

ARTICLE INFO

Submitted: 04-08-2019

Accepted: 25-10-2019

Online: 31-12-2019

Readability as a Source of Measurement Error in Medical Education Assessment

Dylan Harver¹, Kenneth D. Royal²¹Office of Academic Affairs, North Carolina State University,
UNITED STATES²Department of Clinical Sciences, North Carolina State University,
UNITED STATES

To cite this article: Harver D, Royal KD. Readability as a source of measurement error in medical education assessment. *Education in Medicine Journal*. 2019;11(4):71–74. <https://doi.org/10.21315/eimj2019.11.4.7>

To link to this article: <https://doi.org/10.21315/eimj2019.11.4.7>

ABSTRACT

Readability is a measure of the accessibility of a text to a reader. Readability scores should not exceed the readability levels of the intended audience. To date, the topic of readability has rarely been explored in the context of medical education assessment. Thus, the purpose of this pilot study was to investigate the potential relationship between readability measures and item difficulty estimates. We used two readability formulas, FOG and FORCAST, based on each formula's intended purposes and requirements for shorter texts. A sample of 853 multiple-choice questions (MCQs) were obtained and the difficulty values for each item were plotted relative to each item's readability score. Results indicate an association was present between items with greater difficulty (items answered correctly by 70% or fewer examinees) and items with a readability measure greater than 12.0. We conclude that empirical evidence was discernible to support long-standing theoretical evidence that readability issues may introduce measurement error and consequently threaten score validity.

Keywords: *Assessment, Evaluation, Measurement, Bias, Validity, Psychometrics*

CORRESPONDING AUTHOR

Dr. Kenneth Royal, Department of Clinical Sciences, North Carolina State University, 1060 William Moore Dr., Raleigh, NC 27607, United States | Email: kdroyal2@ncsu.edu

INTRODUCTION

Readability is a measure of the accessibility of a text to a reader and is used to inform writing appearing in medical resources, educational materials, newspaper articles and more (1). Readability formulas use variables such as sentence length and average syllables per word to measure the difficulty level of text. These measures of readability can help create materials that are more easily understood by the intended audience. Although multiple-

choice questions (MCQs) continue to make up a large part of many programmes' assessments, the topic of readability has rarely been explored in the context of medical education. This is a problem because measurement error stemming from readability may lead to biased assessments that result in invalid scores and inferences. Thus, the purpose of this study was to report on the findings from a small, yet innovative, pilot study conducted at a large college of veterinary medicine in the United States.

METHODS

We sought to explore the potential relationship between readability measures and item difficulty estimates using a convenience sample of MCQs. We began by carefully reviewing the criteria used to establish common readability formulas, including the New Dale-Chall, Flesch-Kincaid, SMOG, FOG, Modified Coleman-Liau Index and FORCAST formulas. Formulas that used variables such as number of sentences, specific text length, and number of difficult words as determined by primary school students were removed from consideration due to construct irrelevance variance issues (2). Ultimately, the FOG and FORCAST formulas (3–4) were deemed the most robust as the FOG formula compares syllables and sentence length and the FORCAST formula was designed specifically for MCQs.

We defined the criteria for a difficult item as a percent correct of ≤ 0.70 , as this is a common criterion used in the field of psychometrics (5). We defined the criteria for discerning an appropriately targeted readability measure at 12.0, in which the grade level represents the typical minimum requirement for college admission in the United States. A sample of MCQs was obtained by pooling all items ($n = 853$) on mid-term and final examinations administered to 100 first-year students in a college of veterinary medicine in the United States during the 2018 academic year. Data were analysed using SPSS statistical software (version 25.0).

RESULTS

Spearman's *rho* (ρ) correlation coefficients indicate a rather negligible statistical relationship between FOG scores and item difficulty estimates, $\rho = -0.053$, and FORCAST scores and item difficulty estimates, $\rho = -0.073$. However, this

appears a classic case of the “defense attorney fallacy” (6) in which large samples obscure small subsets of data. When evaluating the relationship using scatterplots, a clear association was present between items with greater difficulty (0.70) and items with a readability measure greater than 12.0 (see Figures 1 and 2). Items with difficulty estimates exceeding 0.80 (e.g., items answered correctly by 80% or more students) exhibited a wide range of readability levels. Further, even those items with the highest readability scores (indicating the most difficult to read content) often were answered correctly. However, this trend dissipates when examining items with difficulty estimates less than 0.70 (e.g., items that were answered correctly by 70% or fewer students). Interestingly, in most instances readability measures typically exceeded the minimum targeted measure of 12.0.

DISCUSSION

It is important to note that it is nearly impossible to discern why items may be deemed easy or difficult (e.g., instructional familiarity, deliberate study, guessing, etc.), (5, 7–8) thus a causal link cannot be established. Yet despite some, albeit limited, empirical evidence in this work that readability levels may present a source of measurement error, there is considerable theoretical evidence to support this assertion given an exhaustive body of research in the fields of language assessment and testing (9). If a link is ultimately established between excessively high readability scores and item difficulty, educators would be wise to identify these potentially biased items as candidates for potential revision before administering to students. This is particularly important in situations that carry moderate to high-stakes for examinees, such as mid-term and final examinations, progress tests, licensure and certification examinations, etc.

