Spatial Ability Instrument Ceiling Effect and Implications

Mr. Benjamin James Call, Utah State University

Benjamin Call graduated with his Masters of Science degree in Mechanical Engineering (Aerospace Emphasis) in 2006 from Utah State University. After eight years with NAVAIR, he has returned to pursue a PhD in Engineering Education. He is funded by the Presidential Doctoral Research Fellowship. His research interests range from sophomore-level engineering curricula to spatial ability and creativity to student entrepreneurship.

Dr. Wade H Goodridge, Utah State University

Wade Goodridge, Assistant Professor in the Department of Engineering and Technology Education at Utah State University, has taught Solid Modeling, CAD, Introductory Electronics, Surveying, Statics, teaching and Learning, Assessment and Evaluation, and Introductory Engineering courses at Utah State University. Goodridge has been teaching for the Utah State College of Engineering for more than 15 years. He holds dual B.S degrees in industrial technology education and civil engineering from Utah State University, as well as an M.S. and Ph.D. in civil engineering from Utah State University. His research interests include spatial thinking/spatial ability at a course specific level in engineering, conceptual and procedural knowledge interplay in novice engineering students, and entrepreneurship.

Thayne L Sweeten Ph.D., Utah State University

Thayne Sweeten earned a Ph.D. in Medical Neurobiology and Indiana University – Purdue University at Indianapolis. He teaches biology courses for Utah State University and publishes research in pedagogy and the etiology of autism and related disorders focusing on immune mechanisms.

Spatial Ability Instrument Ceiling Effect and Implications

Abstract

This research paper investigates the potential existence of and implications for a ceiling effect observed in sophomore engineering students' spatial ability scores when using a common spatial ability instrument. Repeated use of the Purdue Spatial Visualization Test: Visualization of Rotations (PSVT:R) – shown herein over the course of two semesters – has revealed potential limitations when using the assessment with undergraduate engineering students during their sophomore year. The correlation between spatial ability and academic performance in engineering education has been thoroughly established. The PSVT:R and its revision are commonly used in academic spatial ability research. However, with the observed high average performance typical of engineering students on the PSVT:R, a ceiling effect may pose limitations to its utility. Sophomore engineering students in a Statics class - the first class in the Engineering Mechanics series - were each given the PSVT:R and Mental Cutting Test (MCT) assessments twice per semester. Results showing that the MCT may be more capable of differentiating student abilities, despite having a lower possible maximum score, are presented. Scores from similarly aged students in an Anatomy class are provided for comparison. The impact of ceiling effects for the education of high-performing populations, such as Engineering Mechanics students, will be discussed and actions for improvements in spatial ability measurement will be proposed. An argument is also put forth to understand how these tests relate to students' engineering capabilities.

Introduction

As we seek to identify means of predicting and tracking student success in engineering education, one of the key differentiators for predicting student success in engineering – spatial ability – is showing limits in its differentiability. In this paper, the term “spatial ability” refers to the measure of an individual’s spatial aptitude. Spatial ability has been repeatedly correlated with students’ success in engineering majors and professionals’ success as engineers[1-3]. Native spatial ability scores, as measured via occupational instruments (see below), have proven to be a valuable predictor for success in engineering. Indeed some research indicates their potential as more substantive than SAT math and verbal scores[4, 5]. Current work is now just beginning to investigate why and where spatial ability is so important within engineering curricula[6].

