Canadian Medical Education Journal

Brief Reports

The impact of systematically repairing multiple choice questions with low discrimination on assessment reliability: an interrupted time series analysis

L'impact de la correction systématique des questions à choix multiples ayant une faible discrimination sur la fiabilité de l'évaluation : une analyse via des séries chronologiques interrompues

Janeve Desy,[1] Adrian Harvey,[1] Sarah Weeks,[1] Kevin D Busche,[1] Kerri Martin,[1] Michael Paget,[1] Christopher Naugler,[1] Kevin McLaughlin[1]

[1]Office of Undergraduate Medical Education, Cummings School of Medicine, University of Calgary, Alberta, Canada.

Correspondence to: Dr. Kevin McLaughlin, Office of Undergraduate Medical Education, Cummings School of Medicine, University of Calgary, Health Sciences Centre, 3330 Hospital Drive NW, Calgary, Alberta, Canada T2N 4N1; email: kmclaugh@ucalgary.ca

Published ahead of issue: Mar 12, 2024 CMEJ 2024 Available at https://doi.org/10.36834/cmej.77596

Abstract

At our centre, we introduced a continuous quality improvement (CQI) initiative during academic year 2018-19 targeting for repair multiple choice question (MCQ) items with discrimination index (*D*) < 0.1. The purpose of this study was to assess the impact of this initiative on reliability/internal consistency of our assessments. Our participants were medical students during academic years 2015-16 to 2020-21 and our data were summative MCQ assessments during this time. Since the goal was to systematically review and improve summative assessments in our undergraduate program on an ongoing basis, we used interrupted time series analysis to assess the impact on reliability. Between 2015-16 and 2017-18 there was a significant negative trend in the mean alpha coefficient for MCQ exams (regression coefficient -0.027 [-0.008, -0.047], $p = 0.024$). In the academic year following the introduction of our initiative (2018-19) there was a significant increase in the mean alpha coefficient (regression coefficient 0.113 [0.063, 0.163], $p = 0.010$) which was then followed by a significant positive post-intervention trend (regression coefficient 0.056 [0.037, 0.075], $p = 0.006$). In conclusion, our CQI intervention resulted in an immediate and progressive improvement reliability of our MCQ assessments.

Résumé

Dans notre centre, nous avons introduit une initiative d'amélioration continue de la qualité (ACQ) au cours de l'année académique 2018-19 ciblant la correction des questions à choix multiples (QCM) dont l'indice de discrimination (D) est < 0,1. Le but de cette étude était d'évaluer l'impact de cette initiative sur la fiabilité/cohérence interne de nos évaluations. Nos participants étaient des étudiants en médecine au cours des années académiques 2015-16 à 2020-21 et nos données provenaient d'évaluations sommatives par QCM au cours de cette période. Comme l'objectif était de revoir et d'améliorer systématiquement les évaluations sommatives dans notre programme prégradué sur une base continue, nous avons utilisé une analyse basée sur des séries chronologies interrompues pour évaluer l'impact sur la fiabilité. Entre 2015-16 et 2017-18, il y a eu une tendance négative significative dans le coefficient alpha moyen pour les examens utilisant des QCM (coefficient de régression -0,027 [-0,008, -0,047], p = 0,024). Au cours de l'année académique suivant l'introduction de notre initiative (2018-19), il y a eu une augmentation significative du coefficient alpha moyen (coefficient de régression 0,113 [0,063, 0,163], p = 0,010) qui a été suivie d'une tendance positive significative après l'intervention (coefficient de régression 0,056 [0,037, 0,075], p = 0,006). En conclusion, notre intervention d'ACQ a entraîné une amélioration immédiate et progressive de la fiabilité de nos évaluations par QCM.

Introduction

"Reliability is the precondition for trust" –
Wolfgang Schauble

Those of us involved in creating assessment tools must be prepared to defend the *validity* of these tools. Validity is best viewed as a hypothesis with multiple sources of evidence contributing to the decision to accept or reject the hypothesis of validity of a specific assessment in a specific context.[1] While validity frameworks have evolved over time,[1-4] *reliability* has formed the core of validity arguments since the origins of Classical Test Theory (CTT) at the start of the 20th century. Maximal validity coefficient for any criterion is the square root of the product of the reliability coefficients of the assessment and criterion under study and therefore validity is directly limited by reliability.[1,2,5,6]

