Beyond gender differences:

Using tests of equivalence to evaluate gender similarities

Laura C. Ball, Robert A. Cribbie, and Jennifer R. Steele

York University

Author Note

Laura C. Ball, Robert A. Cribbie and Jennifer R. Steele, Department of Psychology, York University.

While still a student at York University, Laura C. Ball is now located in the Research and Academics Division, Waypoint Centre for Mental Health Care.

The authors would like the Jennifer Steele's graduate students for their helpful comments and suggestions during the development of this manuscript.

Correspondence concerning this article should be addressed to Laura C. Ball, Research and Academics Division, Waypoint Centre for Mental Health Care, 500 Church St., Penetanguishene, ON, Canada, L9M 1G3. Email: lball@waypointcentre.ca

Abstract

Proponents of what has been termed the Gender Similarities Hypothesis (GSH) have typically relied on meta-analyses as well as the generation of non-significant tests of mean differences to support their argument that the genders are more similar than they are different. In the present paper, we argue that alternative statistical methodologies, such as tests of equivalence, can provide more accurate (yet equally rigorous) tests of these hypotheses and therefore might serve to complement, challenge, and/or extend findings from meta-analyses. To demonstrate and test the usefulness of such procedures, we examined SAT-M data to determine the degree of similarity between genders in the historically gender-stereotyped field of mathematics. Consistent with previous findings, our results suggest that men and women performed similarly on the SAT-M for every year that we examined (1996-2009). Importantly, our statistical approach provides a greater opportunity to open a dialogue on theoretical issues surrounding what does and what should constitute a meaningful difference in intelligence and achievement. As we note in the discussion, it remains important to consider whether even very small but consistent gender differences in mean test performance could reflect stereotype threat in the testing environment and/or gender biases in the test itself that would be important to address.

Keywords: equivalence testing, statistical analysis, hypothesis testing, human sex differences

Beyond gender differences:
Using tests of equivalence to evaluate gender similarities

In 2005, Janet Hyde formally presented what is known as the Gender Similarities Hypothesis (GSH), which states that men and women are more similar than they are different in most respects. This was not a completely novel theory—the GSH has been a part of psychology for almost as long as competing theories suggesting that men and women are essentially different. Proponents of the GSH, beginning with such academics as Helen Thompson Woolley (1910) and Leta Stetter Hollingworth (1914), have tried to call attention to the similarities between the genders (see also Hyde, 2005, 2007; Lindberg, Hyde, Petersen, & Linn, 2010). By contrast, those who support the gender differences model[1] (GDM; see Hyde, 2005; Teo, 2005) have claimed the opposite, arguing instead that that men and women differ fundamentally on a number of important psychological characteristics (e.g., Irwing & Lynn, 2005; Lynn, 1992; Mau & Lynn, 2001).

An interesting piece that managed to straddle both positions is the *Psychology of Sex Differences* (1974) by Eleanor Maccoby and Carol Jacklin. It remains one of the most comprehensive studies of gender differences to date, and even though it is now over 30 years-old, it continues to be cited by leaders in the field (for example, see Eagly, 1995; Hyde, 1990, 1994, 2005; Lynn, 1992; Mau & Lynn, 2001). Using a methodology that was an early precursor to meta-analysis, Maccoby and Jacklin looked at over 1600 articles that examined gender differences across a number of domains. They wanted to document which of our beliefs about gender differences were supported by empirical evidence, as well as which were not. In the end, they concluded that only four variables repeatedly showed significant gender differences: aggression, verbal ability, visual-spatial ability, and mathematical ability. The remaining variables, which ranged from suggestibility and self-esteem to achievement motivation and learning styles, showed no significant differences. Interestingly, proponents of both GSH and GDM have used this landmark work to justify their theories (e.g., Hyde, 2005; Lynn, 1992; Mau & Lynn, 2001).

Gender Similarities Hypothesis

Following in the tradition of Maccoby and Jacklin (1974), Hyde (2005) has often made use of meta-analyses to examine gender differences. Like other feminist empiricists, she has done her work based on the belief that scientific methodology—if applied properly and without bias—will produce results that ultimately undermine the GDM (e.g., Else-Quest, Hyde, & Linn, 2010; Hyde, 1990, 1994, 2005, 2007; Hyde, Lindberg, Linn, Ellis, & Williams, 2008; Hyde & Linn, 2006; Hyde & Plant, 1995; also see, Eagly, 1995; Riger, 1992; Teo, 2005).