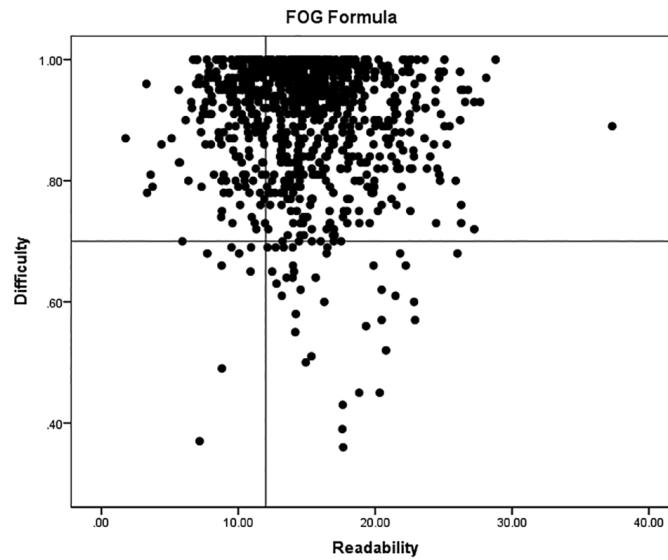


Figure 1: Scatterplot of FOG formula vs. difficulty.

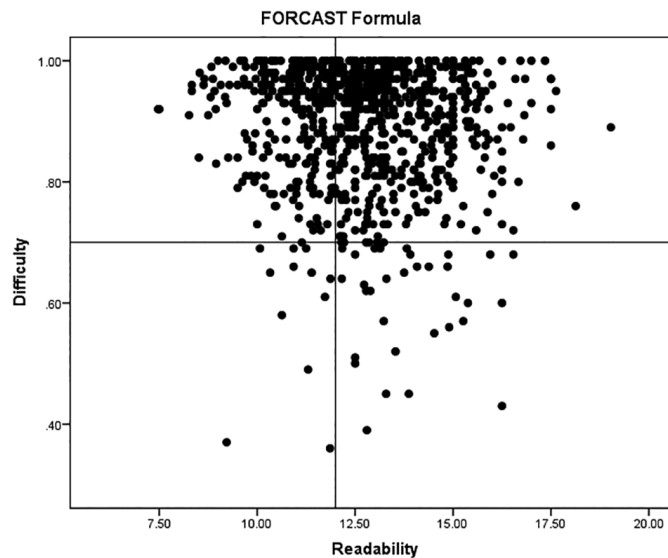


Figure 2: Scatterplot of FORCAST formula vs. difficulty.

A limitation of this pilot study was that it did not consider the reading levels of examinees and how various ability levels may impact examination performance. Future research should explore the interaction between examinees' reading levels and items' readability levels to determine the degree to which readability may pose a validity threat to score accuracy and meaning.

CONCLUSION

This innovative pilot study presents empirical evidence to support long-standing theoretical evidence that readability issues may introduce measurement error and consequently threaten score validity. Given this is the first study of its kind in medical education, more research is necessary to better understand the role readability issues may play in creating fair and defensible assessments.

REFERENCES

1. Oakland T, Lane HB. Language, reading, and readability formulas: implications for developing and adapting tests. *Int J Testing*. 2004;4(3):239–52. https://doi.org/10.1207/s15327574ijt0403_3
2. Messick S. Validity. In: Linn RL, editor. *Educational measurement*. 3rd ed. New York, NY: Macmillan; 1989. p. 13–103.
3. Gunning R. *The technique of clear writing*. New York: McGraw-Hill; 1952. p. 36–7.
4. Caylor JS, Stitch TG, Fox LC, Ford JP. *Methodologies for determining reading requirements of military occupational specialties: technical report no. 73-5*. Alexandria, VA: Human Resources Research Organization; 1973.
5. Royal KD. Using the nudge and shove methods to adjust item difficulty values. *J Vet Med Educ*. 2015;42(3):239–41. <https://doi.org/10.3138/jvme.0115-008R>
6. Thompson WC, Schumann EL. Interpretation of statistical evidence in criminal trials: the prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behaviour*. 1987;11(3):167–87. <https://doi.org/10.1007/BF01044641>
7. Haladyna T, Roid G. The role of instructional sensitivity in the empirical review of criterion-referenced test items. *J Educ Meas*. 1981;18:39–53. <https://doi.org/10.1111/j.1745-3984.1981.tb00841.x>
8. Royal KD, Hedgpeth MW, Smith KW, Kirk D. A method for investigating “instructional familiarity” and discerning authentic learning. *Ann Med Health Sci Res*. 2015;5(6):428–34. <https://doi.org/10.4103/2141-9248.177990>.
9. American Educational Research Association, American Psychological Association and National Council on measurement in education. *Standards for educational and psychological testing*. Washington, DC: Amer Educational Research Assn; 2014.