SHORT COMMUNICATION

Volume 12 Issue 4 2020

DOI: 10.21315/eimj2020.12.4.6

ARTICLE INFO

Submitted: 04-05-2020
Accepted: 18-09-2020
Online: 30-12-2020

On Misinterpretation of Course and Instruction Evaluation Data: How Relying Solely on Mean Scores Can Distort Score Meaning

Kenneth D. Royal

Department of Clinical Sciences, North Carolina State University, UNITED STATES OF AMERICA

To cite this article: Royal KD. On misinterpretation of course and instruction evaluation data: how relying solely on mean scores can distort score meaning. Education in Medicine Journal. 2020;12(4):43–45. https://doi.org/10.21315/eimj2020.12.4.6

To link to this article: https://doi.org/10.21315/eimj2020.12.4.6

ABSTRACT

One of the fundamental components of basic statistics is to examine a data distribution, namely its centre (median, mean, etc.), shape (skewness, symmetry, modality, etc.) and spread (variability, range, etc.). When examining research data, most educators are keenly aware of these fundamentals, but curiously seem to forget these fundamentals when examining course and instructor evaluation data. What often occurs is evaluators rely solely on mean score ratings as the basis for making inferences about a course and/or its instructor(s). This is problematic because a mean score alone does not illustrate the underlying score distribution, which in turn could completely alter the meaning of the data. The aim of this article is to present an illustrative example from basic statistics illustrating how course and instructor evaluation score inferences may be distorted by the underlying distribution of scores, thus threatening the validity of the measures. Suggestions for improving data reporting are provided.

Keywords: *Assessment, Evaluation, Measurement, Teaching evaluation, Statistics*

CORRESPONDING AUTHOR

Kenneth D. Royal, PhD, Department of Clinical Sciences, North Carolina State University, 1060 William Moore Dr., Raleigh, NC 27607, United States of America | E-mail: kdroyal2@ncsu.edu

INTRODUCTION

A fundamental components of basic statistical analysis involves examining data distributions, namely centre (median, mean, etc.), shape (skewness, symmetry, modality, etc.) and spread (variability, range, etc.). Although researchers routinely utilise this practice when examining research data, many seem to forget these fundamentals when examining course and instructor evaluation data (1). Unfortunately, many consumers of course and instructor evaluation data rely solely on mean score ratings as the basis for making inferences about a course and/or its instructor(s). This approach presents a considerable problem, as a single mean score does not illustrate the underlying score distribution. Thus, the purpose of this article is to provide an illustrative example from basic statistics illustrating how course and instructor evaluation score inferences may be distorted by the underlying distribution of scores.

AN ILLUSTRATIVE EXAMPLE

Most course and instructor evaluation instruments utilise either a 4-point or 5-point Likert-type rating scale (2–3). Although it is technically a statistical violation to treat Likert-type (ordinal level) data (e.g., strongly agree, agree, etc.) as an interval measure, (4) it remains a common practice in the social and behavioural sciences to calculate means (with accompanying standard deviations). Now, let us now consider a scenario in which the same course evaluation instrument utilising a 5-point rating scale was administered across six different courses. Next, assume that 50 students from each course provided ratings. Finally, assume that all six courses yielded the same mean score of 3.0 for an item evaluating the "overall quality of the course".

Imagine that this limited information was then presented to a curriculum committee that was charged with reviewing course quality in a medical training programme. By solely reviewing mean scores, committee members may be tempted to conclude each of the six courses averaged a 3.0/5.0, perhaps indicating "average" quality. Here, simulated data were used to construct six graphical distributions that all share a common mean score of 3.0 (see Figure 1).

A closer look at the score distributions for each course would identify additional, critical information necessary to make an informed judgement. The uniform distribution is mostly flat indicating a wide range of opinions. The normal distribution indicates most students provided ratings of 3, with fewer students selecting ratings of 1, 2, 4 and 5. The bimodal distribution indicates students had polarising opinions of the course, with an equal number of students selecting ratings of 1 and 5. Multiple peaks indicates two ratings were selected most predominantly. Edge aversion indicates respondents avoided selecting extreme ratings of 1 and 5. The flat distribution indicates students selected each category in equal quantities. In nearly all instances, the interpretation of a mean score of 3.0 could be very different depending on its underlying distribution of scores.

