THE MULTIPLE CHOICE QUESTION: ASSESSING STATISTICAL LITERACY AND CRITICAL THINKING IN THE INTRODUCTORY STATISTICS COURSE AT THE COLLEGE LEVEL

Rossi A. Hassad[1] and Gerald Iacullo
[1]Mercy College, New York, USA
rhassad@mercy.edu

Statistical literacy is recognized by most disciplines as a necessary competency for success in college and the workplace. Toward this end, there is much emphasis and debate on assessment approaches, including whether the multiple choice question (MCQ) format is appropriate for assessing statistical literacy, which encompasses critical thinking. This study analyzed data from 4 different introductory college-level statistics classes, which used the same MCQ exam designed to assess statistical literacy. Psychometric analysis was performed. The results suggest that MCQ questions can be effective in assessing statistical literacy if they facilitate multilogical thinking, or connected understanding, that is, using multiple concepts simultaneously in problem-solving, including conceptual hierarchies or nested concepts.

INTRODUCTION

As Resnick noted, "*we get what we assess, and if we don't assess it, we won't get it*" (quoted in Wiggins, 1992, p. 152). This quote underlines the importance of being clear about what we want students to learn, and assessing what we value most (Chance, 2002). Ideally, such learning outcomes should be informed by the needs of academic programs and the workforce. However, there is a long tradition of the introductory statistics course being managed by mathematics departments, with a standardized curriculum focused on discrete and compartmentalized knowledge and skills, rather than integration and conceptual understanding (Moore, 1997; Hedges & Harkness, 2017). This pedagogical approach is consistent with the behaviorist philosophy, which is geared primarily toward determining how much information students can memorize and recall (surface knowledge), and is generally assessed by drill and practice exercises (Cobb, 1992; Garfield, delMas, & Zieffler, 2010).

A popular assessment format for statistics and mathematics is the multiple choice question (MCQ), given the ease and efficiency in scoring, and the scope to assess a broad range of course content (Davies & Marriott, 2010), facilitated by dedicated scoring software (such as Scantron), which also provides psychometric analysis. Nonetheless, MCQ exams for introductory statistics courses remain largely focused on lower order thinking skills, primarily recognizing and recalling information (delMas, Garfield, Ooms, & Chance, 2007).

THE MCQ FORMAT: STATISTICAL LITERACY AND CRITICAL THINKING

The multiple choice question (MCQ) format as an assessment approach has been the subject of much debate over the years. For example, the National Council of Teachers of Mathematics (NCTM, 1991) noted that MCQ exams can have a negative impact on student learning "*since student scores are generated solely on the basis of right and wrong answers with no consideration or credit given to students' strategies*" (p.8). And, Garfield (2003) observed that:

> "*traditional test questions involving statistical content often lack appropriate context and tend to focus on accuracy of statistical computations, correct application of formulas, or correctness of graphs and charts ……. and therefore provide only limited information about students' statistical reasoning processes and their ability to construct or interpret statistical arguments*" (p. 24).

Furthermore, according to delMas et al. (2007), exam questions are not appropriate for assessing statistical literacy if they focus on procedures and definitions, rather than conceptual understanding.

Statistical literacy is typically defined as: "*People's ability to interpret and critically evaluate statistical information and data-based arguments appearing in diverse media channels, and their ability to discuss their opinions regarding such statistical information*" (Gal, 2000 as cited in Rumsey, 2002, p. 2). Underpinning statistical literacy is critical thinking (Aizikovitsh-Udi

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

& Kuntze, 2014), which is recognized by most disciplines as a necessary competency to better prepare students and graduates for effective and engaged citizenship (Engel, 2017). Bloom's Taxonomy (Bloom, 1956) characterizes critical thinking as a higher-order skill set. It is typically defined as "*the process of gathering information, analyzing it in different ways, and evaluating it for the purposes of gaining understanding, solving a problem, or making a decision*" (Carter, Bishop, & Kravits, 2007).

Assessment items for statistical literacy including MCQs should assess students' ability to make connections and explain the interrelationships among statistical concepts, as well as create representations of statistical data (Garfield & Ben-Zvi, 2008). Proponents of this approach argue that what matters, is not the format but the construction of the question (Shete, Kausar, Lakhkar, & Khan, 2015; Khan, Danish, Awan, & Anwar, 2013), with attention to the level of reasoning or cognitive skills required (Garfield, 2003). Some educators have noted that exam questions addressing confounding, variability, and multivariate thinking, are effective in engendering critical thinking and problem-solving skills (GAISE, 2016; Schield, 2010; Pfannkuch, & Wild, 2004). The quality of MCQs is commonly assessed using classical items analysis (Kehoe, 1995).

