks awarded objectively for each student. While there was a high coefficient of correlation between the marks awarded by marker 1 for the oral communication task and those awarded objectively (correlation coefficient 0.66), there was no such correlation for any of the other five assessors (correlation co-efficients ranged between 0.00-0.19). Furthermore, the variance in awarded marks was not significantly correlated between marker groups (Kruskal-Wallis one-way ANOVA, Figure 3b). These data strongly suggest assessor bias is at least a contributing factor to the poor correlation between OA and SA.

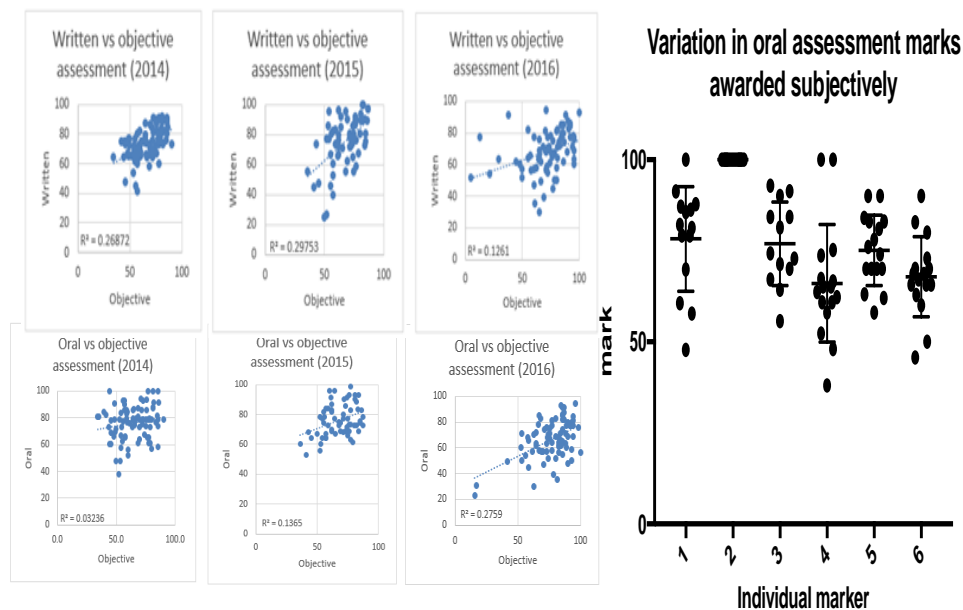


Figure 3. (a) Comparison of performance in written (top panels) and oral (bottom panels) assessment tasks versus objective assessment. (b) Variation in mark awarded by individual markers.

4. Discussion

EMQs were introduced to the capstone units of the final year Immunology course for two reasons. First, EMQs are superior to traditional OA and SA methods in their ability to probe student knowledge and its application to problem solving –this advantage of EMQs is important because problem solving ability is essential for students preparing to enter the professional community. The second motivation for introducing EMQs was to relieve the pressure on academics during peak work loads. The reported ability of EMQs to deeply penetrate the knowledge and problem solving ability of students has been documented (Bullens et al., 2002). However, the introduction of changes to assessment methodologies has the potential to introduce new variables that must be managed. Here we show there was a high degree of correlation between performance in EMQs and MCQs, and conclude that EMQs are as valid an OA tool as traditional MCQs. Because of the increased challenges inherent in problem-solving style questions, and student perception that EMQs are more difficult than MCQs, we further analysed whether the mean mark was decreased by the introduction of EMQs. The finding that there was no difference in the mean mark associated with the introduction of EMQs is important; not only does it validate the fairness of this approach, it also was an important statistic to provide students at the beginning of semester to develop their confidence in OA by EMQ. The shift towards an increased percentage OA improved the correlation of student performance between related units,

demonstrating the reproducibility of OA. Conversely, the poor correlation of student performance in written and oral assessment tasks measured subjectively, and EMQs measured objectively, raises important questions. How much of the lack of correlation is due to variance in student aptitude for the tasks under assessment, and how much is due to marker bias? It is difficult to independently measure student aptitude without compounding the question by the method of assessment (Damon, 2007). However, it was possible to independently measure marker bias by stratifying the data into individual marker groups. It is clear that marker bias significantly influences the scatter of marks and drives the poor correlation between OA and SA. Such unreliability of marking has been reported to occur despite provision of clear marking criteria, and is proposed to occur because of the complexity of assessment decisions (Bloxham et al., 2016).