Spatial ability is typically measured with one of a host of instruments each originally occupationally designed[7, 8]. These instruments target or situate themselves around a construct of spatial ability that is to be assessed. Some of the most popular[8] and well known instruments are the Purdue Spatial Visualization Test: Visualization of Rotations (PSVT:R)[9] and the Mental Cutting Test (MCT)[10]. This study has chosen to focus on the PSVT:R and MCT instruments and their application into the assessment of sophomore level engineering students spatial ability. These instruments were chosen due to previous work that was conducted in an engineering graphics course for mechanical engineers where data analysis pointed to the potential existence

of a ceiling effect with engineering students[11]. While the spatial ability research community generally treats the MCT and PSVT:R as if they both measure spatial ability equally, a review of the instruments makes it apparent that the PSVT:R is focused on a rotations facet of spatial ability. The MCT also includes the rotational facet, but adds in proportionality as a facet as well. Thus it may be expected that the MCT scores reveal it to be a more challenging test of spatial ability. It is anticipated that rotations of vectors – e.g. for trigonometric functions – in Statics may relate to the rotations facet, while the length and relative magnitude of force vectors in Statics may relate to the proportionality facet.

Unfortunately, existing metrics such as the PSVT:R may not have a sufficient range of questions and depth (i.e. difficulty) within the existing question set in order to appropriately assess engineering students' spatial ability improvements. At least not within a measurement construct using pre- and post-testing to bracket spatial instructional interventions. Literature indicates students entering and pursuing engineering degrees often have higher native spatial ability, but some work also points to potential ceiling effects that may exist using certain spatial metrics[11]. There are a number of relatively young engineering students achieving top scores on spatial ability instruments. This creates a dilemma in that we are unable to measure how much they continue to improve over the course of their academic career when they may natively perform at an instrument's top score threshold. This also prevents researchers from identifying which portions of the curriculum are highly correlated with spatial ability improvement for those students. In educational settings, quasi-experimental statistics are generally utilized due to the cost and complexity of treating each application of the treatment (i.e. each application course instruction – once per course section) as a single data point, a technique advocated in fully experimental statistical analysis[12]. Moving forward, the researchers hope to collect data from sufficient separate instructional deliveries to perform a full experimentally-designed research plan with corresponding statistics; however, the presence of a ceiling effect would also impact the average score measurement that each experimental unit (i.e. section) would receive.

The research question for this paper is "Is there a tendency for spatial ability instrument scores in engineering sophomore students to indicate a ceiling effect that will prevent accurate statistical measures from being achieved?" The hypothesis, simply based on average scores for PSVT:R and MCT, is that the PSVT:R is more likely to see a ceiling effect within engineering populations since the average score within the class is so close to the maximum score possible. The goal of this work is to evaluate specific spatial instruments for their application into spatial ability research in sophomore engineering courses as well as to illuminate a potential limitation that researchers in engineering education should be aware of.

Literature Review

The impact and predictive qualities of 3-D spatial ability regarding engineering education have been studied for some time, and various instruments have proven useful for measuring spatial ability in these studies[8]. The PSVT:R and MCT, mentioned above, are very prominent in that list. In fact, Sorby and Baartmans[8] primarily used the PSVT:R for recruitment and assessment in their research and positively impacting interventions. When using data to create predictions or to

identify the impact of certain factors, multiple regression/correlational analysis (MRC) is often the statistical approach of choice[13]. Of course, basic MRC approaches require data that are "well-behaved", inferring requirements like normality.

Unfortunately, real-world data is often not normal – particularly real-world, academic, standardized test data[14-17]. In the data collected for this study, one subset of participants seemed to score especially high on the PSVT:R. This led to the observation by researchers of a potential ceiling effect. "Ceiling effect" is the term used to describe the situation when many participants obtain a maximum score[18]. This is a type of censoring – where censored data occurs when there is a lower bound, a; an upper bound, b; or a situation with bounds a and b[19]. Kruskal and Tanur[19] point out that censored data will result in sample means and standard deviations that are poor predictors of the population mean and standard deviation. Since nearly every basic statistical measure relies upon good estimates for the mean and standard deviation, further statistical analyses may suffer when based on censored data. To help identify a lack of well-behaved data, checks for normality will be performed. Ho and Yu[14] looked into the impact of ceiling effect on test scores and recommend skewness and kurtosis as factors for identifying normality in standardized test scores. Of note with regard to standardized test scores, "easy" test scores will likely have negative skew, and symmetric raw score distributions will have kurtosis < 3 (i.e. platykurtic)[15, 16]. Censored data are susceptible to failing checks for normality, and violations of normality may also be easily identified through visual inspection. Skew, kurtosis, and visual inspections of the data will be shown below.

