

An Elaborate Scheme to Investigate the Psychometric Properties of the Clinical-Decision Making Exam Questions

Isaac Li PhD¹, Qionggiong Liu MS¹, Yi Wang MS¹, Edward Tsai PhD¹

¹National Board of Osteopathic Medical Examiners, Chicago, Illinois

Introduction

With a focus on competency-based outcomes¹⁻³ in graduate medical education and assessment, a novel way to assess clinical decision-making (CDM) skills has emerged.^{2,3} CDM cases center on key features (KFs), critical steps required to diagnose and treat patients in clinical case scenarios.^{2,3} NBOME supplements multiple choice questions (MCQs) with CDM cases in the COMLEX-USA Level 3 exam to assess residents' ability to choose appropriate tests, procedures, and treatments and ensure patient safety. (Online practice can be found at <http://www.nbome.org/cdm.asp>.)

CDM/KF cases include a clinical scenario comprised of a patient presentation, vital signs, and laboratory or instrument findings, followed by Short Answer (SA) or Extended Multiple Choice (EMC) questions ordered in a clinical sequence. Multiple responses are allowed. If responses which might harm the patient—"killer" options—were chosen or provided, zero points are awarded; an excessive number of acceptable responses imply over-treatment and earns zero credit. A question may consist of one or more KFs scored dichotomously or polytomously. The multi-layered design calls for nonconventional but consonant analysis strategies to warrant the validity and reliability of the test. Primarily, using a variety of statistics analysts strive to maximize the amount of information available for content developers to enhance item quality at reduced costs. Research staff at NBOME has established a scheme of psychometric analysis that includes both case- and item-level overview and in-depth response-level feedback that offers traditional and alternative indices extended from those used for MCQs. The scheme maintains an emphasis on visual display of statistical indices.

Objectives

1. Illustrate graphically the various analyses employed to investigate the psychometric performance of the CDM/KF items for the purpose of item development.
2. Demonstrate the complexity and multitude of evaluative information extracted from the CDM cases.

Methods

The illustration here is based on response data from 54 CDM/KF cases administered in 2016, including 74 SA and 101 EMC items. 2,327 candidates responded to these cases. Overall item performance is given in the table below.

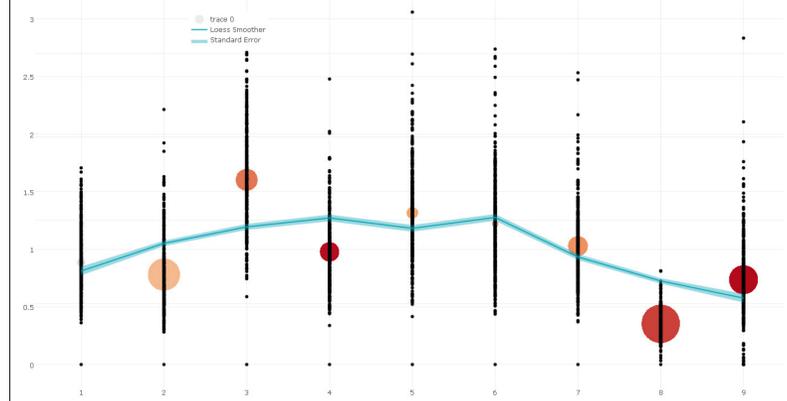
	P-Val_MN	P-Val_SD	Rpbis_MN	Rpbis_SD
EMC	.613	.214	.241	.064
SA	.610	.233	.259	.062

Analyses at the case level involve difficulty and discrimination indices averaged over all KFs embedded in the case. In addition, factors like response time, speededness, question length, and position effect are taken into account because they are instructive to test forms assembly. At the item (KF) level, indices like p-value, point-biserial correlation (r-value), and response time, as well as their relationship, are examined. For partial credit (polytomous) items, polyserial correlation is used to evaluate the discrimination power. The Rasch model is fitted to the data to yield item measures, different item fit statistics, and item response function (not shown here due to limited space). Analysis of options, including "killer" and "exceed-the-limit", is the most informative for item reviewers for revealing the proportion of candidates picking each response and their correlation with the sum score.

Results

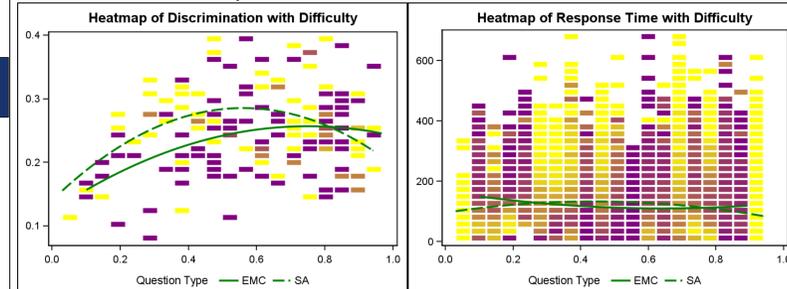
Case Analysis

Each CDM exam form contained 9 cases. In the graph below, the average per-word response time for each case is fitted with the Loess function by appearance order. The slight drop-off after position 6 suggested that the candidates in general intensified their efforts towards the end. Such efforts, however, did not appear to have led to more mistakes as the last two cases were relatively easy (larger bubbles). The discrimination power of the cases did not suffer from their positions either. (Redder bubbles reflect higher point-biserials.)