Hyde's (2005) article, entitled "Gender Similarities Hypothesis," can be seen as a culmination of her career's work to that point. In this piece, where she presented a meta-analysis of other meta-analyses, Hyde examined the domains where the most consistent differences between the genders had previously been found. These domains included cognitive abilities (e.g., intelligence, perceptual speed, and spatial visualization), verbal and nonverbal communication, social and personality variables (e.g., aggression, helping behaviors, and sexuality), psychological well-being (e.g., self-esteem, life satisfaction, and coping), motor behaviors, (e.g., balance, grip strength, and flexibility), and other miscellaneous constructs (e.g., moral reasoning, delay of gratification, and computer use). Through her meta-meta-analysis, Hyde showed that the majority of studies have effect sizes either close to zero ($d < 0.10$) or in the small ($0.11 > d <$

0.35) range. The exceptions to this pattern were for motor behaviors, some features of sexuality, and aggression, all of which showed larger effects.

Although it is entirely plausible that Hyde's (2005) gender similarities hypothesis is correct, we believe that further analyses are required to more fully test her argument. Specifically, in order to effectively test the GSH within the theoretical assumptions of feminist empiricism, we suggest that more appropriate (yet equally rigorous) statistical methodologies are necessary.

Tests of Equivalence

In psychology, the most frequently used statistical methods for investigating group means (*t*-tests, ANOVA *F*) determine whether there are significant mean differences between groups, not whether the means are equivalent. This is not to say that researchers are not testing hypotheses of equivalence, but rather that they employ traditional difference-based tests, even when the hypotheses relate to the equivalence of the groups. As a result, researchers interested in group similarity would have no recourse but to conclude that, with their sample, there was insufficient evidence to reject the null hypothesis. It is important to note that this does not indicate that the group means are similar or equivalent; rather, the means are simply not different enough to merit rejection of the null hypothesis. Building on research from other fields, we suggest that a different statistical formulation is required to make a stronger case for group similarity.

In order to provide statistical support for the GSH, tests of equivalence are needed. Tests of equivalence address the research question of whether groups are similar on an outcome variable. Commonly used in pharmaceutical studies, they have rarely been employed for the purposes of psychological research (Cribbie, Gruman, & Arpin-Cribbie, 2004; Rogers, Howard,

& Vessey, 1993). However, tests of equivalence have been gaining popularity in psychology and have recently been applied to, for example, assess clinical significance (e.g., Cribbie & Arpin-Cribbie, 2009; Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999; Manzoni et al., 2010) and evaluate the equivalence of paper-and-pencil and computer-based scale administrations (e.g., Lewis, Watson, & White, 2009). The most popular test of equivalence for two independent samples is the procedure developed by Donald Schuirmann (1987).

Schuirmann's (1987) test uses two simultaneous one-tailed *t*-tests. The formula is based on Student's *t* and allows the researcher to set an acceptable equivalence interval (-*D*, *D*), where *D* represents the maximum allowable difference between group means that would still be considered meaningless or inconsequential within the framework of the research. Establishing an appropriate value of *D* represents one of the most important and challenging aspects of conducting equivalence tests because it is often difficult to pinpoint the maximum difference that is not meaningful for the effect of interest. It is also very difficult to address this problem from a purely methodological viewpoint. For example, *D* values based on raw mean differences or a proportion of the standard deviation have been proposed, but these depend on the specific nature of the outcome variable (Rogers et al., 1993). It is usually more valuable for a researcher to select *D* based on consideration of the phenomena of interest, as outlined below, rather than relying on suggested intervals.

When using Schuirmann's (1987) test, an observed *mean difference* that falls within the equivalence interval would be considered unimportant within the nature of the study (Cribbie et al., 2004; Rogers et al., 1993; Stegner, Bostrom, & Greenfield, 1996; Tryon, 2001). When specifying the statistical hypotheses, group difference in either direction therefore becomes the null hypothesis; group similarity, the alternative hypothesis. For Schuirmann, the null hypothesis has two components (where μ_1 and μ_2 refer to the population means for the two groups being compared):

$$Ho_1: \mu_1 - \mu_2 \geq D$$

$$Ho_2: \mu_1 - \mu_2 \leq -D$$

In order to demonstrate that the group means are equivalent, both null hypotheses must be rejected. Rejecting both null hypotheses would imply that $\mu_1 - \mu_2$ falls within the bounds of -*D* to +*D* (Cribbie et al., 2004; Schuirmann, 1987). The formula for the degrees of freedom is the same as the formula for an independent samples *t*-test ($n_1 + n_2 – 2$).

