Bayesian Analysis in Educational Psychology Research:

An Example of Gender Differences in Achievement Goals

Steven Peterson and Avi Kaplan

College of Education, Temple University

Accepted for publication in *Learning & Individual Differences*

Correspondence regarding this manuscript should be sent to Steven Peterson at steve@wikipeterson.org

Bayesian Analysis in Educational Psychology Research:

An Example of Gender Differences in Achievement Goals

Abstract

Much research in educational psychology concerns group differences. In this study, we argue that Bayesian estimation is more appropriate for testing group differences than is the traditional null hypothesis significance testing (NHST). We demonstrate the use of Bayesian estimation on gender differences in students' achievement goals. Research findings on gender differences in achievement goals have been mixed. We explain how Bayesian estimation of mean differences is more intuitive, informative, and coherent in comparison with NHST, how it overcomes structural and interpretive problems of NHST, and how it offers a way to achieve cumulative progress toward increasing precision in estimating gender differences in achievement goals. We provide an empirical demonstration by comparing a Bayesian and a traditional NHST analysis of gender differences in achievement goals among 442 7^{th}-grade students (223 girls and 219 boys). Whereas findings from the two analyses indicate comparable results of higher endorsement of mastery goals among girls and higher endorsement of performance-approach and avoidance goals among boys, it is the Bayesian analysis rather than the NHST that is more intuitively interpreted. We conclude by discussing the perceived disadvantages of Bayesian estimation, and some ways in which a consideration of Bayesian probability can aid interpretations of traditional analytical methods.

Bayesian Analysis in Educational Psychology Research: An Example of Gender Differences in Achievement Goals

Gender differences in academic motivation have been of interest for researchers aiming to explain differences between girls and boys in academic decision-making and performance. Researchers have sought to understand, for example, why boys and girls elect different courses of study and perform at different levels in language arts and in math and science (Eccles, 1983). Research in the past three decades has fruitfully investigated gender differences in perceived abilities and also in task values (Wigfield & Eccles, 2002). However, research findings have been much less consistent regarding gender differences in the motivational orientations that students adopt *for studying* in different domains—their achievement goals—leading to uncertainty regarding gender differences in these important motivational processes that have been related to quality engagement, development of interests, and performance (Hulleman et al., 2010; Linnenbrink-Garcia et al., 2008). We propose that one reason for the uncertainty may be the reliance of researchers on normative Null Hypothesis Significance Testing (NHST) as the primary method for drawing conclusions about gender differences from the data. In this paper, we illustrate interpretive and structural problems with traditional *t* tests. In addition, we discuss how these problems may be addressed by employing a Bayesian analysis as an alternative method for understanding gender differences within the framework of achievement goal theory. We illustrate the use of Bayesian analysis to investigate gender differences in achievement goals among a sample of Junior-High school students.

Gender Differences in Achievement Goals

Achievement goal theory is an important perspective for understanding student motivation in school (Ames, 1992; Elliot, 2005; Nicholls, 1989). Researchers distinguish

between three primary achievement goals: mastery-approach, performance-approach, and performance-avoidance goals[1]. Mastery-approach goals refer to a focus on development of competence, have been found to be associated with adaptive patterns of learning including self-regulation, persistence, and preference for challenging activities, and are considered desirable motivational goals (Maehr & Zusho, 2009). Performance-approach goals refer to a focus on demonstrating high competence, particularly relative to others. This motivational orientation has been associated with some positive patterns of learning, such as high efficacy and achievement, which have been associated with the normative comparison goal, but also with somewhat less positive patterns such as disruptive behavior and unwillingness to cooperate, which have been associated with the demonstration of ability goal (Hulleman et al., 2010; Kaplan & Maehr, 2007; Senko, Hulleman & Harackiewicz, 2011). Performance-avoidance goals refer to a focus on avoiding demonstrating low competence, particularly relative to others, and have been commonly associated with maladaptive patterns of learning, including low efficacy, negative emotions, self-handicapping strategies, and low performance (Kaplan & Maehr, 2007; Maehr & Zusho, 2009).