In a multiple choice question (MCQ) assessment, each individual question/item contributes to overall reliability and validity of the assessment, and for this reason item-level statistics, such as *discrimination*, are calculated for each question and reviewed alongside the alpha coefficient.[5,7,8] Cronbach's alpha coefficient, which ranges from 0 to 1, is one of the most commonly used measures of reliability for MCQ assessments, and the higher the alpha coefficient the greater is the covariance between the individual items making up the assessment and the overall variance of the assessment.[8] The discrimination index (*D*) for each question is typically assessed by dividing students into high and low performers on the assessment overall and then comparing the proportion of high versus low performers that answered correctly.[9] MCQ items may have low *D* due to a keying error, but alternative explanations include discrepancy between what students were taught and the designated correct answer to the question; a poorly worded, misleading, or cueing question stem; or poorly functioning distractors.[10-12] (The role of distractors in an MCQ is to attract poor performers and if the distractors fail to do so then the number of students answering correctly will be high in both low and high performance and *D* will be low.)

At our centre we introduced a post-exam continuous quality improvement (CQI) initiative during academic year 2018-19 where we targeted for repair items considered to have poor discrimination ($D < 0.1$).[13] We used interrupted time series analysis (ITSA) to assess the impact of our CQI initiative on the reliability/internal consistency of our MCQ assessments.[14] We predicted that if our intervention improves reliability then we should see a positive trend over time in the mean alpha coefficient of these assessments following the introduction of our CQI initiative.

Methods

This was a retrospective intervention study using an ITSA design. The Conjoint Research Ethics Board at the University of Calgary approved our study (REB21-1455).

Participants

Our participants were medical students at the Cumming School of Medicine, University of Calgary during academic years 2015-16 to 2020-21. We have a three-year undergraduate curriculum during which the first 18 months is a pre-clerkship Clinical Presentation Curriculum followed by the typical clinical clerkship for the remainder of the curriculum.[15]

Materials

The data source for our study was summative MCQ assessments on the seven core preclerkship courses and eight clerkship rotations. Each course or clerkship rotation has two or three MCQ exams that can be used as the summative assessment for any given academic year. Typically, our MCQ assessments are Single Answer (SA) format with four options, three of which are distractors. Student performance on these assessments and psychometric analysis of each assessment are stored on our undergraduate medical education assessment database. To calculate *D*, we divide the group of students into approximate quarters while ensuring that the number of students included in the top and bottom quarters are the same. We then use the formula: D = UG-LG/n, where UG is the number of students in the upper group answering correctly, LG is the number of students in the lower group answering correctly, and n is the number of students in each group.

Procedure

Prior to academic year 2018-19, the post-exam review personnel included the Director of Student Evaluation and student representatives for each mandatory assessment. This group would review student performance and feedback on each question and decide on whether to remove items or adjust the minimum performance level for questions that were not removed. Beginning in academic year 2018-19, we revised the post-exam review process so that student representatives were no longer involved in this process and, instead, we followed the CQI process

illustrated in Figure 1 after all summative MCQ assessments. Since the goal of our CQI initiative was to improve validity in general, in addition to targeting items with poor reliability our intervention also included components designed to strengthen other sources of validity evidence, such as "content" (ensuring the content of the assessment matched the course blueprint) and "response process" (reviewing students' comments as part of the post-exam review) validity.[1]

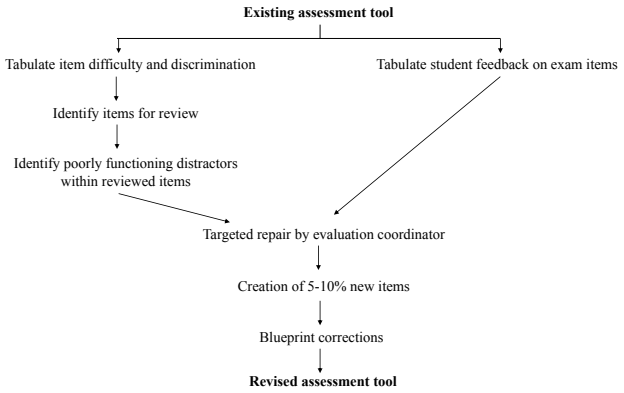

Figure 1. Cycle for improving/maintaining exam validity and quality (introduced during academic year 2018-19)