The equations for Schuirmann's (1987) tests of equivalence are:

$$t_1 = \frac{(M_1 - M_2) - D}{\sqrt{\frac{(n_1 + n_2)[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{n_1 n_2 (n_1 + n_2 - 2)}}}$$

$$t_2 = \frac{(M_1 - M_2) - (-D)}{\sqrt{\frac{(n_1 + n_2)[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{n_1 n_2 (n_1 + n_2 - 2)}}}$$

For the purposes of this equation, *M* represents a sample mean, *-D to D* is the equivalence interval, *n* represents the group sample size, and *s* represents the sample standard deviation (Cribbie et al., 2004; Schuirmann, 1987).

Schuirmann's (1987) test has some distinct advantages for psychologists interested in providing statistical support for the GSH. Traditional statistics frame the analysis in terms of difference, and the interpretation of those statistics is usually framed in an analogous manner. By contrast, equivalence tests allow the researcher to use the language of gender similarity in the analysis of their data. This perspective is important because it provides a stronger rhetorical base for interpretation of statistics. Rather than reporting that a non-significant difference was observed and/or that the confidence interval (CI) includes zero, a researcher using equivalence testing can explicitly test whether a difference between groups falls within bounds set a priori. In addition, studies that have simply found group means to be statistically similar have had to report that there is not enough evidence to reject the null hypothesis, a conclusion that indirectly suggests similarity. Furthermore, because there is a publication bias in psychology in favor of significant results (Ferguson & Brannick, 2012; Rosenthal, 1979), making arguments based on results that failed to reach significance can lead to a publication bias in favor of the GDM. By using equivalence statistics, researchers can report more directly relevant findings for gender similarity.

Providing a Test of the GSH

Hyde's (2005) meta-meta-analysis has shown that past research has not sufficiently confirmed the existence of reliable gender differences. However, from a statistical perspective, it is also important to have tests available for determining if the male and female populations are equivalent. In general terms, one cannot fully claim that the population means are equivalent by not rejecting the traditional null hypothesis or by the presence of small effect sizes alone. We suggest that researchers interested in studying gender similarities should consider making use of additional statistical methodologies, such as tests of equivalence, to complement the findings from meta-analyses. To demonstrate the usefulness of such procedures, fewe have used the Scholastic Aptitude Test-Math (SAT-M) data to determine if men and women score equivalently on this standardized test of mathematical ability. Our goals in doing so are neither to challenge previous findings that have sought to address this question, nor to present conclusive findings in support of the GSH, but instead to suggest new avenues for future research and to open a dialogue about the theoretical and practical issues surrounding the use of statistics to explore what constitutes a meaningful gender difference.

Method

Data Collection

In an effort to provide the most complete dataset possible, we used the annual SAT archived data from the College Board which is made publicly available on their website. We also chose the math section specifically for our analysis because it was the only section of the test that was not affected by the restructuring of the SATs in 2006. Until 2005, the test was divided into Verbal and Mathematics sections, whereas in 2006 it was changed to reflect a new tripartite structure: Critical Reading, Mathematics, and Writing. The Mathematics section is the only part which remained unchanged at that time (College Board, 2009). In addition, the SAT-M is perhaps the most contentious of the sections, following in the wake of former Harvard president Lawrence Summers' (2005) comments on the underrepresentation of women in mathematics, engineering, and the sciences (see discussion in Spelke, 2005; Spelke & Ellison, 2009; Spelke & Grace, 2006).

The archived data from the College Board had several advantages over the published literature that would typically be used in a meta-analysis. First, the SAT data represent an extremely large dataset which is readily available to the public on their website.[2] Because we were primarily interested in testing the utility of this statistical approach, the archived data from the College Board allowed for an open dataset and transparent analyses. Second, we found the published literature on intelligence, aptitude, and scholastic achievement to be too varied to be useful for our purposes. The populations tested were dissimilar across any number of demographic variables (e.g., age, SES, and IQ), and the methods of testing (SAT, WISC/WAIS, Raven's Matrices, Stanford-Binet), and provision of information we required (*n*s, *M*s, and *SD*s) were inconsistent. Therefore, the published peer-reviewed literature was not a good fit given our main purpose; by contrast the comprehensive and public data from the College Board seemed ideal.

