A MEASURE OF BASIC MATH SKILLS FOR USE WITH UNDERGRADUATE STATISTICS STUDENTS: THE MACS

LAURA RABIN
Brooklyn College of the City University of New York
Lrabin@brooklyn.cuny.edu

LAUREN FINK
Brooklyn College of the City University of New York
Lauren.Fink7@gmail.com

ANJALI KRISHNAN
Brooklyn College of the City University of New York
Anjali.Krishnan05@brooklyn.cuny.edu

JOSHUA FOGEL
Brooklyn College of the City University of New York
Jfogel@brooklyn.cuny.edu

LORIN BERMAN
Brooklyn College of the City University of New York
Berman_Lorin@yahoo.com

ROSE BERGDOLL
Brooklyn College of the City University of New York
RoseBergdoll@gmail.com

ABSTRACT

Mathematical competency is related to performance in introductory statistics courses and may be a roadblock to successful course completion. We developed a new measure (Math Assessment for College Students, MACS) of basic mathematics skills that improves upon measures previously used in undergraduate settings. The MACS is freely available and contains items not typically included on standardized measures of mathematical ability. We administered the 44-item MACS to 414 undergraduate psychology statistics students, and used a multiple correspondence analysis to eliminate 14 items, resulting in a 30-item measure with strong psychometric properties. MACS scores showed statistically significant moderate correlations with a commonly used standardized measure of basic mathematics skills and with overall statistics course grade. We discuss the utility of the MACS and how the MACS may help course instructors identify areas of mathematical deficiency that require remediation.

Keywords: *Statistics education research; Mathematics assessment; Mathematical knowledge*

1. INTRODUCTION

An introductory course in statistics is one of the requirements for an undergraduate degree in psychology, business, and many health and social science majors, as well as a prerequisite for graduate school (Chew & Dillon, 2014; Lalonde & Gardner 1993; Stoloff et al., 2009). Such a course typically covers descriptive and inferential statistical approaches (Pagano, 2012), which together provide a comprehensive understanding of the analytic processes behind scholarly research (Galli, Chiesi, & Primi, 2008). One of the overall goals of an undergraduate statistics course is to understand how to

analyze and interpret data (GAISE College Report ASA Revision Committee, 2016). These skills are necessary for conducting research and evaluating published research. Moreover, basic statistical literacy is necessary to appraise and critique information in the world (delMas, Garfield, Ooms, & Chance, 2007; Gal, 2002). Being able to evaluate claims based on statistics is crucial for individuals to be educated citizens of a democracy, especially because many policy and healthcare decisions are determined by statistics (Chew & Dillon, 2014). In fact, much of the statistical information that the media reports is in the form of graphs and/or percentages; thus, having an understanding of the mathematical connection between these summary statistics and the sample from which they were determined can help individuals be more cognizant of the challenges in their environment (Gal, 2002).

Previous research has demonstrated a connection between mathematics ability and statistics course performance; moreover, mathematical competency appears to be a roadblock to the successful completion of undergraduate statistics courses for some students. Therefore, statistics instructors may benefit from having tools to identify areas of weakness with basic mathematics that they can then remediate to improve student performance and retention. This paper describes the development and preliminary validation of a new measure (i.e., the Math Assessment for College Students, MACS), including content validity, dimension structure, and convergent validity. Guided by previous literature, we also investigate gender differences in MACS scores.

In developing the MACS, we aimed to improve upon existing measures. First, we strove to develop a measure that was easy to administer and score, and able to gauge students' understanding of basic mathematics without study or preparation. Second, the MACS was designed to test similar domains to previous measures (Johnson & Kuennen, 2006; Mulhern & Wylie, 2004), but incorporate greater use of items that involve applying symbol notation (i.e., < and >, Σ), numerical interval categorization, rounding, analysis of greatest common factors, elapsed time, and ability to define a mathematical concept (i.e., parallel lines) both in words and through illustration. We also included an item assessing the ability to extract information from a line graph to calculate the mean. We felt that incorporating these items would be helpful, as students would be revisiting many of these domains of mathematics during their statistics course. Third, in contrast to similar measures (e.g., Galli et al., 2008; Johnson & Kuennen, 2006), we avoided a multiple-choice format in favor of fill-in responses to have greater assurance that responses were not based on guesswork. Finally, as opposed to many standardized mathematics achievement and basic skills measures (e.g., KeyMath™-3 Diagnostic Assessment, Connolly, 2007; Kaufman Test of Educational Achievement, Third Edition, Kaufman & Kaufman, 2014; Wide Range Achievement Test-Fourth Edition (WRAT4), Wilkinson & Robertson, 2006), we plan to make the MACS freely available to instructors.

2. BACKGROUND

To succeed in an introductory statistics course, one should possess basic mathematical fluency and the ability to engage in critical thinking (Aron, Coups, & Aron, 2011; Pagano, 2012; Tomasetto, Matteucci, Carugati, & Selleri, 2009). Some studies have found moderate to strong relationships between statistics course performance and basic mathematics knowledge—but most of these studies used a composite mathematics skills score that included items within various mathematical domains (e.g., Chiesi & Primi, 2010) or a proxy for basic mathematics skills such as high school mathematics grade point average (e.g., Dupuis et al., 2012). There is some literature examining the relationship between specific mathematics skills and understanding of statistical concepts within the context of performance in introductory statistics courses. Bourne (2014) investigated mathematical ability related to success in a year-long undergraduate research methods/statistics course. Overall, higher scores on mathematical procedures (equations/basic operations, square and square root, rounding, number sequences) and interpretation (graphical interpretation, tables interpretation, converting decimals and percentages) predicted significantly higher course grades (though effect sizes were small); in contrast, higher scores on mathematical semantics (defining terms, < > symbols, negative numbers) were not a significant predictor of course grade.

Johnson and Kuennen (2006) explored the relationship between final grade in an introductory business statistics class and performance on a 15-item multiple choice mathematics quiz (total score and scores on each item). Results revealed an effect for overall mathematics score and for questions covering basic concepts in arithmetic, algebra, and geometry (e.g., manipulating simple systems of

equations, manipulating ratios, dividing fractions, two-step word problems to find the area of a rectangle, and estimating square roots). Notably, mathematics skills were an important determinant of statistics success regardless of the relative emphasis on mathematical computation (versus interpretation) by the specific course instructors. Also, neither taking calculus nor standardized mathematics test scores (measuring mathematics skills such as algebra, geometry and trigonometry) had a significant effect on course performance (Johnson & Kuennen, 2006). Overall, the existing literature supports a connection between basic mathematics skills and statistics course performance with some support for the value of items tapping mathematical equations, square and square root, rounding, number sequences, decimals and percentages, and understanding graphs/tables.

For many students, the basic mathematics skills required for introductory statistics courses pose a challenge (Conners, Mccown, & Roskos-Ewoldsen, 1998; Tomasetto et al., 2009). Previous research has sought to understand deficiencies in mathematical understanding among university students and to evaluate whether such deficiencies influence course grades. One study assessed for 10^{th} grade mathematics skills in undergraduate students in a business core "production" class and found difficulties with basic mathematics skills (Jones, Price, & Randall, 2011). This study was then replicated for both a business statistics and a quantitative analysis course at a different university. Though students varied in college level (e.g., sophomore, junior, senior), campuses, and courses (in addition to prerequisite requirements), their difficulty with basic mathematical skills was consistent. Moreover, there was a significant positive relationship between course performance and percent correct on the skills test, with stronger mathematics skills associated with better letter grades earned in the courses (Jones et al., 2011).