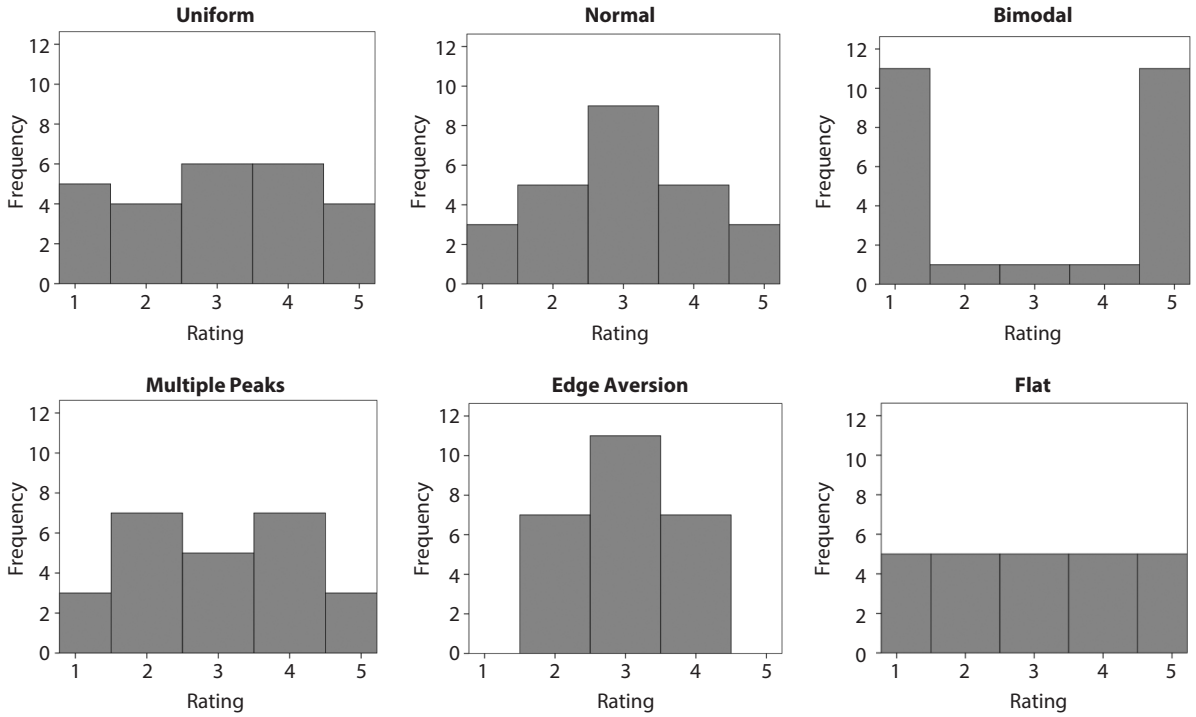

Figure 1: Six distributions with a mean score of 3.0.

RECOMMENDATIONS AND CAVEATS

Figure 1 illustrates how the meaning of a set of scores can vary dramatically depending upon its distribution. It should serve as a reminder to educators and evaluators that merely relying on mean score measures alone is an inadequate, and arguably irresponsible, practice. One way to ensure educators and evaluators are mindful of distributional properties is to report a summary of descriptive statistical information. For example, if a graphic of a score distribution is not feasible for a course and/or instructor evaluation report, then it is critical that other numeral data be provided. Such data should include frequency statistics indicating how often each rating scale category was selected, the standard deviation that accompanies each mean score, minimum and maximum ratings, and skewness and kurtosis measures to help describe the shape of the distribution in the absence of a graphic. Collectively, these pieces of information coupled with relevant training of statistical concepts will equip educators and evaluators with most all the statistical tools they need to properly discern course and instructor evaluation data.

It is also important to note the role of criterion-referenced versus norm-referenced inferences when making judgements about course and instructor evaluation data. Criterion-referenced inferences refer to those in which judgements are based relative to a standard; norm-referenced inferences refer to those in which judgements are based relative to other courses or instructors. In the context of course and instructor evaluations, norm-referenced inferences are inappropriate. Courses should not be compared because courses vary in content, difficulty, and other very important ways that destroy the integrity of such comparisons. However, some caution is also necessary for criterion-referenced inferences. That is, even if an institution uses a standard (e.g., a course must meet or exceed a quality score of 3.0), it is possible that the standard may be different across courses because many of the aforementioned factors (e.g., content, difficulty, etc.) can also influence ratings.

CONCLUSION

Relying solely on mean score measures is a problematic practice. Additional data (e.g., frequency statistics, standard deviations, minimum and maximum ratings, skewness and kurtosis measures) are necessary for evaluators to truly discern the meaningfulness of data. When effective training is coupled with responsible data interpretation practices evaluators will be equipped to properly discern course and instructor evaluation data.

REFERENCES

1. Krzywinski M, Altman N. Points of significance: error bars. Nat Methods. 2013;10(10):921–2. https://doi.org/10.1038/nmeth.2659

2. Royal KD. A guide for assessing the interpretive validity of faculty and course evaluations in medical schools. Med Sci Educ. 2016;26(4):711–7. https://doi.org/10.1007/s40670-016-0325-9

3. Royal KD. A guide for making valid interpretations about student evaluation of teaching (SET) results. J Vet Med Educ. 2016;44(2):316–22. https://doi.org/10.3138/jvme.1215-201R

4. Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. 3rd ed. Routledge; 2015. https://doi.org/10.4324/9781315814698