Linear regression techniques hold promise in identifying key factors in students' native spatial ability and identifying any spatial ability improvement that may occur during the semester, while controlling for a variety of factors, demographic and otherwise. In particular, lagged regression (aka residual gain) is appropriate, when provided with the opportunity to collect pre- and post-intervention measurements. Lagged regression presents an interesting case, if a ceiling effect is present, as the dependent variable (i.e. the post-intervention measurement) and at least one of the independent variables (i.e. the pre-intervention measurement) may be restricted by the ceiling effect. Austin & Brunner[20] and Austin & Hoch[21] point out that most research into the impact of ceiling effects or censored data deal with the impact of censoring the dependent variable. Their research identifies the possibility that a censored independent variable may increase the likelihood of a Type I error (i.e. identifying a change that is not really there) in data analysis. Spatial ability scores constitute the independent variable in studies such as those treated in this report.

Various analytical techniques are available for dealing with ceiling effects, including Tobit models for censored dependent variables[14, 20], or a simple discard method and the maximum likelihood method for independent variables[20, 21]. While these may be effective at minimizing the impact of the ceiling effect - like the maximum likelihood method does for an increase in Type 1 error - the statistical power is reduced by the technique, and it is recommended that data be collected that is not subject to ceiling effects for best results[20].

Additionally, there are other limitations imposed by censored data. Vogt[18] provides a fitting example of the impact of ceiling effect:

> For example, suppose a group of statistics professors wants to see whether a new method of teaching increases knowledge of elementary statistics. They give students in their classes a test, try the new method, and then give the students another version of the same test to see whether their scores went up. If one of the professors had students who knew a lot of statistics already, and scored at or near 100% on the first test, she could not tell whether the new method was effective in her class. The scores of her students were so high (at the ceiling) that they could hardly go up, even if the students learned a great deal using the new method.[18]

This scenario is essentially what was identified in this study for a subset of the participants. Such a condition is of great concern to educators who are seeking to develop and validate new methods of instruction as well as to those looking to use existing instruments with populations that may have native abilities above what existing instruments were originally designed to test.

Methods

Participants were recruited from an Engineering Statics class and an Anatomy class, just as in work by Wood, Goodridge, Call, & Sweeten[6]. Previous investigations with the Statics class had revealed that scores for engineering students were quite high on the PSVT:R and MCT. Additional data collected and presented in previous work[11] also identified higher scores on the PSVT:R by freshmen mechanical engineering students. An Anatomy class was incorporated into this study to serve as a non-engineering control group. Data were collected over the course of two semesters, Fall 2014 - Spring 2015 for the Anatomy and Statics courses at a western research university. The racial demographic of the study participants is relatively homogeneous – predominantly Caucasian – with some deviation in gender. The Statics courses had a 3:19 female-to-male ratio while the Anatomy courses had an 8:3 female-to-male ratio. Participants took the PSVT:R and MCT assessments at the beginning and end of the semester in which they were enrolled. The time in between pre- and post-curriculum assessments was approximately three months, well over the one month minimum recommended to avoid the risk of improved performance by repeating the same test[22]. The advantage of repeating the same test is to gauge improvement in spatial ability. While individual participants may not receive much direct benefit – other than having evidence of their personal improvement in a cognitive capability – the ability to determine a difference score, representing improvement, is advantageous to researchers of spatial ability and educators who desire to understand the impact of their curricula.

The Anatomy class served as a control group to provide insight on whether testing repetition or age-related maturation were responsible for changes in Statics students' performance. 198 Anatomy students and 154 Statics students participated during these two semesters. An opt-out design was given where students were given the opportunity to drop out of the study at any time. Appropriate Institutional Review Board (IRB) protocols were developed and implemented.