Despite the meaningful association of achievement goals with academic outcomes, researchers have failed to identify differences between boys and girls in achievement goals that would help explain gender differences in academic patterns such as performance in math versus language arts. Some studies concerning gender differences in achievement goals report that girls are more mastery-oriented and less performance-oriented than boys are (e.g., Anderman & Young, 1994). Yet, Meece and Jones (1996) reported that boys in the low-ability groups were

[1] An additional achievement goal, mastery-avoidance goals, has been added to the more prevalently studied three mentioned here. The conceptual meaning and prevalence of this motivational orientation among young students is still under investigation (Madjar et al., 2011), and it was not included in the current study. For brevity, mastery-approach goals in the current manuscript are labeled simply mastery goals.

more mastery-oriented than girls were. Some studies find no difference between the genders (e.g., Greene, DeBacker, Ravindran, & Krows, 1999), or gender differences in one ethnic group but not in another (e.g., Middleton & Midgley, 1997). In their review of this literature, Meece, Glienke, and Burg (2006) concluded that there was "no clear pattern of gender differences in students' achievement goal orientations" (p. 360) and that gender differences (when they are detected at all), are moderated by race, ability, age, and classroom context.

One potential reason for the state of uncertainty regarding gender differences in achievement goals is the reliance on NHST. While NHST is the most prevalent statistical analysis in educational research (and social science research more broadly), the literature has emphasized its structural and interpretative problems (e.g., Rozeboom, 1960; Cohen, 1994; McLean & Ernest, 1998; Dienes, 2011) with a lamentable influence on normative practice (Gigerenzer & Sedlmeier, 1989; Halk & Greenbaum, 1995, Lecoutre, 2006). In the next sections, we elaborate on these critiques and their meaning to investigating gender differences in achievement goals. We then present an alternative approach to the analysis of mean differences that overcomes many of these issues—Bayesian estimation.

Null Hypothesis Significance Testing (NHST)

NHST refers to the orthodox practice of assessing the evidence against a null hypothesis by first assuming that the null hypothesis in question is true and then by comparing the data actually observed to hypothetical data that could have been observed if the researchers repeatedly drew random samples from the population. The evidence is measured using a *p*-value—the probability of getting a result at least as extreme as that observed assuming that the null hypothesis is true. Ronald Fisher promoted the use of such probabilities (*p*-values) as measures of statistical significance, i.e., the extent to which the observed data are inconsistent

with the null hypothesis (Fisher, 1935). In practice, a *p*-value smaller than .05 is commonly taken to indicate statistical significance.

Jerzy Neyman (the inventor of the confidence interval [Neyman, 1937]) and Egon Pearson developed an alternative procedure based on Type I and Type II error rates and comparing a null model and a specified alternative model for the data (Neyman & Pearson, 1933). In a Neyman-Pearson test, there are two hypothesized models for the data, H_O and H_A. A Type I error rate is chosen, and a variety of tests are evaluated. The test that minimizes the probability of a Type II error (and maximizes power) is chosen (Christensen, 2005).

Education researchers today compute *p*-values to measure the strength of evidence against the null hypothesis, and they are encouraged to also consider Type I errors, Type II errors, and power (Huck, 2007). While Neyman/Pearson and Fisher were each in turn critical of one another's approaches (Berger, 2003), "current practice has become an amalgamation of the two incompatible theories" (Wagenmakers, 2007).

Comparing Mean Differences Using NHST

In traditional practice, researchers assess the degree to which the data support a null hypothesis of exactly equal population means across the compared groups. However, even before collecting the data, a hypothesis of exactly equal population means is generally not realistic or plausible (Cohen, 1994; Wagenmakers, 2007; Gelman & Tuerlinckx, 2000; Gelman, Hill, & Yajima, 2012). We can generally assume that differences do exist even if they are extremely small and of no practical importance. Rather than asking *is there a difference*?, questions that may be more relevant to researchers investigating group differences are *what is the direction of the difference?* and *how large is the difference?* NHST can help establish confidence in the *direction* of a difference between the groups or leave us uncertain about the direction (Tukey,

1991). That is, rejection of the null hypothesis supports confidence with regard to which group mean is larger, Though a failure to reject the null hypothesis is commonly interpreted as supporting the conclusion that no difference exists (McLean & Ernest, 1998; Wainer & Robinson, 2003), it should be taken to indicate diffidence about the direction of the difference.