The importance of basic mathematics skills extends beyond the mere ability to perform calculations appropriately. A sample of psychology undergraduate students was tested on major mathematical domains related to understanding statistics including calculations involving decimals and fractions, algebraic reasoning, graphical interpretation, proportionality and ratio, probability and sampling, and estimation (Mulhern & Wylie, 2006). Student scores on the 32-item mathematics test indicated poor performance on probability and sampling, estimation, proportionality, ratio, and calculation (Mulhern & Wylie, 2006). The authors discussed the serious implications of these deficiencies for the teaching and learning of statistics and the unfortunate increased reliance on statistical software packages in introductory statistics courses.

Galli and colleagues (2008) developed the Mathematical Prerequisites for Psychometrics (PMP: Prerequisiti di Matematica per la Psicometria), which measures mathematical ability needed by psychology students to complete introductory statistics courses. As mathematical ability plays a role in statistics achievement (Chiesi & Primi, 2010; Harlow, Burkholder, & Morrow, 2002; Lalonde & Gardner, 1993; Schutz, Drogosz, White, & Distefano, 1998), the PMP was designed to identify students with low levels of mathematical ability as well as ascertain specific areas of difficulty. Using the statistics course curriculum as their basis, the authors identified six domains (operations, fractions, set theory, first order equations, relations, and probability) tested by their 30-item scale to gather information on students' arithmetic abilities. Students with lower mathematical ability had difficulty with all six domains, specifically due to poor mathematical skills of absolute value, percentages, square roots, and decimals (Galli et al., 2008).

When assessing mathematical competency in undergraduate students, it is important to consider the literature on gender differences. At the elementary and middle school levels, research suggests that girls are superior to boys in computation and equal to boys in understanding mathematical concepts (Tsui, Xu, & Venator, 2011). Observable gender differences that favor boys begin in high school on problem-solving tasks and these differences carry into performance on the mathematics portion of the SAT (Tsui et al., 2011). There is also evidence for better performance by undergraduate males than females on measures of calculation, proportionality, ratio, and estimation (Mulhern & Wylie, 2006). This gender gap may be attributable to sociocultural factors including the internalization of social stereotypes regarding male superiority in domains of mathematics that influence female performance (Brown & Josephs, 1999; Eccles, 1987; Johnson, Barnard-Brak, Saxon, & Johnson, 2012; Spencer, Steele, & Quinn, 1999; Steele, 1997). More specifically, females may be aware of negative stereotypes related to their gender's performance in mathematics. This awareness can lead to a stereotype threat that engenders a self-fulfilling prophecy in which female students acknowledge defeat before even partaking in the mathematical task, and thus succumb to the threat by performing worse on the assessment despite efforts to do well (Johnson et al. 2012, Spencer et al., 1999). Alternatively, some suggest that weaker

female performance in mathematical areas in which males perform well may relate to females having poorer spatial-mechanical skills (Casey, Nuttall, & Pezaris, 2001).

Also, although female scores on the SAT and other standardized mathematics tests are often lower than male scores, females tend to achieve better grades in high school and college courses (Halpern et al., 2007; Wilder & Powell, 1989). Factors such as competitive test-taking environments or mathematics anxiety may produce gender differences in exam scores that exaggerate underlying gender differences in mathematics skills (Devine, Fawcett, Szűcs, & Dowker, 2012; Niederle & Vesterlund, 2010). The discrepancy between gender differences observed on achievement tests and actual course performance, however, implies that course grades reflect learning in the larger social context of the classroom (Voyer & Voyer, 2014). Grades attained in semester-long classes likely reflect effort and persistence for a greater duration compared to standardized tests that assess academic abilities at one point in time.

3. METHODOLOGY

3.1. SETTING AND SAMPLE

Participants were 414 students enrolled in introductory undergraduate Statistical Methods in Psychological Research courses, commonly known as "psychology statistics," at an urban public college in the northeast United States. During the 15-week semester, each class meets twice a week for 75-minute lectures (taught by a faculty member) and once a week for a 110-minute laboratory session (taught by a graduate student). Typically, the first part of the course covers descriptive statistics, mean, variance, standard deviation, *z* scores, correlation, prediction, the normal curve, sampling, and basic probability theory. The second part covers basic principles of hypothesis testing, decision errors, effect size, power, *z* tests, and *t* tests for a single sample and dependent means. The third part covers *t* tests for independent means, analysis of variance, chi square tests, strategies for non-normal populations, and an overview of advanced statistical procedures (e.g., multivariate analyses, reliability, causal modeling).

Over three semesters (Fall 2013, Spring 2014, Fall 2014) students who enrolled in psychology statistics taught by a single instructor (the first author) were invited to complete the MACS. During the Spring 2015 semester, additional students from five different psychology statistics courses at the same college (taught by other instructors) were invited to complete the MACS. To reduce participant burden, only a subset of students (n = 156) was asked to complete the WRAT4. Participation was completely optional and voluntary, and students were not compensated for their participation nor penalized for non-participation. All participants were treated in accordance with American Psychological Association ethical guidelines (American Psychological Association, 2002) with an IRB-exempted protocol.

The response rate of 55.9% was calculated from the 740 students approached with 290 declining and an additional 36 MACS questionnaires not fully completed (414/740 ×100%). The sample consisted of 324 females and 90 males; thus, more than three-fourths (78.3%) of the sample was female. More than one-third (39.6%) of students were in their third year of undergraduate study. The detailed distribution of education levels among the sample was: 11 (2.7%) students in their first year of undergraduate education, 91 (22.1%) students in their second year of undergraduate education, 164 (39.8%) students in their third year of undergraduate education, 107 (26.0%) students in their fourth year of undergraduate education, 17 (4.1%) students in their fifth year of undergraduate education, 17 (4.1%) students in postgraduate education, and 5 (1.2%) non-degree students; two students did not report their year in school. Although we did not ask participants to report their age in years, most students were of typical college age (i.e., 18–24). Additionally, race/ethnicity was equally distributed between white and non-white: 206 (50.4%) White/Caucasian, 72 (17.6%) Black/African American, 58 (14.2%) Asian, 57 (13.9%) Hispanic/Latino, and 16 (3.9%) biracial; five students did not report their race/ethnicity.

3.2. PROCEDURE

Students completed the MACS and WRAT4 during the first week of the academic semester. At this time, students also provided basic demographic information including their gender, race/ethnicity, and year in school. The allotted time to complete the MACS was 20–25 minutes. A small number

(approximately 20) students required an additional 15 minutes and these students were accommodated. Students taking the WRAT4 in conjunction with the MACS were given a total of approximately 60 minutes to complete the assessments. The ordering of the MACS and WRAT4 was randomized. We did not permit students to use calculators during administration of the MACS or WRAT4.

Highly trained research assistants scored the MACS and WRAT4 by hand using objective procedures to derive a score for each item and a total score for each measure. After scoring was completed, data were entered into an SPSS Statistics Version 22 (IBM Corporation, 2013) database by one researcher and checked by a second researcher.

3.3. MEASURES

WRAT4 As noted above, a subset of students (n = 156) completed the WRAT4 (Wilkinson & Robertson, 2006) in addition to the MACS. We used the mathematics computation subtest of the WRAT4, which consists of 40 questions, to assess proficiency with basic mathematics (Wilkinson & Robertson, 2006). The WRAT is a series of tests first published in 1946 as a simple, rapid assessment of essential academic skills. The WRAT4 is norm-referenced and has been standardized on a national sample of over 3,000 participants ranging in age from 5 to 94 (Wilkinson & Robertson, 2006). The computation subset contains 20 questions related to addition, subtraction, multiplication, and long division of whole numbers, and eight and student work where a group of students used an applet to simulate a sampling distribution then used a z-statistic and calculator to find a p-value on a formative assessment questions related to the addition, subtraction, multiplication, and division of fractions. Additionally, there are three questions related to basic multiplication of decimals, one question on percentage values, two questions related to inter-conversion between fractions, decimals and percentage values, four questions testing competency in algebraic expressions, one question on numeric pattern recognition, and one question on rounding. The WRAT4 was administered in paper-and-pencil format; students computed and manually wrote responses on the test. Two forms of the WRAT4 mathematics computation subtest are available, a Blue Form and a Green Form, and we administered the Blue Form. For the Blue Form of the Mathematics Computation subtest, the median Cronbach's alpha (Cronbach, 1951) reliability coefficient (measuring internal consistency across various age ranges) is 0.89 (Wilkinson & Robertson, 2006). On average, scoring time per WRAT4 protocol is approximately 2–3 minutes.