Identification of ceiling effects in this paper will largely be demonstrated through visual graphics, with quantitative measures, skewness and kurtosis, provided to reference against the guidelines in Ho and Yu[14]. Statistics will be done separately for the Statics and Anatomy classes to identify when the impact of ceiling effects should be considered.

Results

It is noted that the engineering student participants (in the Statics class) scored higher, on average, in every spatial ability assessment, than their non-engineering peers. This is not unexpected as seen in previous literature[5]. However, and of particular importance to this work, one can observe that the difference is sufficient that ceiling effects are regularly more visible for the engineering students than for the non-engineering students. Results on the two instruments are reported in the figures in this section where the top bar charts (Plots A and C in both figures) reflect Anatomy course scores and the bottom bar charts (Plots B and D in both figures) reflect Statics course scores. Data is initially presented with regards to the MCT instruments application in a pre- and post-testing format for both classes and then data is presented for the PSVT:R in the same fashion. Kurtosis and skewness will be discussed as relevant descriptive statistical data for each bar chart and comparisons can then easily be made between the Anatomy and Statics pre and post-performance on both instruments. A typical bell curve centered on the mean has been provided to aid visual confirmation of data normality.

<u>MCT Results</u>

Based on the pre-MCT results, the Anatomy course (Fig. 1, Plot A) had kurtosis = 0.439 and skewness = 0.658. Based the pre-MCT results, the Statics course (Fig. 1, Plot B) had kurtosis = -0.724 and skewness = 0.215. Visually, normality is not perfectly obtained, nor is it grossly violated.

Based on the post-MCT results, the Anatomy course (Fig. 1, Plot C) had kurtosis = 1.378 and skewness = 0.955. Based on the post-MCT results, the Statics course (Fig. 1, Plot D) had kurtosis = -0.747 and skewness = 0.020. Visually, normality is poorer than for the pre-MCT data, and the data from the Statics class appears to indicate a ceiling effect at the maximum MCT score of 25 points – see Fig. 1, Plot D (note that more participants in Statics scored 25 than scored 24, and nearly as many scored 25 as scored 23, 21, 19, 18 and 15). It is easy to imagine a better fit on the right side of the curve if the count scoring 25 points was distributed more normally throughout higher scores.

<u>PSVT:R Results</u>

Based on the pre-PSVT:R results, the Anatomy course (Fig. 2, Plot A) had kurtosis = -0.433 and skewness = -0.008. Based on the pre-PSVT:R results, the Statics course (Fig. 2, Plot B) had kurtosis = 1.417 and skewness = -0.882. Visually, normality is not perfectly obtained, nor is it grossly violated for the Anatomy scores, but it is for the Statics scores, with a clustering of perfect and near-perfect scores (30 and 29, respectively – seen in Fig. 2, plot B).