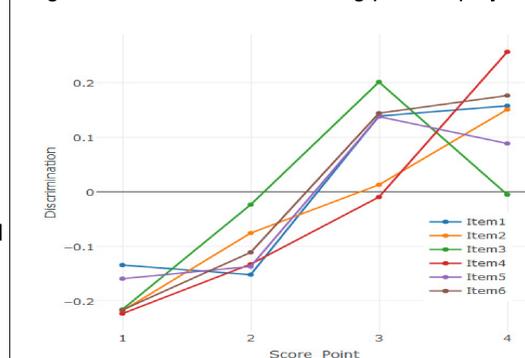


Item Analysis

From the distribution of the p-values and r-values in the left plot below the overall performance of items were good. Items of medium difficulty (~.6) discriminated the most and more items were easier (~.8) than hard. The few items with very low r-value were obvious for item reviewers. When p-values are plotted against response time, it is shown that the overall difficulty of the two item types are about the same but there is more variation with SA response time.



In general, more competent candidates do better on a partial-credit item. We expect candidates with the highest points on the item obtain the highest total score. The resulting positive polyserial correlation implies

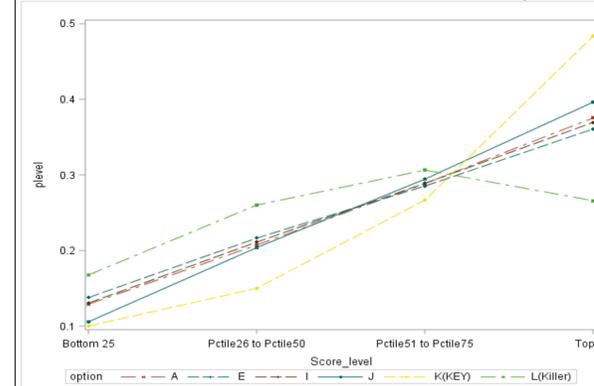


r-values of each category increasing monotonously. For example, the 6 items (1-4 points) graphed on the left hold this trend except for two items at point 4, indicating potential collapsing of point 3 and 4 during item revision.

Results (continued)

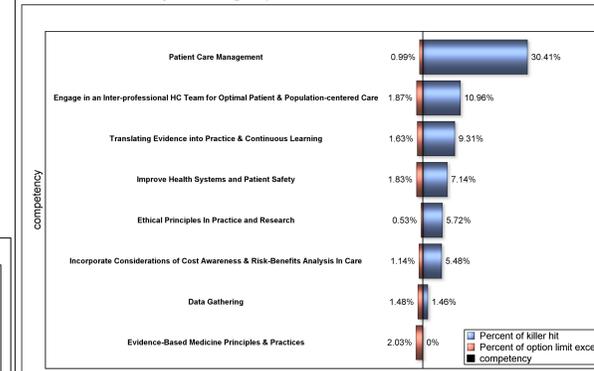
Response Analysis

Studying what responses are popular in a multiple-response item can supply specific information to content developers. In the item plotted below, the killer was answered more frequently at lower score levels than the key but far less by the top group. Empirical data have shown that candidates who answered killer or exceeded response limits had lower



MCQ and CDM scores than those who did not.⁴ Correlation analysis also pointed out that hitting a killer is associated with worse performing candidates consistently. Options A, E, I, and J were

picked by about the same percentages of candidates at each score level. Neither key features nor harmful to patients, they are considered acceptable responses not credited for that capable candidates were more likely to be able to identify. The graph below reveals that a killer response was given by



candidates to over 30% of the items under "Patient Care Management", more often than others. Killers reflect harmful or detrimental outcome, which fits with this competency best.

Conclusions

This presentation illustrates a comprehensive scheme of psychometric analyses devised to collect pertinent statistics from different sources inherent in the computer-based CDM/KF questions: cases, key features, options, and in particular, killer options and response limits. Test development will benefit from the feedback in revising and improving CDM/KF cases to help this innovative assessment realize its full potential in evaluating clinical competence and patient safety management.

References

- 1 Holmboe ES, Ward DS, Reznick RK, Katsufakis PJ, Leslie KM, Patel VL, Ray DD, Nelson EA. Faculty development in assessment: The missing link in competency-based medical education. *Academic Medicine*. 2011; 86(4): 460-467.
- 2 Bordage G, Brailovsky C, Carretier H, Page G. Content validation of key features on a national examination of clinical decision-making skills. *Academic Medicine*. 1995; 70(4): 276-81.
- 3 Page G, Bordage G, Allen T. Developing key-feature problems and examinations to assess clinical decision-making skills. *Academic Medicine*. 1995; 70(3): 194-201.
- 4 Song H, Kalinowski K, Bates BP. *Clinical Decision-making Questions for Assessing Competence in Managing Patient Safety*. Paper presented at the 2016 annual meeting of the American Association of Colleges of Osteopathic Medicine (AACOM): Baltimore, MD.