Procedure

We compiled the overall number of men and women who completed the SAT-M portion of the test for each year in which archival SAT data were available (1996-2009). We also gleaned the mean score and standard deviation for each gender from these reports. Next, we wrote a program for the R statistical package (R Development Core Team, 2011) that was designed to assess the similarity of the data from males and females for each year using Schuirmann’s (1987) two one-sided test procedures.

In order to conduct Schuirmann’s (1987) test, we chose three comparison points for assessing group similarity (small, medium, and large). Given the uniqueness of using equivalence testing on these kinds of data, however, there was no past literature from which to draw appropriate equivalence intervals. Because ours is an exploratory study designed to demonstrate the usefulness of a procedure, rather than a study designed to produce conclusive findings on the subject, we somewhat arbitrarily set the testing intervals at 1/3, 2/3, and 1 *SD*, where *SD* represents the average of the males’ and females’ standard deviations. The arbitrariness of choosing these particular numbers is a point to which we will return in the discussion. However, to aid researchers facing the difficult task of setting an appropriate equivalence interval, two important points can be made. First, it should be obvious that what constitutes a meaningful difference is not generally an obvious quantity. Differences fall on a continuum, making the selection of a logical equivalence interval a challenging task. Second, when establishing an appropriate interval, it is necessary to consider what would be the largest difference that would be inconsequential. In other words, what is the largest difference between the population means that would not be meaningful within the context of the study.

When setting an appropriate interval for comparing males and females on the SAT-M, considerations might include the effect size, percentage of distribution overlap, percentage of each sex that is in the top/bottom 5%, and so on for each potential interval. For example, from Whissell (2003), a comparison with an effect size of d = .33 (1/3 *SD*), there is about 97% shared variability (i.e., $1\text{-}r^2$), whereas for a comparison with an effect size of d = 1 (1 *SD*), there is about 80% shared variability. With regard to the equivalence tests conducted in our study (and any null hypothesis significance test for that matter), we note that low power is not an issue of concern because of the large sample sizes we included. For the test statistics computed below, the sample *SD*s (which were slightly larger than the normed *SD* of 100 provided by the College Board) were utilized.

Results

As can be seen in Table 1, between 1996 and 2009, women had SAT-M mean scores ranging from 492 to 504, with an overall mean of 498.64 (*SD* = 3.46; *SE* = 0.92). Men, on the other hand, had SAT-M mean scores between 527 and 538, with an overall mean score of 533.36 (*SD* = 3.03; *SE* = 0.81).

Table 2 shows the correlations between gender and SAT-M scores for each year (r), the proportion of variance in scores explained by gender (r^2), the amount of overlap in scores between the genders (λ), and the effect size for the correlation (d). We found that slightly more than 2% of the variance in scores each year was accounted for by gender, which amounted to slightly less than 98% overlap between the two populations. As could be expected based on previous meta-analyses, we found that the effect size in every case was in the small range (*ds* $\leq$

.32). In terms of how these results should be read, an article by Whissell (2003, p. 720) examined the implications of correlations where the effect sizes fell into the small range ($d \leq .33$), stating that such comparisons have “very little explanatory power.” This conclusion supports the notion that men and women perform similarly on the SAT-M. In the analyses that follow, we build upon these results.

Table 3 presents the results of the equivalence tests on the SAT-M data, with alpha set at .05. The columns of Table 3 represent the three levels of equivalence intervals that were used (1/3, 2/3 and 1 *SDs* of the average scores for males and females). The rows of the table represent the different years that we analyzed, and t_1 and t_2 represent the two *t*-tests used in Schuirmann’s (1987) procedure (described above). Through the comparison of the yearly SAT data, we found that men and women were consistently similar at all of our chosen intervals. This effect was significant in each of the analyses ($|t|$s > 1.66, *p*s < .05) and was consistent for each of the yearly datasets. Because the datasets were found to be consistently equivalent at 1/3 *SD* (the smallest testing interval), there was no need to test further.

We should note that, although the math section of the SAT has undergone fewer changes than other sections (e.g., verbal), certain items on the SAT-M that showed a female advantage were eliminated in 2005, and items showing a male advantage remained untouched (Chipman, 2005; Spelke, 2005). As such, it is possible that our results from the post-2004 data could have reflected this change, making it appear that the mathematical ability of women is on the decline. Interestingly, there was no change in the pattern of the data in 2005 and any year thereafter, despite the gender-biased revisions.