ITEM ANALYSIS

Item analysis generally refers to a process which uses a set of statistical techniques to examine students' responses to individual test items or questions on an exam, in order to assess the quality of each item, and the test in general (Kehoe, 1995). Item analysis is particularly useful for improving or eliminating ambiguous or misleading items, or those that lack meaningful discriminant value (Gajjar, Sharma, Kumar, & Rana, 2014). In other words, this process helps to assess and improve the reliability and validity of the test. There are two recognized statistical frameworks which guide item analysis; classical test theory (CTT) and item response theory (IRT). CTT focuses on aggregate test level performance whereas IRT addresses "*the relationship between ability (or trait) and performance for each individual item*" (Reid, Kolakowsky-Hayner, Lewis, & Armstrong, 2007, p. 179), which makes IRT the generally preferred approach for item analysis.

Item analysis provides measures of item difficulty, discriminant value, and reliability. It is recognized that the single best measure of the effectiveness of an item is its ability to separate students who vary in their degree of knowledge of the material tested. A common measure of item discrimination is the point biserial correlation coefficient, which indicates the strength and direction of the relationship between performance on an item (dichotomous variable; correct or not) and the total score (continuous variable) on the test or exam.

OBJECTIVE, RATIONALE, AND THEORETICAL FRAMEWORK

The objective of this study is to present and analyze a selected multiple choice question (MCQ), which consistently demonstrated high discriminant value in differentiating between students who mastered the introductory statistics material and those who did not. The course is designed to foster statistical literacy, and it is posited that this particular question is measuring critical thinking, which is the core component of statistical literacy. This study offers a realistic model of an effective MCQ question for assessing statistical literacy (including critical thinking). Students' perspective on the nature of the question (mathematical versus conceptual) was also ascertained in order to triangulate the results. In general, this study is guided by Bloom's Taxonomy (1956), specifically higher-order or critical thinking, as well as the Guidelines for Assessment and Instruction in Statistics Education report (GAISE, 2016), and the constructivist philosophy of teaching and learning (Hassad, 2011).

METHODOLOGY

This study utilized mostly secondary data from the psychometric analysis of a 20-item MCQ exam, which was administered to four different groups of students over four semesters (Table 2). The same exam was used each semester, and was intended to assess statistical literacy (and critical thinking) in an introductory statistics course for college students in psychology, media, culture, nutrition, and the humanities. Specifically, the statistics course was designed and administered in accordance with the Guidelines for Assessment and Instruction in Statistics

Education (GAISE, 2016), and the first exam, which data were analyzed for this study, covers all material up to and including the one-sample t-test.

The course emphasizes concepts over calculations, with a focus on telling the story of the data. The material covers descriptive and inferential statistics, including measures of central tendency, measures of variability, sampling, the normal distribution, as well as t-test, ANOVA, correlation, regression, and chi-square. Study designs (including association, causation, confounding, and interaction) are also addressed, and students are required to complete a small-group project.

Item Analysis Reports (generated by the Scantron software)

The item analysis reports were reviewed specifically to determine which of the 20 questions consistently demonstrated high discriminant value, that is, meaningfully differentiated between students who mastered the material and those who did not. The "*extreme group method*" (Kline, 2005); that is, a comparison of the upper and lower 27% of the distribution was used, and item discrimination was quantified using the point biserial correlation coefficient. Positive values are desirable because they indicate that a student who performed well on the exam (as a whole) also answered this question correctly. Point biserial correlations between 0.30 and 0.70 should be the goal (Allen & Yen, 1979), and very high coefficients are counter-productive, as they indicate redundancy among items in the test. Also extracted from each report, was the Kuder-Richardson Formula (KR-20), a measure of internal consistency or reliability for binary or dichotomous variables. It measures the extent to which the exam is composed of items measuring a single subject area or underlying ability or trait such as quantitative reasoning or statistical literacy.

Primary Data

Another group of statistics students (N = 57, Fall 2017) answered the selected MCQ (Table 1), and also rated it on a scale of 1 to 7, with 1 being more mathematical (requiring calculations) and 7 being more conceptual (requiring reasoning). This information was used to triangulate the results.

RESULTS

This study identified a single MCQ considered effective for assessing statistical literacy and critical thinking (Table 1). The variability in the percent of correct response to this question for each semester (Table 2) shows that this item meaningfully discriminated between those students who did well (upper 27%) and those who did not (lower 27%). This is also reflected in the point biserial correlation coefficients, which are high, indicating that top performing students were more likely than low performing students to get this question correct.