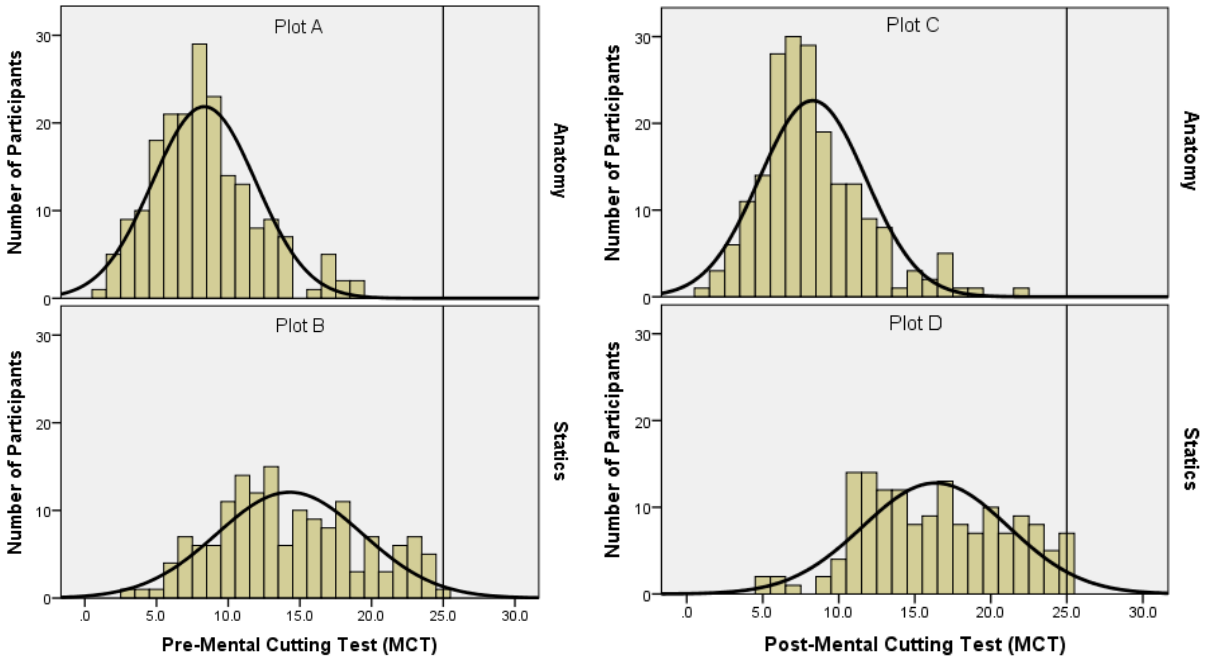

Figure 1. Score Distribution Plots: Mental Cutting Test

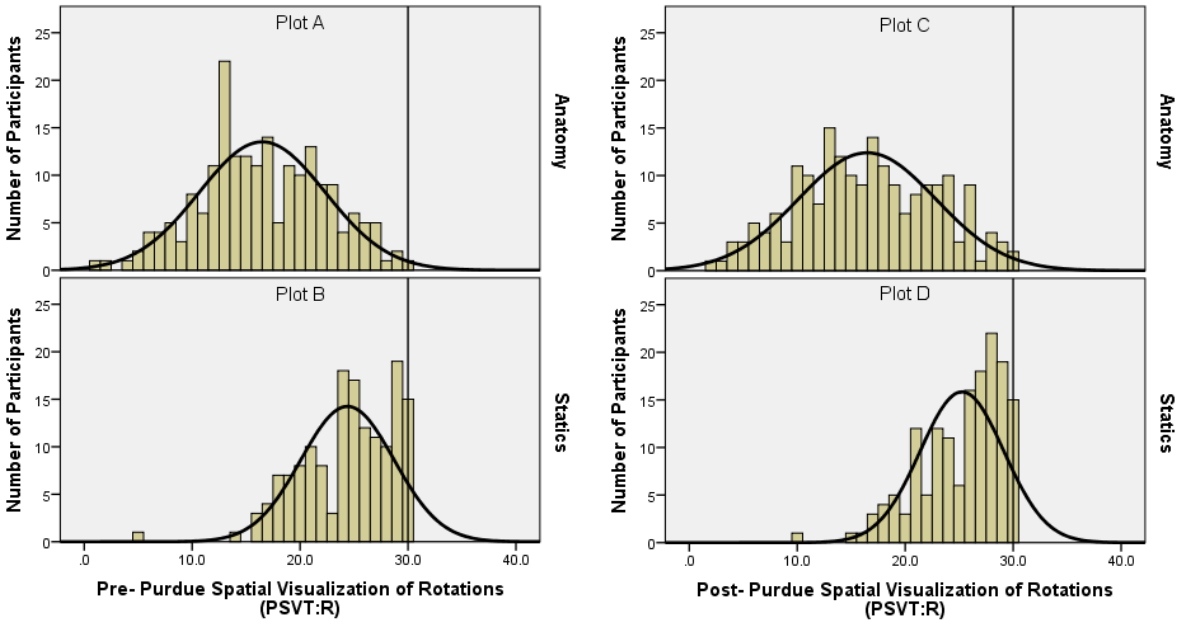

Figure 2. Score Distribution Plots: Purdue Spatial Visualization: Visualization of Rotations