Discussion

The results of our study were consistent with past research that has documented only trivial differences between men and women on the SAT-M (Casey, Nuttall, & Pezaris, 1999; Hyde, 2005; Spelke, 2005). More importantly, ours is the first known study to examine whether men and women perform similarly on the SAT-M based on tests of equivalence. In the present analyses, we found evidence that indeed men's and women's performance on the SAT-M from 1996 to 2005 was statistically equivalent. This finding is consistent with the GSH and extends previous findings from this literature. More importantly, in line with the primary goal of our study, the present analyses provide another statistical tool that can be used when considering whether gender differences, as well as directly tested equivalence, exist in a given domain.

Although our finding is consistent with the GSH, it is important to emphasize that the main goal of our study was to demonstrate the utility of equivalence testing for psychological studies of gender similarities. Because this is a relatively new approach to data analysis for psychologists, the three equivalence intervals were somewhat arbitrarily selected. Although, in a variety of real-world settings, differences that are less than one-third of a *SD* are considered trivial, in some instances small effect sizes can have different forms of importance that cannot be dismissed through statistical tests of difference or equivalence (e.g., Prentice & Miller, 1992), and small effect sizes – when compounded over time – may be considered a meaningful result (Martell, Lane & Emrich, 1996). For example, with the SAT-M, the difference between males and females is consistently around 35 points, and therefore it is highly unlikely that the 35 point difference reflects random variability. The equivalence tests conducted in our paper have demonstrated that this difference, relative to the standard deviations, is small; but, it remains important that future research using these and other statistical tools consider what the threshold is for a meaningful difference, particularly when considering potential gender differences in ability or achievement. As we discussed earlier, establishing an appropriate equivalence interval is one of the most challenging aspects of equivalence testing. However, we anticipate that this choice will present less of a challenge as more researchers start utilizing equivalence testing in their studies and ensuing discussions surrounding appropriate intervals increase.

When discussing what constitutes a meaningful difference to use when examining standardized test scores, such as the SAT, it is also important to consider that very small gender differences in mean test performance—particularly at the high end of performance—could be the result of biases in the testing environment and/or biases in the test items. Furthermore, if those differences are seemingly trivial, yet consistent and unacknowledged, changes to the test or the testing environment which might eliminate such differences will not be made. For example, as discussed above, the average performance of men was consistently numerically higher than the average performance of women every year. What can we take from this discrepancy? Although there could be a host of contributing factors, extensive research on stereotype threat suggests that women who are highly identified with mathematics can underperform in contexts that make their gender salient (Davies & Spencer, 2005; Good, Aronson, & Harder, 2008; Spencer, C. M. Steele, & Quinn, 1999; C. M. Steele, Spencer, Aronson, 2002; J. R. Steele, Reisz, Williams, & Kawakami, 2007; Walton & Spencer, 2009). It is therefore plausible that mathematically-gifted women underperformed relative to their potential in some testing environments due to the activation of gender stereotypes. This possibility is in line with the findings of a recent meta-analysis by Walton and Spencer (2009) that found that measures of mathematical ability, such as the SAT, reveal a consistent bias against negatively stereotyped minority groups (including women) that can be accounted for by stereotype threat. In the analyses by Walton and Spencer, stereotype threat was found to account for approximately one-fifth of a standard deviation (approximately 19-21 points) of the gender differences between scores on the SAT-M.

If differences in SAT-M achievement are seen as analogous to differences in ability, and some women are underperforming on this test because gender stereotypes are evoked, then women may be disproportionately denied scholarships, entrance to competitive science programs and universities, and later to competitive and sometimes lucrative math-based occupations. They may also face discrimination throughout their education in mathematics and in the workplace (Else-Quest et al., 2010; see also Nosek et al., 2009). It is therefore important that researchers who are attempting to compare the abilities of men and women use the statistical tool that best complements their research hypotheses. We also believe that, in the future, equivalence testing might be a useful technique for comparing males and females at the high end of the ability spectrum—in order to test what has become known as the *variability hypothesis* (see Hyde, 2005) by examining whether the math performance of high achieving men and women is equivalent.