Math Assessment for College Students (MACS) The course instructor (author LR) with assistance from two students (authors LF, LB) developed the items for the MACS after consulting various standardized achievement measures and/or basic mathematics skills tests such as the

- WRAT4
- PMP
- California Basic Educational Skills Test™
- Wonderlic Basic Skills Test
- Stanford Achievement Test Series

Test items are administered using a paper-and-pencil method in which participants calculate and write responses manually on the test form. We initially administered possible test items to 10 undergraduate students who were psychology, biology, or chemistry majors known to the authors (i.e., classmates of authors LF and LB or students in the lab of author LR). These individuals provided feedback about the understandability of test items. Approximately 10 items were revised based on feedback (minor word changes or numerical changes) and 4 items were eliminated. For example, some items were perceived as being too easy (e.g., writing the title of a figure, simple addition) and others were either too difficult or confusing (e.g., questions about probability and interpreting a complex formula). This early development phase resulted in an initial MACS version that consisted of 44 items.

Following intital development, we engaged in a data analysis process, described below, which led to the eliminiation of 14 items. The resulting final measure consisted of 30 items and will subsequently be referred to as the 30-item MACS. Table 1 presents sample MACS items. To help establish content validity, the authors consulted with a professional mathematics tutor for grades K–12, who helped assign each item to a corresponding mathematical content domain. The test protocol was then

distributed to five psychology instructors with expertise in mathematics and/or who teach mathematics intensive courses (e.g., psychology statistics, experimental psychology) who agreed to serve as content raters. Raters were asked to indicate the content domain(s) best represented by each item. Table 2 presents the mathematical content area tested by the 30-item MACS and corresponding item numbers. As shown in Table 2, the expert raters agreed strongly (100% agreement) with the intended content domain for 29 of the 30 items and agreed strongly (80% agreement with the intended contain domain for 1 item—i.e., item #29); notably one of the raters felt that this item belonged within basic arithmetic skills instead of the decimals, fractions, and percent content domain.

Table 1. Sample questions from the 30-item MACS

Item number	Item
Item 1	(12–2) × 3 – 8 ÷ 2 =
Item 8	Convert 10 ± 15 to a range:
Item 26	What percent of 80 is 100?
Item 29	Put the numbers 0 to 99 into 10 uniform categories (for example 0-3, 4-6, …)
Item 34	State whether the fraction on the left is less than or greater than the fraction on the right. Use > or <. 2/3 3/5

Note. The MACS will be provided to instructors by e-mailing the corresponding author.

Table 2. Mathematical content domains and expert rater agreement for MACS items

Intended content domain	MACS item	Rater agreement (%) with intended content domain
Basic arithmetic skills: includes addition, subtraction, multiplication, division, rounding, and greatest/least common factors	1	100
	15	100
	16	100
	17	100
	23	100
Basic algebraic skills: includes prealgebra, basic algebra, squares, square roots, exponents, summation, and operations with negative numbers	6	100
	9	100
	10	100
	11	100
	12	100
	13	100
	14	100
	24	100
Decimals, fractions, and percentages: includes calculation, conversion, and comparison of fractions, decimals, and percentages	2	100
	3	100
	4	100
	5	100
	8	100
	18	100
	19	100
	25	100
	26	100
	27	100
	28	100
	29	80
Categorization and ranges: includes conversion of whole numbers, ranges, and elapsed time	7	100
	21	100
	22	100
Visual understanding: includes parallel lines and graphical interpretation	20	100
	30	100

For all 30 MACS items, no partial credit was awarded and all items received a score of 0 = incorrect or 1 = correct, resulting in scores ranging from 0 to 30. Of the 30 MACS items, 28 have a single, objective right or wrong answer—e.g., for the item $(-2)^3 = ?$, the only correct response is −8. For two items (Q18, Q27, see Appendix), there is a degree of subjectivity to scoring, and we strove to develop criteria that minimized subjectivity and maximized interrater reliability. For example, for the item: What is the inverse (opposite) of the square root function?, we accepted several responses as correct including: square of a number, a number to the power of two, a number multiplied by itself, X^2, etc. On average, scoring time per MACS protocol was approximately 2–3 minutes.

3.4. PLANNED ANALYSES

As noted above, scoring of the MACS protocols was accomplished by a single rater and rescored by a second, independent rater. Inter-rater reliability analyses for the two items with slightly subjective scoring criteria used the Cohen's kappa coefficient (Cohen, 1960). Internal consistency of the original 44 items was assessed using the Kuder-Richardson Formula 20 [KR-20] for dichotomous data (Kuder & Richardson, 1937). We next examined properties of individual MACS items using descriptive measures such as item difficulty (proportion of students who answered each item correctly) and item discrimination (computed as a corrected point-biserial coefficient of correlation for each item and the overall score, excluding the specific item being correlated); these values are reported in the Appendix.

An important study goal was to investigate the dimension structure of the measure and engage in item selection. To account for the categorical nature of the data in our analyses (0 = incorrect, 1 = correct), we used a technique related to principal component analysis (PCA) that was designed specifically for categorical data; this technique is known as multiple correspondence analysis (MCA; Abdi & Valentin, 2007; Greenacre & Blasius, 2006; Sourial et al., 2010). Although a PCA would demonstrate the variability across participants based on the overall number of correct answers, we were interested in the variability across participants based on their pattern of correct and incorrect responses on the MACS. Just like a PCA, MCA generates dimensions that explain the variance in a given dataset. Using the dataset consisting of 44 items, we performed a permutation test to determine which dimensions contributed significantly to the observed variance in our data (Peres-Neto, Jackson, & Somers, 2005). For the permutation test, re-sampling without replacement was performed within each variable. The MCA was then performed on the permuted data, and eigenvalues were retained per dimension. This procedure (repeated 10,000 times) provides a data-driven distribution for the eigenvalues under the null hypothesis, which can then be used to derive *p*-values in order to determine whether the eigenvalue for each dimension is statistically significant (i.e., $p < 0.0001$ for 10,000 permutations (Berry, Johnston, & Mielke, 2011; Peres-Neto et al., 2005). Using this information we were able to decide which MCA dimensions to retain for interpretation. If the eigenvalue for a dimension was significant, the dimension was retained and if the eigenvalue was not significant, the dimension was ignored.

Bootstrap resampling—which, contrary to permutation, resamples participants with replacement (Efron & Tibshirani, 1994)—was used to identify the MACS items that contributed significantly to each dimension through a procedure called the "bootstrap ratio" test (BSR; Hesterberg, 2011; McIntosh & Lobaugh, 2004) as described by Beaton and colleagues (2014). The BSR is the mean of a bootstrap distribution divided by its standard deviation. Bootstrap resampling was performed 10,000 times to create distributions of variation around each response in the MCA dimension space. BSRs are interpreted like Student's *t* where a typical statistical threshold (i.e., $\alpha = 0.05$) corresponds to a *t* or *z*-value (i.e., ±1.96 corresponds to $p = 0.05$). We applied a more conservative BSR cutoff of ±3.75 (which corresponds to $p = 0.000101$) to select significant and highly stable MACS items. This method allowed us to eliminate items that contributed least to the variance in our data and arrive at the shorter (30 item) assessment. As a descriptive measure of how each item contributes to the variance explained by a particular dimension, we used percent contribution (see Abdi & Valentin, 2007, for computation details).