Based on the post-PSVT:R results, the Anatomy course (Fig. 2, Plot C) had kurtosis = -0.689 and skewness = 0.027. Based on the post-PSVT:R results, the Statics course (Fig. 2, Plot D) had

kurtosis = 0.709 and skewness = -0.958. Visually, the normality approximation is potentially worse for the post-test Anatomy scores than for the pre-test scores (Fig. 2, Plot C vs Plot A). As for the Statics course, with a large increase of near-perfect scores (27 and 28, respectively - seen in Fig. 2, Plot D), normality is clearly not achieved. The abundance of high scores on the PSVT:R may indicate that it is an easier test for the average engineering student. Additionally, it may be worth considering, that for $\alpha = 95\%$ (or $p \leq 0.05$), a typical level used in most statistical studies, a score of 95% or above is essentially indistinguishable from a score of 100%. This would result in treating scores of 29 (96.7%) or 30 (100%), and even potentially 28 (93.3%), as a grouped maximum score allowed by the ceiling effect. It is difficult to state with confidence that participants achieving a score of 29 are significantly differentiable from students who would have scored 30, in terms of spatial ability.

Skewness is positive for most of the curves shown above. However, the pre- and post- PSVT:R scores for the engineering students have negative skew, which further promotes the idea that the PSVT:R is an easier test for engineering students[15, 16]. It is noted that the kurtosis is less than 3 for every set of data above, matching the predictions of Lord[15] and Cook[16] for standardized test scores. Overall, the MCT, exhibiting less of a ceiling effect, may be the most useful at differentiating between engineering students' spatial ability levels and at tracking increases in spatial ability as students progress through engineering curricula or through spatial ability interventions.

Conclusion

In past uses of the PSVT:R and MCT tests – either with a broader sample of the population that was not so engineer-dominant, or in use to determine if a test taker was inclined to certain occupations due to a high spatial ability – the impact of ceiling effects may not have been an issue for spatial ability assessments. Unfortunately for engineering education, due to ceiling effects, it is impossible to separate the spatial ability for rotations among students who all score highly – e.g. 30 (or even, significantly, those who score 29) on the PSVT:R. This means that the progress of those higher scoring students cannot be tracked as we seek to understand how their spatial ability improves through any intervention that may have a positive impact on spatial ability. This has been shown to be an issue present in sophomore-level engineering students for the PSVT:R, and the issue may also be slightly present in MCT scores after a single semester of coursework. The ceiling effect issue may be even more limiting as students progress through the engineering curriculum, gaining spatial ability along the way. More investigation in higher-level coursework (e.g. junior-level courses) is needed to determine if that is the case. It is also recommended that other engineering courses not associated with the mechanics series of classes (which includes Statics) be investigated to gain insight into the broader population of engineering students.

More insight into the statistical impact of censored data when present in the dependent variable and an independent variable may be desirable to the academy, and may be aided through simulation work. But given the obvious ceiling effects displayed in the figures above, the more pressing issue is to locate, develop, or define spatial ability instruments that provide a higher

range of possible values before a potential ceiling effect can censor the data - particularly for engineering students. Not only would this improve the normality of the data and decrease the need for additional analytical processes that will reduce the statistical power, but it would also allow for improved understanding of student learning and improved assessment of curriculum impact on student abilities.

Funding and Acknowledgements

Benjamin Call is funded by Utah State University's Presidential Doctoral Research Fellowship. We would like to thank all of the students who participated in the study.