Table 1: *The Selected MCQ*

With reference to estimation in statistics, the larger the sample size:

A. the wider the confidence interval
B. the lower the degree of precision of the confidence interval
C. a and b
D. the smaller the standard error
E. a and d

Table 2: *Data Extracted from the Item Analysis of the Selected MCQ*

Semester	Class Size (N)	(KR-20) Reliability of the Exam	Correct Group Responses			Point Biserial Correlation ($p < .01$)
			Total %	Upper 27%	Lower 27%	
Fall 2015	38	0.68	66	90	20	0.61
Spring 2016	56	0.72	73	93	33	0.58
Fall 2016	56	0.70	73	100	40	0.56
Spring 2017	53	0.76	76	100	50	0.49

As evidenced in Table 2, the exam demonstrated acceptable levels of internal consistency or reliability over four semesters, suggesting that the test is measuring a single underlying concept, intended to be statistical literacy and reasoning (including critical thinking), albeit not specifically validated in this study.

As noted, primary data were obtained from another group of students (Table 3) who rated the selected MCQ as more conceptual (requiring reasoning) rather than mathematical (requiring calculations), and there was no statistically significant difference between those who answered it correctly and those who did not.

Table 3: *Comparison of Students' Rating of the Selected MCQ (N = 57)*

	N	Mean*	SD
Correct Response	27	5.74	0.86
Incorrect Response	30	6.03	1.10
Overall	57	5.89	0.99

*t (55) = 1.11, p = .27

We observed that the point biserial correlation coefficients for the 20 items ranged from .13 to .69, with 18 items being approximately at least .3 (consistently). And for two items which had a coefficient of .13, they both had higher values (greater than .4) for other semesters. This pattern suggests that the questions on this test are meaningfully correlated with the intended underlying construct (statistical literacy, including critical thinking).

DISCUSSION

Although this study presents and discusses a single MCQ (multiple choice question), it is somewhat original, in that, while there is an abundance of information on how to write effective multiple choice questions to assess statistical literacy, there is a dearth of actual examples or models of questions based on sound item analysis. Furthermore, this question addresses core statistical concepts, including confidence interval (delMas et al., 2007), and standard error (Sabbag & Zieffler, 2015), which have been classified as threshold concepts (Dunne, Low, & Ardington, 2003), and challenging for students. Moreover, one of the instruments for which item analysis data are available; the GOALS-2 (Goals and Outcomes Associated with Learning Statistics), has been shown to possess low discriminant value (almost zero) for an item assessing standard error and inference about the mean (Sabbag & Zieffler, 2015). Therefore, the question from the current study can serve as a model to design items to fill this gap.

The selected question was taken from an exam, which was designed to assess statistical literacy (including critical thinking), and the levels of internal consistency (ranging from .68 to .76 over 4 semesters) reported herein, could support that a single underlying construct is being measured, albeit criterion validity of the test was not determined. Nonetheless, content validity was established. Also, a similar group of students rated the question as more conceptual (that is, requiring reasoning) rather than mathematical (requiring calculations). The totality of this evidence, could suggest that this test is measuring statistical literacy, and hence critical thinking about data. Notably, recognized instruments such as CAOS and SRA, used for assessing statistical literacy and reasoning are based primarily on content validity and internal consistency (Sabbag & Zieffler, 2015).

How does this Selected MCQ (Table 1) Assess Statistical Literacy and Critical Thinking?

The correct answer is D (Table 1), and this question requires connected understanding, that is, reasoning with concepts and procedures, in particular, recognizing their interrelationships. Of course, a student can guess it correctly. The major concepts involved are standard error (SE) of the sample mean ($\bar{x}$), and confidence interval (CI) for a population mean based on the normal distribution (as per the course). While this MCQ requires students to identify these concepts and their formulas or representations, calculations for the standard error of the sample mean (SE =

SD/√n), and confidence interval [CI = $\bar{x} \pm z$ (SE)] are not required. Indeed, a comparable group of students rated this question as more conceptual (requiring reasoning) rather than mathematical (Table 3). A key relationship that needs to be understood and applied is that standard error is inversely proportional to the sample size.

Furthermore, given that the response options are all seemingly plausible (with familiar and related terminology), students should recognize that each option needs to be evaluated. This requires them to think about the various connections and relationships between sample size, standard error, and the confidence interval; and this involves quantitative reasoning and critical thinking. In summary, this question requires multilogical thinking or connected understanding, that is, using multiple concepts simultaneously in problem-solving (Montgomery, 1998), including conceptual hierarchies or nested concepts.

CONCLUSION AND IMPLICATIONS

This study reinforces that classical item analysis is helpful in identifying effective multiple choice questions for assessing statistical literacy and critical thinking. The MCQ format for assessment is proliferating in academia at a time when the statistics education community is focused on modifying the curriculum to address and assess statistical literacy and critical thinking. Therefore, it is imperative that we develop reliable and valid tests, and hence questions with acceptable levels of discriminant value. This will require a multifaceted and integrated approach, encompassing instructors, academic institutions, and professional bodies, if we are to change the culture of thinking about and using MCQs. Attention to the following could be helpful, in this regard.

