Performance Metric Distribution Characteristics of Medical School Exam Items

Nicholas B. Sajjadi, BS, OMSII, Lindsay Terry, BS OMSII, Joseph Price III, PhD
Oklahoma State University Center for Health Sciences: College of Osteopathic Medicine

MEDICINE

INTRODUCTION & OBJECTIVES

Academic assessment, commonly in the form of written exams, is used to measure students' progress in the mastery course material, to ensure learning objectives are achieved, and to provide instructors with feedback for improving the instruction efficacy [1,2]. Exam qualities are often ascertained by performing item analysis involving the inspection of individual exam questions, for which 3 metrics are commonly used: Difficulty, Discrimination Index, and Point Biserial.

Difficulty represents the portion students having answered an item correctly, typically expressed as a decimal where a value of "1.00" corresponds to 100% of students answering correctly. An item Difficulty of "1.00" potentially indicates low item difficulty. Discrimination Index (DI) is the capacity of an item to distinguish students based on their varying degrees of proficiency. DI is typically calculated by taking the difference between the number of students in the bottom 27% of performance who answered an item correctly and the number of students in the top 27% of performance who answered the item correctly then dividing by the total number of students in both groups. An item with a DI of 0.30 or greater is considered to have good discrimination, values of 0.29-0.10 signify fair discrimination and may suggest the need for item revision, a value of 0 shows no discrimination, and items with negative DIs should be completely revised. However, as the percent correct for an item tends toward 100% (or 0%), the possible extent of discrimination decreases to zero. Point Biserial (PB) is the Pearson Product Moment correlation of student responses to an item and overall exam performance. Strong correlations suggests that answering the item correctly was associated with high overall performance and vice versa. PB values of less than 0.2 typically indicate poor item quality, values of 0.2-0.3 are fair, values of 0.3-0.4 are good, and values of 0.4-0.7 are ideal, subject to the mentioned limits.

A previous investigation by Terry and Price [3] characterized these metrics for 61 exams from 26 first- and second-year courses administered at the Oklahoma State University College of Osteopathic Medicine finding the mean, median, standard deviation, skewness, and kurtosis for each metric. The findings are shown in Figure 1 for the sake of example. However, the normality of the metric distributions remains unknown and would impact the future choice of inferential statistical methods. Thus, the primary objective of this study is to determine the normality of the item Difficulty, DI, and PB for these 61 exams for use in future investigations seeking to improve student learning and to ensure continual enhancement of instruction.

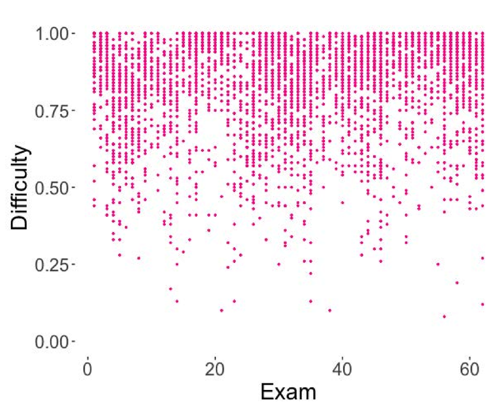

Figure 1a. Item Difficulty for each test item, all exams.

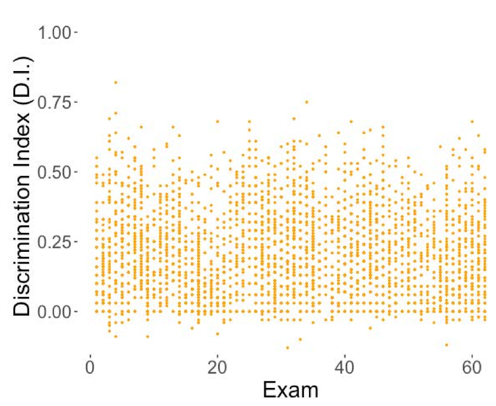

Figure 1b. Item DI for each test item, all exams.