An additional study goal was to establish convergent validity for the MACS. We performed a Pearson's correlation on students' MACS and WRAT4 scores (as percent correct), and an effect size

R^2 was calculated. This correlation was performed on the subset of 156 students who took both the MACS and WRAT4.

To explore the role of the MACS for statistics course performance, we performed a Pearson's correlation between MACS score (as percent correct) and overall course grade (as a percent) for the three semesters taught by the same instructor using a consistent grading rubric (n = 249). Examinations were graded objectively, and although partial credit was awarded for the hypothesis testing problems, point values were assigned using a detailed scoring rubric with predetermined values for all portions of the response. Overall course grade was determined by the following formula: exam I score (22% of grade), exam II score (22% of grade), exam III score (22% grade), laboratory performance and homework assignments (26% of grade), presentation of a research article (4% of grade), and attendance/participation (4% of grade).

Finally, to compare MACS score (as percent correct) in male and female students, we conducted an independent samples t-test (n = 414), and calculated a Cohen's d effect size. We also conducted an independent samples t-test to compare performance in the statistics course overall for male and female students, and calculated a Cohen's d effect size. This independent samples t-test utilized data from the three semesters taught by the same instructor (n = 249), ensuring that the course grading rubric was the same for this subset of students. A differential item functioning analysis (DIF) was also conducted on each of the 44 items to examine the item-wise differences between male and female students based on overall test score (Swaminathan & Rogers, 1990).

The R Program for Statistical Computing was used to perform (R Core Team, 2013) the MCA, and related inference tests were performed with the "ExPosition" and "InPosition" packages (Beaton, Fatt, & Abdi, 2014). Other tests were performed with with SPSS except for Cohen's d, which was calculated manually. All p-values were two-tailed; except the p-values for the BSR tests, which are by definition one–tailed. For the BSR test, each variable was tested for statistical significance separately for correct and incorrect responses based on which end (i.e., positive or negative) of the dimension they loaded. However, a conservative cut–off of ± 3.75 (which corresponds to p = 0.000101) was applied for the BSR test to ensure that we retained only highly significant and stable MACS items.

4. RESULTS

Interrater reliability (kappa) for the two MACS items with slightly subjective scoring criteria was k = 0.97, k = 0.98, respectively, providing evidence of excellent interrater agreement. The KR-20 for the 44 MACS items was 0.894, indicating good internal consistency amongst the items (Streiner, 2003; Streiner & Norman, 1989). Using the dataset consisting of 44 MACS items, we used MCA to analyze the underlying construct of the MACS. MCA generates dimensions to describe the variance among observations (i.e., students), and these dimensions are computed as linear combinations of all the original variables (i.e., the 44 MACS items). Therefore, all 44 items load on all dimensions, but some items may contribute more to specific dimensions than other items. We used a bootstrap ratio (BSR) test to determine the importance of a given item for each dimension of interest.

The MCA revealed that our dataset had 13 total dimensions, and a permutation test indicated that three dimensions had eigenvalues that were each statistically significant at the 0.0001 level. More specifically, the first dimension explained 94.15% and the second dimension explained 3.45% of the variance of our data. The third dimension only explained 0.85% of the variance of our data, thus, we focused our attention on the first two dimensions. The BSR test showed that 30 items loaded significantly on the first dimension and 14 items did not. Many of the 14 items (items 6, 9, 10, 12, 16, 17, 20, 31, 32, 39, 41, 42, 43, and 44) that did not load significantly on the first dimension overlapped in (intended) content domains with items that *did* load significantly on the first dimension. The KR-20 for the 30 items that loaded significantly on the first dimension was found to be 0.886, indicating good internal consistency amongst the items.

Figure 1 presents the MCA coordinate plot for dimensions 1 and 2. Most of the variance in our data was attributed to dimension 1, as this dimension contained the largest difference between students based on their MACS scores. The origin (i.e., the point of intersection of the axes) of the MCA represents the average. Each dimension (i.e., each axis) is interpreted based on observations located at either end of that axis. Dimension 1 (i.e., the horizontal axis) differentiated students with high scores from students with low scores. Figure 1 in the right panel shows that students who scored ≥ 30 of the 44 items correct

are to the right of the origin and students who scored < 30 of the 44 items correct are to the left of the origin. This difference was driven by the number of correct and incorrect responses (Figure 1, left panel), where variability among students increased as the number of incorrect responses increased. Although all MACS questions load on all dimensions, only 30 of the 44 items significantly contributed towards the variance structure of dimension 1. This indicates that correct or incorrect responses to these 30 questions were more likely to be indicative of differences in MACS performance.

Dimension 2 (i.e., the vertical axis) focused on the pattern in which students answered items correctly or incorrectly on particular subsets of items, more specifically, identifying subsets of items such as basic operations of decimals and fractions that students answered correctly (i.e., original MACS items 34, 35, 38), and different subsets of items such as inter-conversion between decimals, fractions, and percentages that students answered incorrectly (e.g., original MACS items 2, 3, 4, 5, 25, and 26). These items have utility in evaluating student performance on the (retained) MACS. Each item's percent contribution to the variance explained by the first and second dimensions of the MACS is displayed in the Appendix, as overall percent contribution per item, and as separate percent contributions for correct and incorrect responses. BSR test values reported in parentheses. Almost all of the statistically significant items (except MACS items 1 and 29) had a percent contribution to dimension 1 of 1.5 or greater. It is important to note that only the BSR test determined the statistical significance of each item's importance to the dimension, and that the percent contributions are reported as descriptive information.

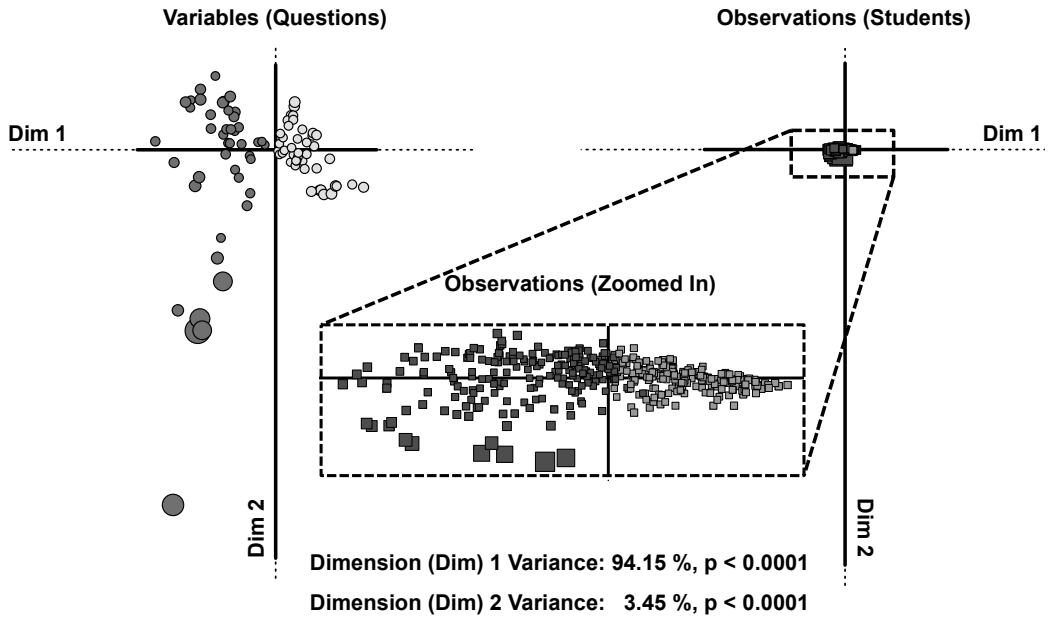

Figure 1. Multiple correspondence analysis plot for dimensions 1 and 2

Figure 1 Legend. The left panel shows the variables (questions/MACS items). Dimension 1 is represented on the horizontal axis and dimension 2 is represented on the vertical axis. Dark gray dots represent the incorrect responses to the 44 items and light gray dots represent the correct responses to the 44 items. Dimension 1 shows the pattern of getting items correct or incorrect on the assessment, whereas dimension 2 shows which specific items students answered mostly correct or mostly incorrect on the whole. The size of the dots represents their percent contribution to the dimension, with larger dots having a greater percent contribution to the dimension. The percent contribution values are reported in the Appendix.