For example, in their study, Meece and Jones (1996) (who did not have access to modern Bayesian methods for data analysis) concluded that "There were no main effects for gender on any of the motivation scales" (p. 400). Importantly, the inability to make a conclusion about a difference on the basis of NHST is not the same as finding that there is no difference. In fact, Meece and Jones's (1996) summary statistics do include differences in achievement goals by gender, but the statistical analysis suggested that the differences were not large enough relative to the sample size in the study to rule out chance variation as a plausible explanation. (It is noteworthy that it is always possible to employ a sample size that is small enough to ensure the failure of detecting differences using NHST.)

Meece and Jones's (1996) report of no main effect for gender was also accompanied by a reported interaction between gender and ability level. Using a traditional procedure for mitigating false alarms in making multiple comparisons[2], they found that boys of low ability had stronger mastery goal orientations than girls did, but that "There were no gender differences in students' mastery orientation among average- and high-ability students" (p. 401). Again, the findings indicated that, in fact, there were differences, but that these were too small to reach

[2] When comparing subgroups (e.g., high-ability males versus high-ability females), statistical power decreases along with the sample sizes, and a penalty is paid for each comparison the researcher intends to make. Kruscke (2013) points out that since the penalties paid in multiple comparison procedures depend (inappropriately) on the researchers' subjective intentions, a researcher motivated to do so could make any observed difference no matter how large statistically *non*significant just by choosing to earnestly *intend* to collect data on enough additional groups and to make additional comparisons at some time in the future.

significance with the sample size in the study. Thus, an accurate description of these findings would be that the researchers were unable to establish confidence in the direction and size of the differences.

Critiques of NHST

NHST has been criticized repeatedly on several grounds (see Christiansen, 2005, Cohen, 1994; Falk and Greenbaum, 1995; Lecoutre, 2006; Wagenmakers, 2007). For example, Cohen (1994) contends that NHST does not test what researchers want to know. While most researchers turn to statistics to ascertain the probability that a certain hypothesis is true in light of the observed data, the NHST *p*-value indicates the probability of obtaining the observed data while assuming that the null hypothesis is true. The fact that researchers would prefer to know the former rather than the latter is evidenced by the fact that *p*-values are so often incorrectly interpreted as the probability that the null hypothesis is true. Whereas the widespread misconceptions among students, teachers of statistics, and applied statisticians regarding the interpretation of NHST may be thought of as indicating a lack of individual mastery of the subject, Lecoutre (2006) and Falk and Greenbaum (1995) argue that the persistence of these misconceptions is a result of the failure of NHST to answer the questions that researchers actually want answered.

Two perspectives on probability

The conception of probability underlying NHST is called frequentism. The frequentist interpretation of the probability of occurrence for an event is the long-term relative frequency of its occurrence. When using NHST, we make a decision based on a p-value—on the *probability of the data under the assumption that the null hypothesis is true*—rather than on the more desired and intuitively appealing option—*the probability that the null hypothesis is false given the data.*

This is done because, from the perspective of frequentism, the latter is simply an incoherent concept. For a frequentist, while samples can be regarded as random (and we can therefore attach probabilities to getting particular results in hypothetical repeated sampling or repeated random assignments of subjects to treatments), population mean differences are not random. They are facts about the world—comprising unknown but fixed quantities. From this perspective, a hypothesis, which constitutes a claim about an unknown but fixed quantity, can be either true or false, with the probability of its truth either 0 or 1. For the frequentist—and hence, within the framework of NHST—claims about the likelihood that the research hypothesis is true are nonsensical.