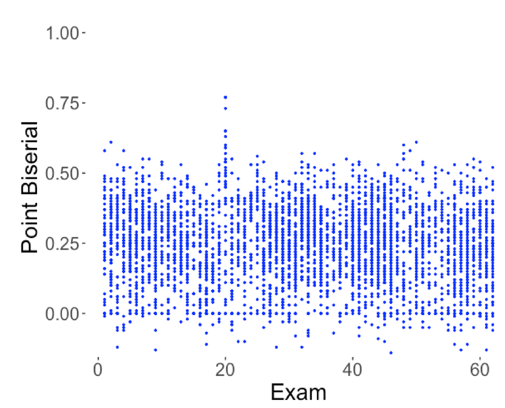

Figure 1c. Item PB for each test item, all exams.

METHODS

61 exam item datasets from the 2018-2019 academic year for the Class of 2021 and the Class of 2022 at Oklahoma State University College of Osteopathic Medicine were recorded, deidentified, then analyzed using the software suite R (version 4.0.2) and the integrated interface RStudio (Version 1.3.959).

Normality was assessed graphically using Q-Q plots and numerically using the Shapiro-Wilk Normality Test. Q-Q plots are classically used to graphically analyze deviation from normality, and better appreciate the effect size of deviation. Graphs were made using the package "ggqqplot," an extension of "ggpubr." The Shapiro-Wilk test was chosen because Monte Carlo simulations have shown that it has the best power for a given significance and accommodates the relatively small sample size for each exam [4,5]. The analysis' null hypothesis is that the data is normally distributed. Therefore, any p-value less than 0.05 indicates statistically significant deviations from a comparable normal distribution with the same mean and standard deviation as the data set. These Shapiro-Wilk p-values were then adjusted for any bias arising from the multiplicity problem using the Benjamini-Hochberg procedure (using the package "p.adjust"), which controls the false discovery rate and is more powerful than methods that control the family-wise error rate [6]. This methodology was applied in a previous study by Sajjadi et al [7].

RESULTS

Figure 2a. Q-Q plots for item Difficulty, all exams.

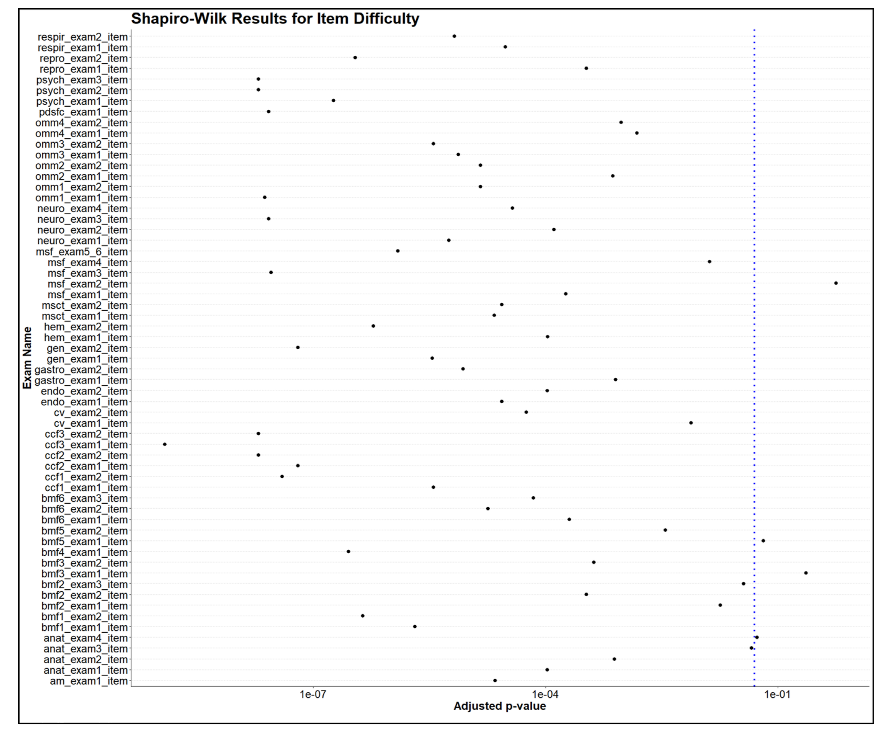

Figure 2b. Shapiro-Wilk BH adjusted p values for item Difficulty, all exams. The x-axis is logarithmic, and the dotted blue line is at x = 0.05.