The right panel shows the observations (students) on the same scale as the variable plot. Dimension 1 is represented on the horizontal axis and dimension 2 on the vertical axis. Each square represents one of the 414 students included in the analysis. The insert zooms in on the students to show the distribution more visibly. The size of the squares represents the percent contribution of each student to the variance explained by the dimension, with larger squares having a greater percent contribution. Dark gray squares represent students who answered < 30 (out of 44) items correctly on the MACS, and light gray squares represent students who answered ≥ 30 (out of 44) items correctly on the MACS. Dimension 2 shows the

students who answered specific subsets of questions correctly (top end of axis) or incorrectly (bottom end of axis) as compared to the other students.

Pearson's correlation between the percent of correct responses on the 30-item MACS and 40-item WRAT4 showed a moderate to strong correlation of $r = 0.78$, $n = 156$, $p < 0.001$. Table 3 presents descriptive statistics (mean and standard deviation) for the MACS and WRAT4. An effect size R^2 of 0.607 was calculated between the percent correct on the MACS and WRAT4. Additionally, the 30-item MACS demonstrated a weak to moderate correlation with statistics course grade, $r = 0.41$, $n = 249$, $p < 0.001$. An effect size of $R^2 = 0.170$ was calculated between the percent correct on the MACS and course grade.

Table 3. Descriptive statistics for the 30-item MACS and WRAT4 analysis (n = 156)

Variable	MACS score (% correct)	WRAT4 score (% correct)
Mean (*SD*)	61.23 (22.20)	76.35 (14.15)
Correlation with WRAT4	$r = 0.78$ ($p < 0.001$)	---

Note. MACS = Math Assessment for College Students; WRAT4 = Wide Range Achievement Test 4

Table 4 presents gender comparisons for the 30-item MACS and overall course grade, including descriptive statistics (mean and standard deviation) and effect sizes. There was a significant difference in mean MACS score (as percent correct) between males and females, with males outperforming females by approximately 10 percentage points. Given the large difference in sample sizes between males and females, we also conducted a Welch test and results were the same: *Welch's* $F(1,153.95) = 15.71$, $p < 0.001$. However, there was no difference in overall statistics course grade for males and females: see Table 4 for *t*-test results and *Welch's* $F(1,75.41) = 0.92$, $p = 0.34$. In addition, the differential item functioning (DIF) analyses on each of the 44 items revealed that only Q8 showed statistical significance between male and female students for uniform DIF ($\chi^2(1) = 23.45$, $p < 0.001$) but not for non-uniform DIF ($\chi^2(1) = 0.895$, $p = 0.34$).

Table 4. Gender comparisons for the 30-item MACS score (n = 414) and overall course grade (n = 249)

Variable	Males mean (*SD*)	Females mean (*SD*)	*t*-statistic	*p*-value	Cohen's *d*
MACS score	68.89 (20.27)	59.10 (22.27)	$t(412) = 3.96$	$p < 0.001$	0.47
Overall course grade	78.70 (14.58)	80.88 (13.98)	$t(247) = -2.18$	$p = 0.340$	0.16

Note. MACS = Math Assessment for College Students

5. DISCUSSION

The 30-item MACS is comprised of questions that specifically address mathematical domains related to successful performance in undergraduate introductory statistics (and similar) classes. Understanding the basics of both descriptive and inferential statistics is a central objective of undergraduate introductory statistics courses (Wilson, 2013). Such courses tend to emphasize the accurate use and interpretation of basic statistical procedures, and presume that students can work properly with symbols (Lunsford & Poplin, 2011). Moreover, students learn to perform mathematical computations and report statistical findings using the appropriate notations.

MACS questions were designed to assess broadly these various competencies. The 30-item MACS includes computational questions requiring mathematical calculations (i.e., solving for an unknown variable in an algebraic expression). Student responses to these questions may be prone to errors in numeric calculations. The MACS also includes more conceptual questions (i.e., questions related to visual understanding of mathematical concepts and questions testing correct symbol notation in the form of a less than or greater symbol. Mastery of symbol notation is critical when reporting significant results of statistical tests (i.e., indicating $p < 0.05$). The MACS can serve as a tool to identify the types

of mathematical concepts with which undergraduate introductory statistics students struggle. If the area of difficulty lies in more computation-related skills, exercises allowing students to practice calculations may be useful in helping students become more comfortable with calculations. If the weakness is linked to conceptual understanding of mathematical elements, reviewing symbols and explaining what they stand for and when they are to be used can provide a stronger conceptual foundation.

In order to gauge competency with basic mathematics skills effectively, MACS items should be able to differentiate between students based on their performance on the assessment. For example, if the majority of students answered a certain item wrong, or if everyone answered a certain item right, those items would not be useful to instructors. We began with 44 items and were able to eliminate 14 items that did not contribute meaningfully to the variance in student performance. Only three dimensions were significant at the 0.001 level—and of these, the first two dimensions explained the majority of the variance in students' responses. We determined that the largest variance (94.15%) in our data was given by dimension 1, which differentiated students with high MACS scores from students with low MACS scores (determined by the number of correct and incorrect responses). This difference was driven by the variability of student responses to 30 of the 44 items. We eliminated the remaining 14 items that did not contribute significantly to the differences in MACS performance (i.e., items that on average were all answered correctly or incorrectly).

Dimension 2 accounted for the second largest amount of variance (3.45%), and explained the pattern of being correct versus incorrect for specific subsets of items. For example, the overall response trend demonstrated the difference between answering questions 34, 35, and 38 incorrectly (which involves using symbol notation >/< to determine which fraction and decimal is larger), and answering questions 2, 3, 4, 5, 25, and 26 correctly (which involves inter-conversion between fractions, decimals, and percentages, as well as determining percentage of a number). The fact that students perform differently on these subsets of items suggests that there are differences in students' level of basic mathematics skills upon beginning the course. These differences could relate to recency of mathematics practice, level of mathematical background from high school, and/or college major. Students who have not practiced basic mathematics skills since high school may have trouble performing algebra and inter-conversions between decimals, fractions, and percentages.

The 30-item MACS had a statistically significant moderate to strong correlation with the WRAT4, suggesting that the MACS assesses basic mathematics skills similar to the widely used WRAT4 without being redundant. The MACS includes content areas not covered by the WRAT4, such as graph interpretation, comparing fractions and decimals using symbol notation, algebraic expressions using the summation symbol sigma, determining a number range, finding the greatest common factor between two numbers, placing numbers into uniform categories, and drawing parallel lines and describing what they are in words. Additionally, the WRAT4 contains items with overlapping content (e.g., 12 items on addition and subtraction of whole numbers) and items that arguably are too easy for college students and may frustrate or seem irrelevant to them (e.g., $1 + 1 = ?$; $5 – 1 = ?$). The MACS contains fewer overlapping items and was designed to evaluate foundational mathematics skills that could contribute to statistical literacy and reasoning, which is consistent with efforts in statistical education to foster students' abilities in thinking and reasoning critically (GAISE, 2016).