After flagging questions with low *D*, we then identified "poorly functioning distractors" in the answer choices for these questions. We classified an option as a poorly functioning distractor if this had a selection rate < 5%, or if this appeared to be negatively discriminating (i.e., this was selected more frequently by students who had higher overall performance that those with lower performance).[12,16] (Although we routinely calculate item difficulty (*P* value), this statistic did not contribute to our decision to target items for repair.[5]) Our evaluation team then met with the director and evaluation coordinator for each course or clerkship to explain the finding on item analysis and discuss revisions to the assessment components prior to future use of these assessments. In addition to ensuring that the assessment items matched the published course blueprint, we discussed strategies for repair of items with low *D*. This included reviewing student comments related to flagged items and then deciding whether the evaluation coordinator should focus on replacing/repairing poorly functioning distractors, revising the question stem, or both. In order to preserve content validity evidence, the clinical presentation of each repaired or replaced item did not change. Similarly, the format of repaired/replaced items matched that of the previous version (Single Answer (SA) with four options, three of which are distractors).

Analyses

We used ITSA to assess the impact of our CQI intervention of the reliability of our MCQ assessments. ITSA allows us to follow an outcome variable over multiple, equally spaced time periods – both before and after an intervention – thus allowing us to assess whether the intervention "interrupted" the level and/or trend in this outcome.[14,17,18] Our outcome variable in this study was the mean alpha coefficient for all summative MCQ assessments at the end of each academic year, plotted over time and an intervention term (corresponding to the introduction of our CQI initiative) to study the impact of our CQI initiative on the reliability of our MCQ exams. We gathered data from three years prior to and three years after the introduction of our CQI initiative. Anticipating autocorrelation, we followed the autoregressive integrated moving average (ARIMA) method of Prais-Winsten and Cochrane-Orcutt, which uses generalized least-squares to estimate parameters in a linear regression model where errors are assumed to be autoregressive.[14,17,18] We assessed the impact of our initiative by examining both an immediate "change in level" and then a subsequent "change in slope" following the intervention.[17] We used STATA® version 15.0 (College Station, Texas) for our statistical analyses.

Results

During academic years 2015-16 to 2017-18, there was a significant negative trend in the mean alpha coefficient for MCQ exams (regression coefficient -0.027 [-0.008, -0.047], $p = 0.024$). In the academic year following the introduction of our CQI initiative (2018-19) there was a significant increase in the mean alpha coefficient (regression coefficient 0.113 [0.063, 0.163], $p = 0.010$) which was then followed by a significant positive post-intervention trend (regression coefficient 0.056 [0.037, 0.075], $p = 0.006$). The pre and post-intervention trends and initial change in mean alpha coefficient are shown in Figure 2. The mean alpha for each year and standard deviation is shown in the Table 1.

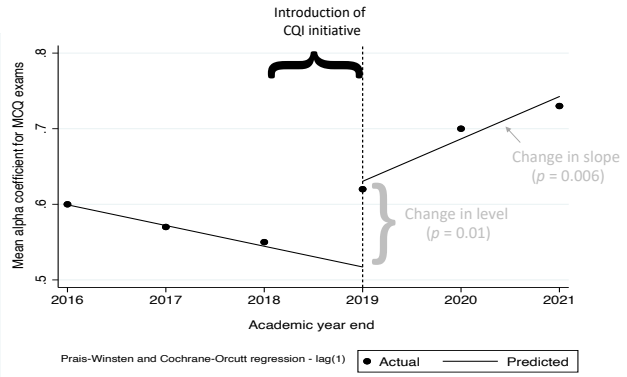

Figure 2. Interrupted time series analysis showing the trend in alpha coefficient before and after the introduction of a QI initiative to improve reliability of MCQ assessment

Table 1. Details of Alpha Coefficients over the course of our study

Year	Mean Alpha	Standard Deviation
2016	0.6	0.18
2017	0.57	0.09
2018	0.55	0.08
2019	0.62	0.08
2020	0.7	0.06
2021	0.73	0.06

Discussion

Given the importance of validity in assessment, and the central role of reliability in the validity argument, assessing and improving reliability are important tasks for those involved in assessment. By understanding the impact of discrimination on reliability and the potential causes of low discrimination, we can design interventions to improve item discrimination with the hope that this will then translate into improved reliability and validity. In this study we describe one such intervention, in the form of a CQI initiative, which led to an immediate and sustained improvement in the reliability of our MCQ assessments. Prior studies using descriptive statistics have suggested such an improvement, but to our knowledge this is the first study to use an ITSA design to formally assess the impact of such an intervention.[16]

Our study has some important limitations: we used CTT rather than Item Response Theory (IRT) to provide data on the quality of test items due to the fact that our sample size for assessments was historically too small to allow us to perform a 3-parameter IRT model;[19] this was a CQI initiative rather than a research trial; although data were gathered prospectively, our data analysis was retrospective; students were exposed to our intervention incrementally over time rather than a hard start to our study with random allocation of our intervention; our study was conducted in a single undergraduate curriculum; and the fact that our initiative was intentionally multifaceted limits our ability to tease apart the independent contribution of each component of our intervention. Finally, although our intervention began approximately two years before the SARS-CoV-2 pandemic was declared, data gathered during the final year of our study may have been affected by this pandemic (although this would not explain the initial change in the alpha coefficient or the first part of the post-intervention trend).