In this study, we have demonstrated the utility of tests of equivalence, which offer important contributions if used along with traditional tests for evaluating the similarity of groups on outcome variables. However, as with other statistical tests, it is important that consideration be given to what constitutes a meaningful difference. The study of potential gender differences in math abilities and other variables is—and will continue to be—a topic of interest to many researchers and the general public alike, with tests of equivalence offering a novel way of approaching the question. It is also important to point out that, although we have presented an innovative null hypothesis testing approach for investigating gender comparisons, very valuable information about gender comparisons has been gained through meta-analytic research, where the focus is on the size of statistical effects. In fact, as highlighted by an anonymous reviewer of our manuscript, it might be less productive to debate whether there are differences or similarities between genders and instead just focus on the effect sizes that are found in gender comparison research. Research that summarizes the sizes of effects over multiple studies is invaluable, and we strongly support the use of meta-analytic approaches for investigating and summarizing research related to gender comparisons. Equivalence testing approaches are in no way meant as a replacement for meta-analytic research, but instead offer an appropriate statistical procedure for addressing whether groups of subjects are equivalent on an outcome—a potentially valuable additional piece of information.

Our goal with the current analyses was to illustrate an additional method for examining gender similarities. It is our hope that we have shown the utility of statistical tests such as Schuirmann's (1987) equivalence testing procedure for researchers who are interested in this topic. In order to improve accessibility to the methods described in our paper, an R program was developed that will conduct the equivalence testing procedure discussed in this paper and is available at http://www.?.? (omitted for blind review). R is open-source software that is free to use and available on almost any platform (e.g., Mac, PC, Linux). Although more debate may be needed to determine what the most useful thresholds for determining equivalence might be, adding such approaches to the statistical toolbox will help strengthen the statements that proponents of the gender similarities hypothesis can make. The result can only be a better and stronger science.

Practice Implications

Numerous research studies have been conducted with the goal of comparing males and females on a specific outcome, and the results of these studies often have important policy implications. This paper highlights the need to consider how the statistical tools chosen relate to the hypotheses and conclusions of a study. This is an important factor to consider because selecting the wrong statistical method, for example evaluating mean equivalence with a test designed to detect differences in means, can lead to misleading or incorrect conclusions. Although the obvious practical implication of this paper is for researchers to become more diligent at identifying the appropriate test statistics given their specific hypotheses, at a more basic level we need to improve the way we teach undergraduate and graduate statistical methods both in methodological courses and in applied courses such as the Psychology of Women/Gender. The methods that we choose for our analyses can substantially alter the conclusions of the study, so making upcoming researchers aware of novel or unfamiliar methods for data analysis will have important implications for the conclusions drawn from future studies.

References

Casey, M. B., Nuttall, R. L., & Pezaris, E. (1997). Mediators of gender differences in mathematics college entrance test scores: A comparison of spatial skills with internalized beliefs and anxieties. *Developmental Psychology, 33*, 669-680. doi:10.1037/0012-1649.33.4.669

Chipman, S. F. (2005). Research on the women and mathematics issue: A personal case history. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 1-24). New York, NY: Cambridge University Press.

College Board. (2009). *Archived SAT data and reports*. Retrieved from http://research.collegeboard.org/ content/archived-data

Cribbie, R. A., & Arpin-Cribbie, C. A. (2009). Evaluating clinical significance through equivalence testing: Extending the normative comparisons approach. *Psychotherapy Research, 19*, 677-686. doi:10.1080/10503300902926554

Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology, 60,* 1-10. doi:10.1002/jclp.10217

Davies, P. J., & Spencer, S. J. (2005). The gender-gap artifact: Women's underperformance in quantitative domains through the lens of stereotype threat. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 172-188). New York, NY: Cambridge University Press.

Eagly, A. (1995). The science and politics of comparing women and men. *American Psychologist, 50*, 145-158. doi:10.1037/0003-066X.50.3.145

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*, 103-127. doi:10.1037/a0018053

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Cambridge, MA: Harvard University Press.

Good, C., Aronson, J., & Harder, J. A. (2008). Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses. *Journal of Applied Developmental Psychology, 29*, 17-28. doi:10.1016/j.appdev.2007.10.004

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*, 120-128. doi:10.1037/a0024445

Hollingworth, L. S. (1914). Variability as related to sex differences in achievement: A critique. *American Journal of Sociology, 19*(4), 510-530.