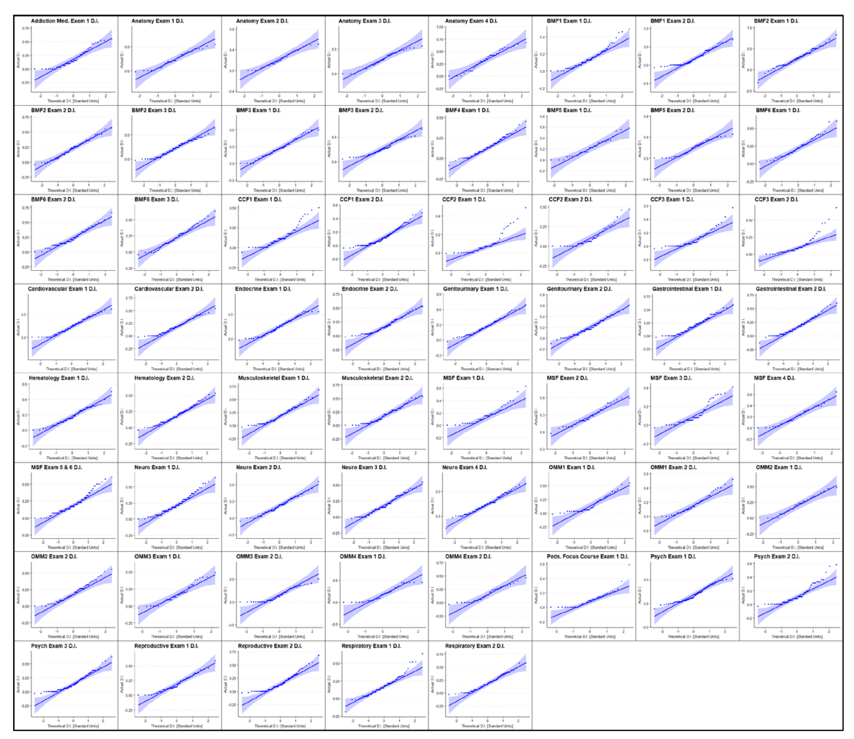

Figure 3a. Q-Q plots for item Discrimination Index, all exams.

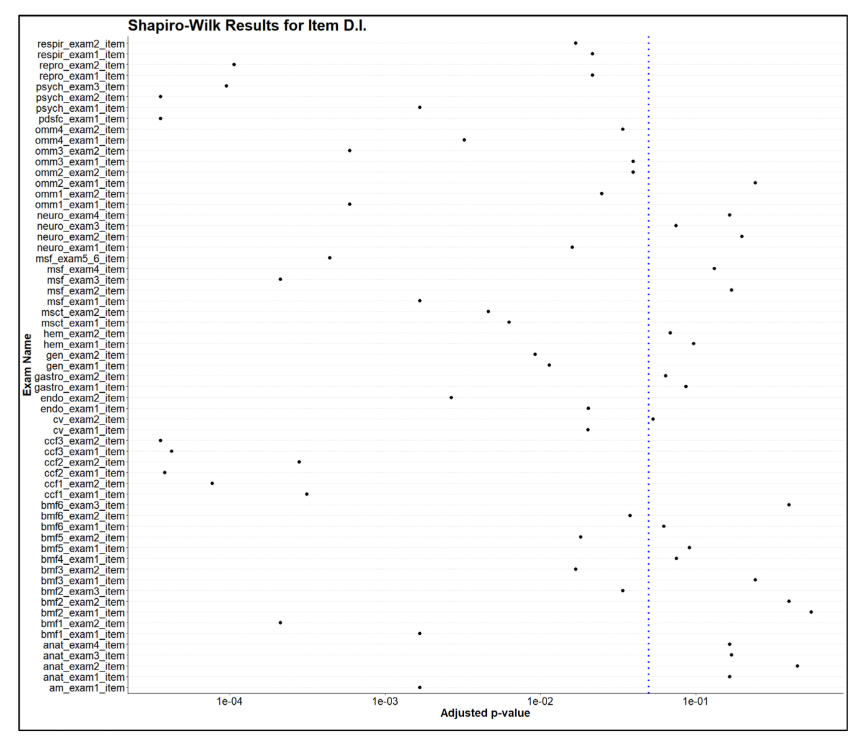

Figure 3b. Shapiro-Wilk adjusted p values for item Discrimination Index, all exams. The x-axis is logarithmic, and the dotted blue line is at x = 0.05.