The Bayesian perspective provides an alternative to the frequentist paradigm. Its premise is the recognition that subjective beliefs about certainty constitute valid probabilities and can be measured according to the mathematics of probability. By treating probabilities as indicating degrees of belief, one can make warranted claims about the probability that a meaningful group difference obtains.

Bayesian estimation begins with the mathematical description of the state of prior knowledge in the absence of new data—what the scientific community may believe to be true in light of previous findings. This description should be made in a way that would satisfy a skeptical audience as in peer review. Then, by applying Bayes' Theorem, researchers can determine how the beliefs ought to be updated in the light of the new data. The need to rely on prior knowledge has been a common point of criticism of Bayesian statistics by the frequentist camp. However, reliance on such professional judgment of the current state of knowledge is inevitable according to the mathematics of probability. There is no logical way for going directly from data to warranted statements about the state of affairs without considering prior knowledge.

Such professional judgment is required in NHST, for example, in deciding what null hypotheses to test and what to conclude about a hypothesis based on a *p*-value that, in itself, says nothing directly about the truth value of the tested hypothesis. An important advantage of this approach is having a mathematically warranted foundation for intuitively meaningful probability statements in research findings (Diennes, 2011) of the sort that are routinely *but erroneously* made on the basis of analyses using NHST.

Erroneous interpretations of NHST

Whereas NHST actually provides *the probability of the data given the hypothesis*, Bayes' Theorem allows a conclusion regarding *the probability of the hypothesis given the data*. However, researchers often misinterpret findings from NHST as indicating the latter. Such an interpretation—that a test based on a significance level of .05 has only a 5% chance of a false alarm—is highly suspect, especially for tests with low power. To illustrate, suppose that we have collected data on endorsements of performance-avoidance goals for boys and girls, and we plan to conduct a traditional *t* test traditionally used to test for the existence of a non-zero difference in population means. Suppose further that based on prior literature, it is believed that a non-zero difference between boys and girls on performance-avoidance goals is unlikely—say, it has a 10% chance of obtaining. We set to collect data from a sample that allows us to run a (not so unusual) low-powered test, where we have only a 40% chance of rejecting the null hypothesis if the difference between the means of boys and girls in performance-avoidance is in fact non-zero. Using the orthodox .05 significance threshold (which allows that in 5% of cases where the null hypothesis is true, the null hypothesis will nevertheless be rejected), suppose that the data lead us to reject the null hypothesis (of no difference between the gender groups). What is the actual probability of a false alarm under these conditions?

If we were to test a set of 1000 mean differences in which only 100 (10%) of them are non-zero, with power at 40%, we will detect only 40 out of those 100 differences. (See Table 1.) Using the 5% significance level, we will incorrectly reject the null hypothesis in 45 of the 900 cases where a difference does *not* exist. Thus, out of the 40 + 45 = 85 tests where the null hypothesis is rejected, there is a difference in only 40 or about 47% of them. Therefore, our false alarm rate is much larger than that assumed in the routine erroneous interpretation of NHST. The actual false alarm rate is about 53%, not at all the same as the 5% many researchers expect when basing their decisions on a 5% significance level. In fact, if we were doubtful about the existence of a substantial gender difference before collecting the data, when finding a statistically significant result in a low-powered NHST, we should still believe that the results is more likely than not just a false alarm.

In short, $p < .05$ does not necessarily mean that the null hypothesis has less than a 5% chance of being true or even that it is most likely false. Likewise, $p > .05$ does not mean that the null hypothesis is most likely true. *p*-values do not tell us anything *directly* about the probability of the truth of any hypothesis as the above numerical example demonstrate. We would need to know the likelihood of the hypothesis prior to the data to compute such a probability in Bayesian fashion.