Hyde, J. S. (1990). Meta-analysis and the psychology of gender differences. *Signs, 16*, 55-73. doi:10.1086/494645

Hyde, J. S. (1994). Should psychologists study gender differences? Yes, with some guidelines. *Feminism and Psychology, 4*, 507-512. doi:10.1177/0959353594044004

Hyde, J. S. (2005). Gender similarities hypothesis. *American Psychologist, 60*, 581-592. doi:10.1037/0003-066X.60.6.581

Hyde, J. S. (2007). New directions in the study of gender similarities and differences. *Current Directions in Psychological Science, 16*, 259-263. doi:10.1111/j.1467-8721.2007.00516.x

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, *321*, 494-495. doi:10.1126/science.1160364

Hyde, J. S., & Linn. M. C. (2006). Gender similarities in mathematics and science. *Science, 314*, 599-600. doi:10.1126/science.1132154

Hyde, J. S., & Plant, E. A. (1995). Magnitude of psychological gender differences: Another side to the story. *American Psychologist, 50*, 159-161. doi:10.1037/0003-066X.50.3.159

Irwing, P., & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta-analysis. *British Journal of Psychology, 96*, 505-524. doi:10.1348/000712605X53542

Jackson, D. N., & Rushton, J. P. (2006). Males have g: Sex differences in general mental ability from 100,000 17- to 18-year-olds on the Scholastic Assessment Test. *Intelligence, 34*, 479-489. doi:10.1016/j.intell.2006.03.005

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 285-299. doi:10.1037/0022-006X.67.3.285

Lewis, I. M., Watson, B. C., & White, K. M. (2009) Internet versus paper-and-pencil survey methods in psychological experiments: Equivalence testing of participant responses to health-related messages. *Australian Journal of Psychology*, *61*, 107-116. doi:10.1080/00049530802105865

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*, 1123-1135. doi:10.1037/a0021276

Lynn, R. (1992). Sex differences in the Differential Aptitude Test. *Educational Psychology, 12*, 101-102. doi:10.1080/0144341920120201

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.

Manzoni, G. M., Cribbie, R. A., Villa, V., Arpin-Cribbie, C. A., Gondoni, L., & Castelnuovo, G. (2010). Psychological well-being in obese inpatients with ischemic heart disease at entry and at discharge from a four-week cardiac rehabilitation program. *Frontiers in Psychology for Clinical Settings, 1*(38), 1-6.

Martell, R. F., Lane, D. M., & Emrich, C. (1996). Male-female differences: A computer simulation. *American Psychologist, 51*, 157-158. doi:10.1037/0003-066X.51.2.157

Mau, W. C., & Lynn, R. (2001). Gender differences on the Scholastic Aptitude Test, the American College Test and college grades. *Educational Psychology, 21*, 133-136. doi:10.1080/01443410020043832

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., … Greenwald, A. G. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences, 106,* 10593-10597. doi:10.1073/pnas.0809921106

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160-164. doi:10.1037/0033-2909.112.1.160

R Development Core Team. (2011). R: A language and environment for statistical computing [software]. Retrieved from http://www.R-project.org/

R Development Core Team. (2011). [software]

Riger, S. (1992). Epistemological debates, feminist voices: Science, social values, and the study of women. *American Psychologist*, *47,* 730-740. doi:10.1037/0003-066X.47.6.730

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113,* 553-565. doi:10.1037/0033-2909.113.3.553

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638-641. doi:10.1037/0033-2909.86.3.638

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15,* 657-680. doi:10.1007/BF01068419

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science? A critical review. *American Psychologist*, *60*, 950-958. doi:10.1037/0003-066X.60.9.950

Spelke, E. S., & Ellison, K. (2009) Gender, math and science. In C. H. Sommers (Ed.), *The science on women and science* (pp. 24-53). Washington, DC: AEI Press.

Spelke, E. S., & Grace, A. D. (2006). Sex, math, and science. In S. Ceci & W. Williams (Eds.), *Why aren't more women in science? Top gender researchers debate the evidence* (pp. 57-67). Washington, DC: APA Publications.

Spencer, S. J., Steele, C. M., & Quinn, D. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*, 4-28. doi:10.1006/jesp.1998.1373

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 379-440). San Diego, CA: Academic Press.

Steele, J. R., Reisz, L., Williams, A., & Kawakami, K. (2007). Women in mathematics: Examining the hidden barriers that gender stereotypes can impose. In R. Burke & M. Mattis (Eds.), *Women and minorities in science, technology, engineering and mathematics: Upping the numbers* (pp.159-183). London, UK: Edward Elgar.