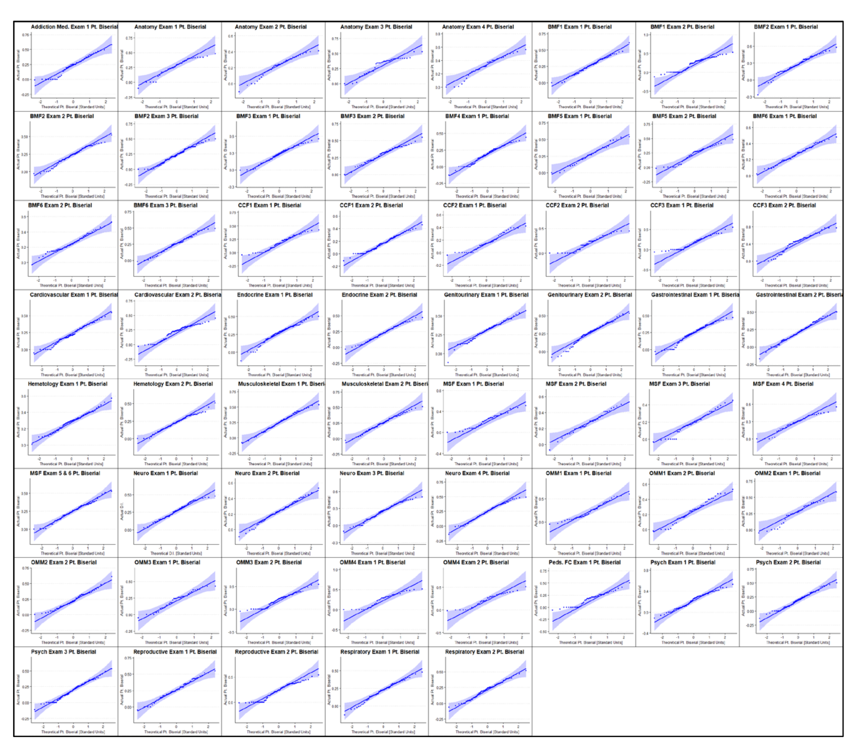

Figure 4a. Q-Q plots for item Point Biserial, all exams.

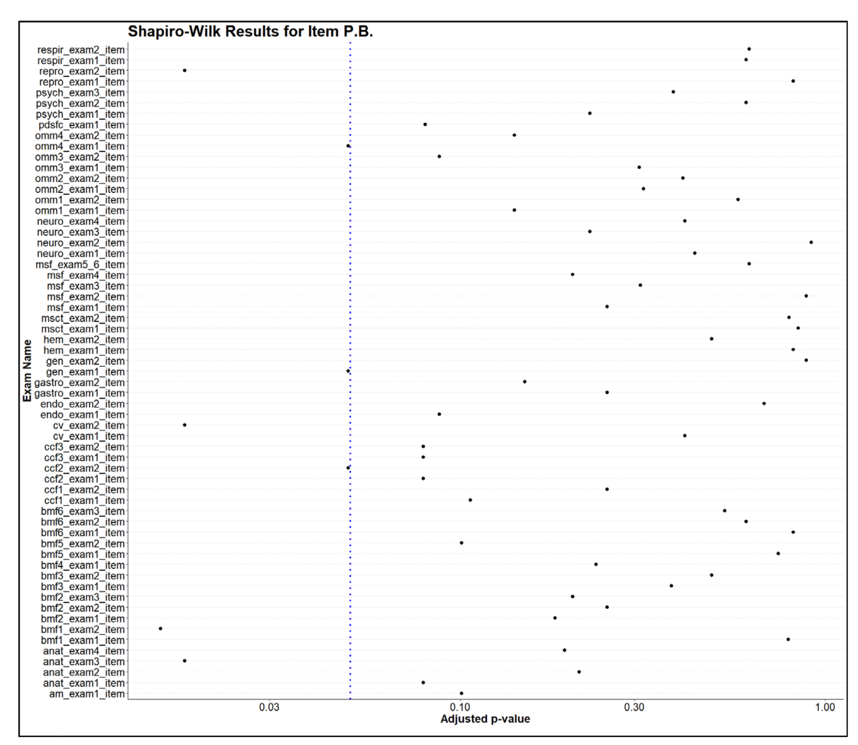

Figure 4b. Shapiro-Wilk adjusted p values for item Point Biserial, all exams. The x-axis is logarithmic, and the dotted blue line is at x = 0.05.