While common misinterpretations of *p*-values may be addressed through education of researchers on proper interpretation, the issue remains that researchers would like to be able to compute the probability that the hypothesis in question is true. Due to the formulation of Markov Chain Monte Carlo sampling algorithms and recent advances in computer software and hardware, Bayesian methods for such computations are now accessible (Kruschke, 2010). An important advantage of 21st century Bayesian methods over 20th century orthodox methods is that

they allow for computation of the probabilities for estimates of parameters such as the difference in means for boys and girls.

Bayesian interval estimation and Confidence Intervals

An important addition to NHST practices in the past decade or so has been the emphasis on reporting confidence intervals (CI). A CI for a difference in means can be defined as the set of mean differences that would lead to the rejection of the null hypothesis in NHST (Tukey, 1991). As such, a CI retains the interpretive problems of NHST mentioned above. In comparison, the Bayesian *highest density interval* (HDI) for a difference in means contains the most credible mean differences. Because they can be interpreted in terms of the probability of the data, HDIs are more intuitive and less likely to be misinterpreted than NHSTs and CIs are (Lecoutre, 2006; Diennes, 2011). A 95% (for example) HDI can be said to have a 95% chance of containing the true difference in means. It would be *statistically incorrect* to make the same interpretation of a frequentist CI, since under NHST, just as a hypothesis is either true or false and has a probability of 1 or 0, so is the true difference in population means considered to either be or not be in the computed CI (Huck, 2007).

In response to the fact that NHST offers no indication of the *size* of a difference when a direction is detected and to the critique that rejecting the null hypothesis can be a matter of sampling to a foregone conclusion (Tukey, 1991; Wagenmakers, 2007), researchers have recently adopted the desirable practice of reporting effect sizes. Still, unfortunately, researchers have been relying on effect size without addressing the uncertainty in effect size estimation. Below we demonstrate a Bayesian approach to estimation including interval estimation of effect sizes in the context of studying gender differences in achievement goals to illustrate the use of Bayesian estimation and to highlight its advantage in more intuitively interpreted results.

An Empirical Investigation

In the current study, a Bayesian estimation method was applied to the investigation of gender differences in achievement goals in a moderate sample of Junior High school students. The research question guiding the study was: What is the direction and size of gender differences in mastery, performance-approach, and performance-avoidance goals in a group of Junior High school students? A traditional analysis using NHST *t*-tests will be compared to a Bayesian approach. While the resulting interval estimates can be anticipated to be very similar in Bayesian and traditional analyses of this sort, we believe the interpretation of results produced by the Bayesian approach are more intuitive and straightforward.

Methods

Sample

Participants in the study were 442 7^{th}-grade students—223 girls and 219 boys—from one large 6-year secondary school in Israel. Students participated in a study on the role of motivational emphases in the school environment and of personal motivational orientations in students' aggressive attitudes and behavior. Students responded to surveys administered by research assistants in their classrooms. Teachers were present in the classroom but were not involved in the administration. Students were allotted as much time as they needed to complete the survey, which took approximately 30 minutes to complete.

Measures

Among other measures (e.g., achievement goal structures, aggressive attitudes), the measures used for the current study included scales assessing mastery goals (sample item: “One of my goals is to master a lot of new skills this year”), performance-approach goals (sample item: “One of my goals is to look smart in comparison to the other students in my class”), and

performance-avoidance goals (sample item: “One of my goals is to keep others from thinking I’m not smart in class”) from the Patterns of Adaptive Learning Survey (PALS) (Midgley et al., 2000). The PALS is one of the most prevalently used instruments for assessing achievement goals among young adolescent students. In the current study, we only analyze mean differences between the boys and girls in the sample on these three achievement goals scales.

Analysis

The analysis was conducted using Kruschke’s (2013) Bayesian Estimation Supersedes the T-test (BEST) software written in the open-source R and JAGS programming languages. It is freely available along with instructions for installation at http://www.indiana.edu/~kruschke/BEST/. Additionally, a web-based version is available that allows the user to paste in data at http://www.sumsar.net/best_online/.