While other influences, such as student aptitude, may also contribute to the discordance of student marks, the fact remains that SA approaches are limited by marker bias (Hathcoat et al., 2016). Automated assessment technology is clearly superior to human grading for written assessment tasks (Heit & Donaldson, 2016; Horn, 2009). Objective approaches need further exploration by the international education community to lead the pedagogical field towards the implementation of cross-validated, objective processes, to assess students' written and oral science communication abilities.

In conclusion we have shown that EMQs are a valid assessment tool and their inclusion strengthens the reliability of the assessment process, because they are objective. Notwithstanding the importance of written and oral communication skills in science graduates, we propose that all discipline-specific components of scientific units be assessed objectively. And that communication skills, which are not assessable by EMQs, be assessed independently by a broad academic audience of sufficient size and diversity to obviate assessor bias. This would more closely represent the global audience with whom our students will professionally engage, once graduated.

Acknowledgements

Thank-you Drs A Barry & K Murphy, A/Profs F Alderrucio & M Wright for provision of unit 1 data and analysis programs & Prof C Doerig for suggestions and editorial assistance.

References

- Beullens, J., Bo Van, D., Jaspaert, H., & Janssen, P. J. (2002). Are extended-matching multiple-choice items appropriate for a final test in medical education? *Medical Teacher, 24*(4), 390-395.
- Blackmur, D. (2007). A Critical Analysis of the UNESCO/OECD Guidelines for Quality Provision of Cross-Border Higher Education. *Quality in Higher Education, 13*(2), 117-130. doi:10.1080/13538320701629137
- Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assessment and Evaluation in Higher Education, 41*(3), 466-481. doi:<http://dx.doi.org/10.1080/02602938.2015.1024607>
- Cramer, N., Asmar, A., Gorman, L., Gros, B., Harris, D., Howard, T., . . . Kibble, J. D. (2016). Application of a utility analysis to evaluate a novel assessment tool for clinically oriented physiology and pharmacology. *Adv Physiol Educ, 40*(3), 304-312. doi:10.1152/advan.00140.2015
- Damon, W. (2007). DISPOSITIONS AND TEACHER ASSESSMENT: THE NEED FOR A MORE RIGOROUS DEFINITION. *Journal of Teacher Education, 58*(5), 365.
- Hathcoat, J. D., Penn, J. D., Barnes, L. L., & Comer, J. C. (2016). A Second Dystopia in Education: Validity Issues in Authentic Assessment Practices. *Research in Higher Education, 57*(7), 892-912. doi:<http://dx.doi.org/10.1007/s11162-016-9407-1>
- Heit, J. P., & Donaldson, R. P. (2016, 2016). *Comparing Human Grading to Automated Assessment Technology: Implications for the 21st Century Classroom*, Singapore.
- Horn, B. (2009). *Developing a computer-assisted writing assessment tool for discipline - specific writing*. (3370623 Ph.D.), Northern Arizona University, Ann Arbor. Education Database database.
- Langan, M. A., Wheeler, P. C., Shaw, E. M., Haines, B. J., & et al. (2005). Peer assessment of oral presentations: Effects of student gender, university affiliation and participation in the development of assessment criteria. *Assessment and Evaluation in Higher Education, 30*(1), 21-34.
- Lederman, M. J. (1988). "Why Test?". *Journal of Basic Writing, 7*(1), 38.
- Malouff, J. M., & Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education, 60*(3), 245-256. doi:<http://dx.doi.org/10.1177/0004944116664618>
- Moss, E. (2001). Multiple choice questions: their value as an assessment tool. *Curr Opin Anaesthesiol, 14*(6), 661-666.
- Schwabe, L., Joëls, M., Roozendaal, B., Wolf, O. T., & Oitzl, M. S. (2012). Stress effects on memory: An update and integration. *Neuroscience & Biobehavioral Reviews, 36*(7), 1740-1749. doi:<http://dx.doi.org/10.1016/j.neubiorev.2011.07.002>
- Tetteh, G. A., & Sarpong, F. A.-A. (2015). Influence of type of assessment and stress on the learning outcome. *Journal of International Education in Business, 8*(2), 125-144.