Table 1: Item Performance Normality Characteristics for 61 Exams

Item Performance Metric	Exams with Statistically Significant Deviations from Normality
Difficulty	57 (93.44%)
Discrimination Index	39 (63.93%)
Point Biserial	7 (11.48%)

DISCUSSION

Our results suggest that these 3 exam item performance indicators vary drastically in their distribution characteristics. Item Difficulty was by far the most deviated parameter as seen graphically by the heavily-tailed curves on most of the Q-Q plots. Many exams exhibited points falling outside of the 95% confidence interval at the far ends of the distribution. Trailing off ends may indicate that there is high variability at the end of the distribution leading to a tight cluster of many values centered about the mean which is suggestive of a leptokurtotic deviation. This peaking is in keeping with previous research where the mean Difficulty for this set was found to be 0.83 with a standard deviation of 0.05. The numerical analysis also supports the finding of non-normality. Taken together, these findings suggest that most exam items from this set are of consistently low difficulty.

The Point Biserial had the least deviations from normality, suggesting that inferential statistics relying on an assumption of normality may be appropriate for further analysis of this metric. The Discrimination Index was of an intermediate degree of deviation but tended towards non-normality. This is again in keeping with previous findings which suggested minimal kurtosis for PB and intermediate kurtosis for DI. We can see that the number of exams meeting statistical significance criteria (p<0.05 – indicated by items falling below the dashed blue lines in Figures 2b, 3b, and 4b) for deviations from normality is different for each parameter.

To our knowledge, not many studies have investigated the ideal distribution characteristics for these parameters when designing exams. It is entirely possible that there are trade-offs between parameters even on the same item, especially when an exam seeks to measure mastery as opposed to seeking to discriminate students based on performance, such as seen with norm-based exams. Either way, these findings may be useful for course instructors and curriculum directors seeking to compare the performance of exams in different courses for the continual improvement of course content and structure.

CONCLUSIONS

This study sought to investigate exam item performance parameter distributions using exam data from first- and second-year medical school courses. The item Difficulty, Discrimination Index, and Point Biserial for 61 exams were recorded and assessed for normality graphically and numerically using R. The analyses revealed that exam item Difficulty had the most deviations from normality, Point Biserial had the least, and Discrimination index had an intermediate value. These findings facilitate further investigation using inferential statistics that rely on knowing the normality of a distribution. These results may also be useful to course instructors and curriculum directors at the College of Osteopathic Medicine seeking to make data-driven curricular decisions.

REFERENCES

1. National council on Measurement in Education http://www.ncme.org/ncme/NCME/Resource_Center/Glossary/NCME/Resource_Center/Glossary1.aspx?hkey=4bb87415-44dc-4088-9ed9-e8515326a061#anchorA Archived 2017-07-22 at the Wayback Machine
2. *Allen, M.J. (2004). Assessing Academic Programs in Higher Education. San Francisco: Jossey-Bass.*
3. Terry L, Price J. Metrics of OSUCOM Exam Question Statistics. Poster presented at OSU-CHS Research Days 2021; Tulsa, OK
4. Journal of Statistical Computation and SimulationVol. 81, No. 12, December 2011, 2141–2155 Comparisons of various types of normality tests. B.W. Yapa and C.H. Sim
5. Razali, Nornadiah; Wah, Yap Bee (2011). "Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests". Journal of Statistical Modeling and Analytics. 2 (1): 21–33.
6. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B, 57, 289–300. http://www.jstor.org/stable/2346101.
7. Sajjadi NB, Bixler KM, Price J. Medical Student Academic Performance Is Not Normally Distributed. Poster presented at OSU-CHS Research Days 2021; Tulsa, OK.