References

1. Halpern, D. F., & Collaer, M. L. (2005). *The Cambridge Handbook of Visuospatial Thinking*. Cambridge: Cambridge University Press.
2. Sorby, S., Casey, B., Veurink, N., & Dulaney, A. (2013). The role of spatial training in improving spatial and calculus performance in engineering students. *Learning & Individual Differences*, *26*, 20-29.
3. Sorby, S., & Veurink, N. L. (2012). *Spatial ability among minority and international engineering students*. Paper presented at the 2012 Annual Conference of the American Society of Engineering Education, San Antonio, Texas.
4. Shea, D. L., Lubinski, D., & Benbow, C. P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: a 20-year longitudinal study. *Journal of Educational Psychology*, *93*(3), 604-614.
5. Webb, R. M., Lubinski, D., & Bendbow, C. P. (2007). Spatial Ability: A Neglected Dimension in Talent Searches for Intellectually Precocious Youth. *Journal of Educational Psychology*, *99*(2), 397-420.
6. Wood, S. D., Goodridge, W. H., Call, B. J., Sweeten, T. L. (2016, June). *Preliminary analysis of spatial ability improvement within an engineering mechanics course: Statics*. Paper accepted to be presented at 2016 ASEE Annual Conference, New Orleans, LA.
7. Hegarty, M. (2010). Components of Spatial Intelligence. *Psychology of Learning and Motivation*, *52*, 265-297. doi:10.1016/s0079-7421(10)52007-3
8. Sorby, S. A., & Baartmans, B. J. (2000). The Development and Assessment of a Course for Enhancing the 3-D Spatial Visualization Skills of First Year Engineering Students. *Journal of Engineering Education*, *89*(3), 301-307. doi:10.1002/j.2168-9830.2000.tb00529.x
9. Guay, R. (1976). Purdue Spatial Visualization Test: Purdue University.
10. CEEB. (1939). CEEB Special aptitude test in spatial rotations. USA.
11. Goodridge, W. H., & Greenhalgh, S. D., & Lawanto, O., & Stewardson, G. A. (2012, June), *Measuring the Differences in Spatial Ability Between a Face-to-Face and a Synchronous Distance Education Undergraduate Engineering Graphics Course* Paper presented at 2012 ASEE Annual Conference, San Antonio, Texas. https://peer.asee.org/21679
12. Oehlert, G. W. (2000). *A First Course in Design and Analysis of Experiments*. New York: W. H. Freeman and Company.
13. Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.). New York: Routledge, Taylor & Francis Group, LLC.
14. Ho, A. D., & Yu, C. C. (2015). Descriptive Statistics for Modern Test Score Distributions: Skewness, Kurtosis, Discreteness, and Ceiling Effects. *Educational and Psychological Measurement*, *75*(3), 365-388. doi:10.1177/0013164414548576
15. Lord, F. M. (1955). A survey of observed test-score distributions with respect to skewness and kurtosis. *Educational and Psychological Measurement*, *15*, 383-389.

16. Cook, D. L. (1959). A replication of Lord's study on skewness and kurtosis of observed test-score distributions. *Educational and Psychological Measurement*, 19, 81-87.
17. Micerri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166.
18. Vogt, W. P. (2005). *Dictionary of Statistics & Methodology: A nontechnical guide for the social sciences* (3rd ed.). Thousand Oaks, CA: SAGE Publications, Inc.
19. Kruskal, W. H., & Tanur, J. M. (1978). *International Encyclopedia of Statistics* (Vol. 2). New York: The Free Press, A Division of Macmillan Publishing Co., Inc.
20. Austin, P. C., & Brunner, L. J. (2003). Type I Error Inflation in the Presence of a Ceiling Effect. *The American Statistician*, *57*(2), 97-104. doi:10.1198/0003130031450
21. Austin, P. C., & Hoch, J. S. (2004). Estimating linear regression models in the presence of a censored independent variable. *Statistics in Medicine*, *23*(3), 411-429.
22. Stangor, C., *Research Methods for the Behavioral Sciences*. 2003, Boston, MA: Houghton Mifflin Harcourt.