A Pearson's correlation between the 30-item MACS and overall statistics course grade was statistically significant and moderate, suggesting that basic mathematics skills are relevant to understanding overall course performance. In future research, it would be useful to target and train basic mathematical knowledge and abilities with the hope of demonstrating improvements in students' understanding of statistical concepts, statistics course performance, and student retention. To this end, refresher sessions could be offered early on in the semester (or in conjunction with the introduction of certain course material) to review foundational computations, proper notations and symbols, and graph interpretations so that students in the course have roughly the same knowledge level in key mathematical concepts (Chiesi & Primi, 2010). One could then gauge changes/improvements in mathematics skills from beginning to the end of the semester and determine how such changes map with course performance. To our knowledge such a study has not yet been carried out. If adding content to the course is unduly burdensome, other options include making online mathematics reviews a required part of (or prerequisite to) the course or offering peer-tutoring sessions outside of class time.

We found a statistically significant gender difference on the 30-item MACS with males outperforming females for percent correct by roughly 10 points, though we did not observe a

corresponding difference in overall course grade by gender. On further investigation using a differential item functioning (DIF) analysis, only 1 of 44 items (i.e., convert 100 $\pm$ 15 to a range) showed a significant effect—with males outperforming females. Additional research would be required to provide a rationale for the gender difference on this specific item but this result suggests a consistent pattern of minimal degree of difference in MACS item performance overall between males and females. In trying to account for the observed 10 percentage point difference, we considered several possible explanations—all of which would require further research to confirm. Although students were told that their MACS assessment would not impact course grades, some students may have found this to be a competitive test-taking environment, which can exaggerate underlying gender differences in mathematics skills (Niederle & Vesterlund, 2010). By contrast, overall course performance depends on continuous studying and preparation and course grade is based on various assessments such as quizzes, exams, homework assignments, and class participation. Thus, course grade is more comprehensive and reflective of proficiency with material over time, as opposed to a one-shot competitive testing situation. Additionally, although we did not assess for gender stereotypes, the literature suggests that discrepancies in mathematics performance may be due to the internalization of social stereotypes that highlight the inferior abilities of women in mathematics, leading them to perform worse than men on mathematical tasks (Brown & Josephs, 1999; Eccles, 1987; Johnson et al., 2012; Spencer et al., 1999; Steele, 1997). Finally, the MACS was clearly a math-intensive assessment, and previous research suggests possible greater mathematics anxiety felt by female students (Devine et al., 2012), which can negatively impact performance.

For those wishing to use the 30-item MACS, which will be made freely available to instructors, we suggest administering the assessment early in the semester. This will enable instructors to gauge students' level of background mathematical knowledge before delving into complex material. Having a grasp on the nature of students' deficits with basic mathematics can inform teaching interventions and direct students' approaches to studying in an effort to reinforce and build upon basic competencies. For example, we found that dimension 2 permits investigation of patterns in the types of items students answer correctly or incorrectly. Thus, dimension 2 elucidates areas of weakness in basic mathematics because it specifies topics that are particularly difficult for students, rather than just detailing items that students answered correctly or incorrectly. Using this information, instructors can provide feedback and targeted assistance.

Our analysis also revealed that students were more likely to answer questions on inter-conversions between decimals, fractions, and percentages incorrectly, which are topics usually covered at the beginning of a statistics course (and also skills related to statistics course grades, Bourne, 2014); therefore, instructors may choose to spend more class time on these areas early in the semester. By contrast, if a student performs poorly on interpretation of graphs, the instructor can link the student to resources that review the theory behind graphical representation of empirical data and instruct about graphing basics. If a student fails basic calculation items, he or she might benefit from tutoring services and additional practice problems. A student whose MACS performance demonstrates difficulty with mathematical symbol notations, but is otherwise proficient in mathematics, may need a quick refresher on what the symbols represent and when and how they are used.

It is important to note study limitations. We only used the WRAT4 Blue Form and future research may wish to explore whether the correlation between the MACS and the WRAT4 remains as strong when using the Green Form. The WRAT4 Green Form contains one item evaluating proficiency with symbol notation <, >, or =, which is not a question assessed on the WRAT4 Blue Form. Also, we were unable to determine the correlation between the WRAT4 scores and the course grade given limitations of data collection. In future research, it will be important to demonstrate the utility of MACS in terms of its value in diagnosing students' mathematical readiness for statistics. In addition, students who completed the MACS in our study were exposed to all 44 original items and not just the 30 items retained following the MCA. The 14 eliminated items may have influenced student performance on the remaining questions. For example, if a student did not know how to answer a graphing question and started worrying about getting that question wrong, this may have affected the student's response to subsequent items. However, students were told that they did not have to go in order on the MACS and the unrestricted nature of the assessment may have reduced any influence from the eliminated items. Finally, we cannot rule out the possibility of selection bias for analyses related to the WRAT4. For the three semesters with data on course performance (n = 249), the retention rate was 94% (in other words,

17 students did not participate for reasons of missing data primarily relating to their being absent for the first 1–2 weeks of the semester or dropping the course). These students did not differ significantly from the remainder of the sample in terms of demographics or performance. However, in the semester where we approached additional students from psychology statistics courses taught by other instructors, only for the purpose of completing both the MACS and WRAT4, approximately two-thirds of those students declined to participate. The likely reason is because we were not offering compensation. We have no way of knowing how those nonparticipants fared in their respective course sections.

In summary, the 30-item MACS has strong psychometric properties and shows promise for the assessment of basic mathematics skills. The MACS taps domains of mathematics that are important for undergraduate statistics courses and are not necessarily represented on other available measures. We hope that the MACS will serve as a valuable resource for instructors who seek to identify students at risk of scoring poorly in mathematics-based undergraduate courses and that the MACS can help in effectively targeting remedial assistance and interventions.

ACKNOWLEDGEMENTS

Funding provided by the TOW Foundation. The authors wish to thank Beliz Hazan, Dr. Jennifer Drake, Susan Chi, Genea Stewart, Kamil Kloskowski, David Lederman, Matthew Fein, Dr. Gail Horowitz, Faigy Mandelbaum, and especially Dr. Rona Miles and Shira Stone for their assistance in establishing content validity for the MACS. We also thank the many students who volunteered their time to participate in this study.

REFERENCES

Abdi, H., & Valentin, D. (2007). Multiple correspondence analysis. In N. J. Salkind (Ed.). *Encyclopedia of measurement and statistics* (pp. 651–657). Thousand Oaks, CA: Sage.

American Psychological Association (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, *57*(12), 1060–1073.

Aron, A., Coups, E., & Aron, E. N. (2011). *Statistics for the behavioral and social sciences: A brief course* (5th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Beaton, D., Fatt, C. R. C., & Abdi, H. (2014). An ExPosition of multivariate analysis with the singular value decomposition in R. *Computational Statistics & Data Analysis*, *72*, 176–189.

Berry, K. J., Johnston, J. E., & Mielke, P. W. (2011). Permutation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(6), 527–542.

Bourne, V.J. (2014). To what extent is mathematical ability predictive of performance in a methodology and statistics course? Can an action research approach be used to understand the relevance of mathematical ability in psychology undergraduates? *Psychology Teaching Review*, *20*(2), 14–27.

Brown, R. P., & Josephs, R. A. (1999). A burden of proof: Stereotype relevance and gender differences in math performance. *Journal of Personality & Social Psychology, 76*(2), 246–257.

Casey, M. B., Nuttall, R. L., & Pezaris, E. (2001). Spatial-mechanical reasoning skills versus mathematics self-confidence as mediators of differences on mathematics subtests using cross-national gender- based items. *Journal for Research in Mathematics*, *32*(1), 28–57.