1. Professional development programs should focus more on psychometric analysis.
2. The statistics education community, professional societies and bodies, as well as journal and conference editors should take a firm stance in support of a more evidence-based approach to test development.
3. Large scale psychometric studies with attention to criterion validity are required.
4. Further studies should include distractor analysis.

REFERENCES

Aizikovitsh-Udi, E. & Kuntze, S. (2014). Critical thinking as an impact factor on statistical literacy. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Proceedings of the Ninth International Conference on Teaching Statistics: Sustainability in Statistics Education.* Voorburg, NL: ISI.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, handbook I: The cognitive domain* (Vol. 19). New York: David McKay Co Inc.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, *10*(3).

Carter, C., Bishop, J., & Kravits, S. (2007). *Keys to college studying: Becoming an active listener*. Upper Saddle River, NJ: Prentice Hall.

Cobb, G.W. (1992). Teaching statistics. In L.A. Steen (Ed.), *Heeding the call for change: Suggestions for curricular action* (pp. 3-43). Washington, D.C.: Mathematical Association of America.

Davies, N., & Marriott, J. (2010). Assessment and feedback in statistics. In P. Bidgood, N. Hunt, & F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 3–19). Chichester, UK: Wiley.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28-58.

Dunne, T., Low, T., & Ardington, C. (2003). Exploring threshold concepts in basic statistics, using the Internet. In *International Association for Statistical Education, Statistics & the Internet*.

Berlin, Germany.
Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal, 16*(1), 44–49.
GAISE (2016). Guidelines for assessment and instruction in statistics education. *College report.* Alexandria, VA: American Statistical Association.
Gajjar, S., Sharma, R., Kumar, P., & Rana, M. (2014). Item and test analysis to identify quality multiple choice questions (MCQs) from an assessment of medical students of Ahmedabad, Gujarat. *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine, 39*(1), 17.
Garfield, J. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*(1), 22–38.
Garfield, J.B., & Ben-Zvi, D. (2008). The discipline of statistics education. In J. B. Garfield, & D. Ben-Zvi (Eds.), *Developing students' statistical reasoning: Connecting research and teaching practice* (pp. 3-19). Springer.
Garfield, J., delMas, R., & Zieffler, A. (2010). Assessing important learning outcomes in introductory tertiary statistics courses. In P. Bidgood, N. Hunt, & F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 75–86). Chichester, UK: John Wiley & Sons.
Hedges, S., & Harkness, S. S. (2017). Is GAISE evident? College students' perceptions of statistics classes as "Almost not math." *Statistics Education Research Journal, 16*(1), 337-35.
Hassad, Rossi A. (2011). Constructivist and Behaviorist approaches: Development and initial evaluation of a teaching practice scale for introductory statistics at the college level. *Numeracy, 4*(2).
Kehoe, J. (1995). Basic item analysis for multiple-choice tests. *Practical Assessment, Research & Evaluation, 4*(10).
Khan, H. F., Danish, K. F., Awan, A. S., & Anwar, M. (2013). Identification of technical item flaws leads to improvement of the quality of single best Multiple Choice Questions. *Pakistan journal of medical sciences, 29*(3), 715.
Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks: Sage Publications.
Montgomery, D. (1998). *Reversing lower attainment: Developmental curriculum strategies for overcoming disaffection and underachievement*. London: David Fulton Publishers.
Moore, D.S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review, 65*(2), 123-165.
National Council of Teachers of Mathematics (NCTM). (1991). *Professional Standards for Teaching Mathematics.* Reston, VA: NCTM.
Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi, & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 17-46). Dordrecht, Netherlands: Kluwer Academic Publishers.
Reid, C. A., Kolakowsky-Hayner, S. A., Lewis, A. N., & Armstrong, A. J. (2007). Modern psychometric methodology: Applications of item response theory. *Rehabilitation Counseling Bulletin, 50,* 177–188.
Rumsey, D. J., (2002). Statistical literacy as a goal for introductory statistics courses, *Journal of Statistics Education, 10*(3).
Sabbag, A. G., & Zieffler, A. (2015). Assessing learning outcomes: An analysis of the GOALS-2 Instrument. *Statistics Education Research Journal, 14*, 93–116.
Schield, M. (2010). Assessing statistical literacy: Take CARE. In P. Bidgood, N. Hunt, & F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 133-152). Chichester, UK: John Wiley & Sons.
Shete, N., Kausar, A., Lakhkar, K., & Khan, S. T. (2015). Item analysis: An evaluation of multiple choice questions in Physiology examination. *Journal of Contemporary Medical Education, 3,* 106-109.
Wiggins, G. (1992), Toward assessment worthy of the Liberal Arts. In L. Steen (Ed). *Heeding the call for change, MAA Notes No. 22*. Washington: Mathematical Association of America.