Stegner, B. L., Bostrom, A. G., & Greenfield, T. K. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. *Evaluation and Program Planning, 19*, 193-198. doi:10.1016/0149-7189(96)00011-0

Summers, L. (2005, January 14). *Remarks at NBER conference on diversifying the science and engineering workforce*. Retrieved from http://www.president.arvard.edu/speeches/2005/nber.html

Teo, T. (2005). *The critique of psychology: From Kant to postcolonial theory*. New York, NY: Springer.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*, 371-386. doi:10.1037//1082-989X.6.4.371

Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science, 20*, 1132-1139. doi:10.1111/j.1467-9280.2009.02417.x

Whissell, C. (2003). What difference does it make? Implications of the size of the difference between the mean of two groups. *Perceptual and Motor Skills, 97*, 716-722. doi:10.2466/PMS.97.7.716-722

Woolley, H. T. (1910). A review of the recent literature on the psychology of sex. *Psychological Bulletin*, *7*, 335-342.

Footnotes

[1] The use of the term “gender differences model” should not be interpreted as a reference to a formally proposed framework, theory, model, or position. Rather, it is meant as a loose descriptor for a collection of work that draws on the assumption of gender differences on a wide array of variables. This could include work stemming from evolutionary psychology (e.g., Jackson & Rushton, 2006) all the way to work on feminist standpoint theory (e.g., Gilligan, 1982).

[2] It is worth noting, however, that unlike other large datasets produced by standardized tests, such as those that might be administered to all students enrolled in public high schools (e.g., the National Assessment of Educational Progress; see Else-Quest et al., 2010 for additional examples), only a select subsample of students complete the SAT. Therefore, there is likely a sampling bias present in the dataset.

Table 1

Compiled Data from the SAT-M

Year		*n*	*M*	*SD*
1996	Male	504598	527	115
	Female	580127	492	107
1997	Male	520338	530	114
	Female	606683	494	108
1998	Male	541692	531	114
	Female	630817	496	108
1999	Male	562911	531	115
	Female	657219	495	110
2000	Male	583331	533	115
	Female	676947	498	109
2001	Male	592366	533	115
	Female	683954	498	109
2002	Male	616201	534	116
	Female	711630	500	110
2003	Male	652606	537	116
	Female	753718	503	111
2004	Male	660270	537	116
	Female	758737	501	110
2005	Male	686298	538	116
	Female	789325	504	111
2006	Male	680725	536	117
	Female	785019	502	111
2007	Male	690500	533	116
	Female	798030	499	110
2008	Male	704226	533	118
	Female	812764	500	111
2009	Male	711368	534	118
	Female	818760	499	112

Note. *n* = number of SAT-M test takers; *M* = mean score; *SD* = standard deviation.

Table 2

SAT-M Scores from 1996-2009

	r	r^2	λ	d
SAT-M (1996)	.156	.024	.976	.32
SAT-M (1997)	.160	.026	.974	.32
SAT-M (1998)	.156	.024	.976	.32
SAT-M (1999)	.158	.025	.975	.32
SAT-M (2000)	.154	.024	.976	.31
SAT-M (2001)	.154	.024	.976	.31
SAT-M (2002)	.149	.022	.978	.30
SAT-M (2003)	.148	.022	.978	.30
SAT-M (2004)	.157	.025	.975	.32
SAT-M (2005)	.148	.022	.978	.30
SAT-M (2006)	.147	.022	.978	.30
SAT-M (2007)	.149	.022	.978	.30
SAT-M (2008)	.143	.020	.980	.29
SAT-M (2009)	.150	.022	.977	.30

Note. r = correlation coefficient; r^2 = variability in scores accounted for by gender; λ = measure of overlap between populations; d = effect size.

Table 3

Equivalence Testing of SAT-M Scores

	1/3*SD*		2/3*SD*		1*SD*	
Year	t_1	t_2	t_1	t_2	t_1	t_2
1996	-7.61	334	-178	505	-355	681
1997	-3.00	345	-177	520	-357	700
1998	-7.91	348	-186	525	-369	708
1999	-5.50	357	-187	539	-374	725
2000	-9.77	359	-194	543	-384	733
2001	-9.84	361	-195	547	-386	738
2002	-16.7	362	-206	551	-401	746
2003	-18.0	372	-213	566	-413	767
2004	-6.77	385	-202	580	-404	782
2005	-18.4	381	-218	580	-424	786
2006	-19.1	379	-218	577	-423	782
2007	-17.7	383	-218	583	-424	790
2008	-25.6	379	-228	581	-436	789
2009	-15.8	391	-219	594	-428	803

Note. t_1 and t_2 represent the two *t*-tests used in Schuirmann's (1987) test of equivalence. All *t*s were significant at $p \leq .01$, which means that males' and females' math test performances on the SAT-M were equivalent at all of our chosen intervals.