Chew, P. K., & Dillon, D. B. (2014). Statistics anxiety update refining the construct and recommendations for a new research agenda. *Perspectives on Psychological Science*, *9*(2), 196–208.

Chiesi, F., & Primi, C. (2010). Cognitive and non-cognitive factors related to students' statistics achievement. *Statistics Education Research Journal*, *9*(1), 6–26. [Online: https://www.stat.auckland.ac.nz/~iase/serj/SERJ9(1)_Chiesi_Primi.pdf]

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Conners, F.A., Mccown, S.M., & Roskos-Ewoldsen, B. (1998). Unique challenges in teaching undergraduate statistics. *Teaching of Psychology*, *25*(1), 40–42.

Connolly, A.J. (2007). *KeyMath-3 diagnostic assessment: Manual forms A and B.* Minneapolis, MN: Pearson.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

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. [Online: https://iase-web.org/documents/SERJ/SERJ6(2)_delMas.pdf]

Devine, A., Fawcett, K., Szűcs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions*, *8*(1), 33.

Dupuis, D. N., Medhanie, A., Harwell, M., Lebau, B., Monson, D., & Post, T. R. (2012). A multi-institutional study of the relationship between high school mathematics achievement and performance in introductory college statistics. *Statistics Education Research Journal*, *11*(1), 4–20. [Online: https://www.stat.auckland.ac.nz/~iase/serj/SERJ11(1)_Dupuis.htm]

Eccles, J.S. (1987). Gender roles and women's achievement-related decisions. *Psychology of Women Quarterly*, *11*(2), 135–172.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton, FL: CRC Press.

GAISE College Report ASA Revision Committee (2016). *Guidelines for Assessment and Instruction in Statistics Education College Report 2016*. [Online: http://www.amstat.org/education/gaise]

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, *70*(1), 1–25.

Galli, S., Chiesi, F., & Primi, C. (2008). The construction of a scale to measure mathematical ability in psychology students: An application of the Rasch Model. *TPM-Testing, Psychometrics, Methodology in Applied Psychology*, *15*(1), 3–18.

Greenacre M., & Blasius J. (2006). *Multiple correspondence analysis and related methods*. Boca Raton, FL: CRC Press.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*(1), 1–51.

Harlow, L. L., Burkholder, G. J., & Morrow, J. A. (2002). Evaluating attitudes, skill, and performance in a learning-enhanced quantitative methods course: A structural modelling approach. *Structural Equation Modeling: A Multisciplinary Journal*, *9*(3), 413–430. [Online: http://doi.org/10.1207/S15328007SEM0903_6]

Hesterberg, T. (2011). Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(6), 497–526.

IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Johnson, H. J., Barnard-Brak, L., Saxon, T. F., & Johnson, M. K. (2012). An experimental study of the effects of stereotype threat and stereotype lift on men and women's performance in mathematics. *Journal of Experimental Education*, *80*(2), 137–149.

Johnson, M., & Kuennen, E. (2006). Basic math skills and performance in an introductory statistics course. *Journal of Statistics Education*, *14*(2), 1–17. [Online: http://doi.org/10.1080/10691898.2006.11910581#.W9U2lmhKg2w]

Jones, T. W., Price, B. A., & Randall, C. H. (2011). A comparative study of student math skills: Perceptions, validation, and recommendations. *Decision Sciences Journal of Innovative Education*, *9*(3), 379–394.

Kaufman, A. S., Kaufman, N. L. (with Breaux, K. C). (2014). *Technical & interpretive manual: Kaufman Test of Educational Achievement* (3rd ed.). Bloomington, MN: NCS Pearson.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160.

Lalonde, R. N., & Gardner, R. C. (1993). Statistics as a second language? A model for predicting performance in psychology students. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, *25*(1), 108–125.

Lunsford, M. L., & Poplin, P. (2011). From research to practice: Basic mathematics skills and success in introductory statistics. *Journal of Statistics Education*, *19*(1), 1–22. [Online: https://doi.org/10.1080/10691898.2018.1483785]

McIntosh, A. R., & Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: Applications and advances. *Neuroimage*, *23*(1), S250–S263.

Mulhern, G., & Wylie, J. (2004). Changing levels of numeracy and other core mathematical skills among psychology undergraduates between 1992 and 2002. *British Journal of Psychology*, *95*(3), 355–370.

Mulhern, G., & Wylie, J. (2006). Mathematical prerequisites for learning statistics in psychology: Assessing core skills of numeracy and mathematical reasoning among undergraduates. *Psychology Learning & Teaching*, *5*(2), 119–132.

Niederle, M., & Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. *The Journal of Economic Perspectives*, *24*(2), 129–144.

Pagano, R. (2012). *Understanding statistics in the behavioral sciences*. Boston, MA: Cengage Learning.

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis*, *49*(4), 974–997.

R Core Team (2013). *R*: A language and environment for statistical computing. *R* Foundation for Statistical Computing, Vienna, Austria.
[Online: http://www.R-project.org/].

Schutz, P. A., Drogosz, L. M., White, V. E., & Distefano, C. (1998). Prior knowledge, attitude, and strategy use in an introduction to statistics course. *Learning and Individual Differences*, *10*(4), 291–308.

Sourial, N., Wolfson, C., Zhu, B., Quail, J., Fletcher, J., Karunananthan, S., Bandeen-Roche, K., Béland, F., & Bergman, H. (2010). Correspondence analysis is a useful tool to uncover the relationships among categorical variables. *Journal of Clinical Epidemiology*, *63*(6), 638–646.

Spencer, S. J., Steele, C. M., & Quinn, D. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613–629.

Stoloff, M., McCarthy, M., Keller, L., Varfolomeeva, V., Lynch, J., Makara, K., Simmons, S., & Smiley, W. (2009). The undergraduate psychology major: An examination of structure and sequence. *Teaching of Psychology*, *37*(1), 4–15.

Streiner, D. L. (2003) Starting at the beginning: An introduction to coefficient alpha and internal consistency, *Journal of Personality Assessment*, *80*(1), 99–103.

Streiner D.L., & Norman G.R. (1989) *Health measurement scales: A practical guide to their development and use* (1st ed., pp. 64–65). New York: Oxford University Press.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370.

Tomasetto, C., Matteucci, M., Carugati, F., & Selleri, P. (2009). Effect of task presentation on students' performances in introductory statistics courses. *Social Psychology of Education*, *12*(2), 191–211.

Tsui, M., Xu, X. Y., & Venator, E. (2011). Gender, stereotype threat and mathematics test scores. *Journal of Social Sciences*, 7(4), 538–549.

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204.

Wilder, G. Z., & Powell, K. (1989). Sex differences in test performance: A survey of the literature. *ETS Research Report Series*, *1989*(1), 1–50.

Wilkinson, G. S., & Robertson, G. J. (2006). *Wide range achievement test* (4th ed.). Lutz, FL: Psychological Assessment Resources.

Wilson, S. G. (2013). The flipped class: A method to address the challenges of an undergraduate statistics course. *Teaching of Psychology*, *40*(3), 193–199.

LAURA A. RABIN
Department of Psychology, Brooklyn College of CUNY
2900 Bedford Avenue
Brooklyn, NY 11210
USA

APPENDIX A.