Conclusion

In this study we describe a CQI initiative that could easily be incorporated into the post-exam review process of most medical schools. Although we cannot generalize the findings of this single-centre study, we found that this simple intervention resulted in an immediate and progressive improvement in the reliability of our MCQ assessments. While MCQ assessments have limitations,[20] they are unlikely to disappear in the near future – so simple and effective interventions that improve reliability and the overall validity argument are worthwhile.

Conflicts of Interest: None of the authors have conflict of interest to declare regarding the work presented in this manuscript.
Funding: There was no funding for this study.
Edited by: Lisa Schwartz (section editor); Cindy Schmidt (senior section editor); Marcel D'Eon (editor-in-chief)

References

1. Messick S. *Validity*. 3rd ed. New York, NY: American Council on Education and Macmillan, 1989.
2. Kane MT. Validation. In: Brennan RL, ed. *Educational measurement.* 4th ed. Westport.: Praeger; 2006:17-64.
3. Messick S. The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher* 1994;32:13-23. https://doi.org/10.2307/1176219
4. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ* 2015;49(6):560-75. https://doi.org/10.1111/medu.12678
5. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. *Med Educ* 2010;44(1):109-17. https://doi.org/10.1111/j.1365-2923.2009.03425.x
6. Thorndike RL, Hagen E. *Measurement and evaluation in psychology and education.* New York: John Wiley and Sons Inc, 1961.
7. Richardson MW. Notes on the rationale of item analysis. *Psychometrika* 1936;1:69-76. https://doi.org/10.1007/BF02287926

8. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16:297-334. https://doi.org/10.1007/BF02310555
9. Glass GV, Hopkins, K.D. *Statistical methods in education and psychology.* 3rd ed. Needham Heights, MA: Allyn and Bacon, 1995.
10. Chiavaroli N. Negatively-worded multiple choice questions: an avoidable threat to validity. *Pract Assessment Res Eval* 2017;22:1-14. https://doi.org/10.1201/9780203739976-1
11. Schuwirth LW, van der Vleuten CP, Donkers HH. A closer look at cueing effects in multiple-choice questions. *Med Educ* 1996;30(1):44-9. https://doi.org/10.1111/j.1365-2923.1996.tb00716.x
12. Rodriguez MC, Kettler RJ, Elliott SN. *Distractor functioning in modified items for test accessibility.* SAGE Open 2014;4(4). https://doi.org/10.1177/2158244014553586
13. Office of Educational Assessment UoW. *Understanding item analyses.* Available from https://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/
14. McDowall D, McCleary R, Meidinger EE, Hay RA. *Interrupted time series analysis.* Newbury Park, CA: Sage Publications, 1980. https://doi.org/10.4135/9781412984607
15. 15. Mandin H, Harasym P, Eagle C, Watanabe M. Developing a "clinical presentation" curriculum at the University of Calgary. *Acad Med* 1995;70(3):186-93. https://doi.org/10.1097/00001888-199503000-00008
16. Ali SH, Carr PA, Ruit KG. Validity and reliability of scores obtained on multiple-choice questions: why functioning distractors matter. *J Schol Teach Learn* 2016;16:1-14. https://doi.org/10.14434/josotl.v16i1.19106
17. Hudson J, Fielding S, Ramsay CR. Methodology and reporting characteristics of studies using interrupted time series design in healthcare. *BMC Med Res Methodol* 2019;19(1):137. https://doi.org/10.1186/s12874-019-0777-x
18. Linden A. Conducting interrupted time-series analysis for single- and multiple-group comparisons. *Stata J.* 2015;15:480-500. https://doi.org/10.1177/1536867X1501500208
19. Jiang S, Wang C, Weiss DJ. Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Front Psychol* 2016;7:109. https://doi.org/10.3389/fpsyg.2016.00109
20. Strauss V. *The real problem with multiple-choice tests.* The Washington Post2013.