Table 5. MACS items properties and contribution to dimensions

Original 44 item MACS	Retained 30 item MACS	Item properties		% Contribution to dimension 1			% Contribution to dimension 2		
		DIF	DISC	Overall	Correct (BSR)	Incorrect (BSR)	Overall	Correct (BSR)	Incorrect (BSR)
1	1	0.66	0.31	1.48	0.5 (4.46)	0.98 (–6.08)	0.31	0.1 (–1.01)	0.21 (1.49)
2	2	0.54	0.51	4.01	1.84 (10.1)	2.17 (–10.11)	3.40	1.56 (–4.82)	1.84 (3.9)
3	3	0.54	0.53	4.18	1.93 (10.49)	2.25 (–10.4)	3.34	1.54 (–4.83)	1.8 (3.81)
4	4	0.46	0.47	3.41	1.85 (9.66)	1.56 (–8.46)	2.95	1.6 (–4.8)	1.35 (3.41)
5	5	0.43	0.52	4.07	2.33 (12.54)	1.74 (–8.79)	2.55	1.46 (–4.91)	1.09 (3.02)
6		0.76	0.29	1.24	0.3 (3.5)	0.94 (–5.45)	0.86	0.21 (–1.5)	0.65 (2.32)
7	6	0.75	0.33	1.63	0.41 (4.04)	1.22 (–6.79)	0.13	0.03 (–0.65)	0.1 (0.82)
8	7	0.38	0.51	3.87	2.39 (12.95)	1.48 (–7.99)	1.42	0.88 (–3.81)	0.54 (2.1)
9		0.97	0.14	0.34	0.01 (0.56)	0.33 (–2.92)	0.11	0 (–0.18)	0.11 (0.88)
10		0.94	0.15	0.35	0.02 (0.88)	0.33 (–3.24)	0.92	0.06 (0.76)	0.86 (–2.29)
11	8	0.65	0.39	2.31	0.81 (6.01)	1.5 (–7.49)	0.40	0.14 (–1.29)	0.26 (1.36)
12		0.93	0.29	1.26	0.09 (1.8)	1.17 (–7.77)	3.41	0.24 (1.66)	3.17 (–4.11)
13	9	0.79	0.40	2.59	0.54 (4.76)	2.05 (–9.66)	1.68	0.35 (–1.91)	1.33 (3.51)
14	10	0.72	0.50	3.79	1.05 (7.27)	2.74 (–11.68)	1.69	0.47 (–2.23)	1.22 (3.38)
15	11	0.74	0.40	2.51	0.64 (5.33)	1.87 (–8.55)	0.27	0.07 (–0.88)	0.2 (1.2)
16		0.94	0.16	0.40	0.02 (0.92)	0.38 (–4.64)	0.62	0.04 (–0.62)	0.58 (2.02)
17		0.93	0.24	0.99	0.07 (1.62)	0.92 (–6.03)	0.24	0.02 (–0.43)	0.22 (1.25)
18	12	0.57	0.47	3.36	1.46 (9.13)	1.9 (–8.54)	0.03	0.01 (0.41)	0.02 (–0.35)
19	13	0.81	0.46	3.29	0.62 (5.33)	2.67 (–11.58)	0.92	0.17 (–1.39)	0.75 (2.34)
20		0.93	0.34	1.83	0.12 (2.2)	1.71 (–10.66)	0.01	0 (–0.08)	0.01 (0.22)
21	14	0.86	0.43	2.82	0.38 (4.09)	2.44 (–11.13)	0.05	0.01 (0.25)	0.04 (–0.55)
22	15	0.76	0.39	2.33	0.56 (4.94)	1.77 (–8.21)	0.67	0.16 (–1.28)	0.51 (2.23)
23	16	0.60	0.41	2.57	1.02 (7.02)	1.55 (–7.55)	0.05	0.02 (–0.51)	0.03 (0.46)

24	17	0.67	0.35	1.81	0.6 (5.17)	1.21 (–6.36)	0.03	0.01 (–0.37)	0.02 (0.45)
25	18	0.56	0.43	2.84	1.25 (7.57)	1.59 (–8.15)	3.89	1.71 (–4.55)	2.18 (4.51)
26	19	0.17	0.43	2.70	2.25 (16.83)	0.45 (–4.22)	0.49	0.41 (–4.07)	0.08 (0.84)
27	20	0.65	0.33	1.55	0.54 (4.91)	1.01 (–5.63)	0.38	0.13 (1.25)	0.25 (–1.35)
28	21	0.52	0.47	3.20	1.53 (9.17)	1.67 (–8.14)	0.67	0.32 (–2.0)	0.35 (1.68)
29	22	0.46	0.31	1.42	0.77 (5.64)	0.65 (–4.85)	0.08	0.04 (0.8)	0.04 (–0.52)
30	23	0.40	0.33	1.59	0.95 (6.89)	0.64 (–4.67)	0.23	0.14 (1.63)	0.09 (–0.83)
31		0.12	0.31	1.44	1.26 (10.42)	0.18 (–2.54)	0.30	0.26 (–2.58)	0.04 (0.57)
32		0.18	0.31	1.46	1.19 (7.68)	0.27 (–3.18)	0.53	0.43 (–3.14)	0.1 (0.91)
33	24	0.39	0.47	3.23	1.98 (11.01)	1.25 (–7.1)	0.09	0.06 (–1.11)	0.03 (0.51)
34	25	0.81	0.41	2.54	0.49 (4.6)	2.05 (–9.25)	13.77	2.66 (7.17)	11.11 (–8.99)
35	26	0.73	0.34	1.77	0.47 (4.37)	1.3 (–6.75)	11.00	2.92 (7.47)	8.08 (–7.65)
36	27	0.68	0.60	5.24	1.66 (10.35)	3.58 (–13.77)	1.07	0.34 (2.38)	0.73 (–1.99)
37	28	0.65	0.61	5.46	1.9 (11.39)	3.56 (–13.66)	0.70	0.24 (2.1)	0.46 (–1.57)
38	29	0.83	0.37	2.04	0.34 (3.86)	1.7 (–7.52)	10.18	1.72 (5.52)	8.46 (–7.05)
39		0.96	0.24	0.84	0.04 (1.16)	0.8 (–5.23)	10.06	0.44 (2.5)	9.62 (–6.47)
40	30	0.58	0.34	1.72	0.72 (5.6)	1 (–5.97)	1.00	0.42 (2.67)	0.58 (–1.94)
41		0.86	0.33	1.54	0.22 (2.9)	1.32 (–7.17)	9.31	1.3 (4.74)	8.01 (–6.18)
42		0.62	0.24	0.87	0.33 (3.62)	0.54 (–4.21)	3.36	1.27 (4.54)	2.09 (–3.78)
43		0.59	0.23	0.78	0.32 (3.39)	0.46 (–4.11)	2.30	0.94 (3.88)	1.36 (–3.04)
44		0.82	0.30	1.32	0.24 (3.03)	1.08 (–6.0)	4.55	0.82 (3.73)	3.73 (–4.03)

Note. DIF = difficulty, DISC = discrimination. MACS = Math Assessment for College Students; Item Properties (reported as difficulty: proportion of correct responses with 0 = incorrect and 1 = correct, and discrimination: corrected point-biserial correlation) and % Contribution of each item to the variance explained by dimension 1 and dimension 2 (reported as overall percent contribution, percent contribution of correct responses, and percent contribution of incorrect responses for each question).

Note. Overall percentage contribution is the sum of correct and incorrect responses for each question. Statistical significance of all items (i.e., correct and incorrect responses for each question) was determined with a bootstrap ratio [BSR] test (BSR values are reported in parentheses). Retained 30 item MACS items are included in the final version of the MACS, based on the BSR test for dimension 1 (i.e., the BSR value for both the correct and incorrect response needed to exceed the cut-off value of ± 3.75 for dimension 1, which corresponded to $p < 0.0001$). Items with high percent contributions to dimension 2 are interpreted as specific topics that students answered correctly or incorrectly, and the statistical significance of these items was determined with a bootstrap ratio [BSR] test on dimension 2 separately for each correct and incorrect response (BSR values are reported in parentheses).