A Bayesian analysis begins with a prior distribution as a model of the uncertainty about the parameters being estimated before collecting the new data. After data collection a posterior distribution based on the likelihood of the data and the prior distribution is computed. The posterior is the mathematically normative way to allocate credibility for different parameter values in light of the data. The posterior distribution cannot be calculated directly, so a computer algorithm based on Markov Chain Monte Carlo (MCMC) sampling approximates the posterior through a process of taking samples within the space of possible parameter values and reallocating belief toward combinations of parameter values that are most consistent with the data and away from the least credible ones.

In the current analysis, a total of five parameters are estimated in this model for each gender comparison of achievement goals including both population means, both population standard deviations, and a normality parameter assumed to apply to both populations. The prior

used is quite broad and vague presuming a large amount of uncertainty about the parameters to be estimated. In order to accommodate data with outliers, the data are modeled in the BEST program based on *t* distributions with degrees of freedom to be estimated from the data rather than on normal distributions. Note that the *t* distribution is not used as a sampling distribution but merely as a convenient choice as a model for the data. The parameter usually thought of as denoting degrees of freedom is used as a parameter describing the range of possibilities from approximately normal to thick tailed. With skewed data other models such as log-normal distributions could be used (Kruschke, 2012).

In light of the uncertainty in the literature concerning gender differences in achievement goals, the prior distributions used are broad and vague. If there were less uncertainty regarding the direction and size of gender differences in achievement goals such as expectation that the difference is likely to be in a particular direction or general skepticism towards very large differences, the degree of prior knowledge could be taken into account in specifying a more informative prior distribution**.** In the current example, the prior is uninformative, which is to say that it will have little influence on the posterior distribution. If we had important prior knowledge, we could take advantage of it, but in the case of great uncertainty, an uninformative prior distribution allowing the data to easily overwhelm it is appropriate. The specification of the uninformative prior used in the BEST software is discussed in Appendix B.

For comparison of results, traditional independent sample *t* tests were run using SPSS.

Findings

Traditional analysis

Independent samples *t* tests (equal variances not assumed) indicate statistical significance of the claim that girls are more mastery-oriented ($t(414.4) = -3.395$, $p < .01$) than boys are. Boys

scored statistically significantly higher on measures of both performance-approach ($t(435.0) = 7.192$, $p < .01$) and performance-avoidance ($t(435.9) = 5.124$, $p < .01$) goals.

Table 1 displays the differences in sample means (Boys – Girls) and 95% confidence interval (CI) estimates for the differences in achievement goals.

Bayesian analysis

Table 2 displays estimates for the differences (Boys – Girls) in population means and standard deviations and for effect sizes (computed as the standardized mean difference $d = (\mu_1 - \mu_2)/\sqrt{(\sigma_1^2 + \sigma_2^2)/2}$) for gender on achievement goal measures with corresponding 95% HDIs. Appendix A contains sample graphical output from the BEST software. According to the Bayesian analysis, if we assume broad uncertainty in the direction and size of gender differences prior to seeing the data, the data warrant nearly 100% certainty in the claim that Girls in this context are more mastery-oriented than boys are. The 95% HDI on effect size, which marks the location of the most believable effect sizes in light of the data, ranges from 0.148 to 0.566 indicating that the difference is most likely small to moderate. Boys scored higher on performance-avoidance as well as performance-approach goals with nearly 100% of credible differences being greater than zero. The greatest gender effect was the extent to which boys versus girls adopted performance-approach goals with a 95% HDI on effect size of 0.503 to 0.900.

While both the traditional and Bayesian approaches produce interval estimates for means, only the BEST software also estimates differences in standard deviations and provides HDIs reflecting the uncertainty in effect size estimation for assessing practical significance.

Discussion

Because the Bayesian analysis in the current study assumed great uncertainty in prior knowledge, the endpoints of the HDI estimates for the difference in population means are nearly identical to those of the corresponding traditional CIs. If the results of the traditional *t* test and CIs and those of the Bayesian estimation were very dissimilar, we should be suspicious of the CIs because it is the Bayesian analysis rather than the traditional one that warrants probabilistic interpretation of the intervals as being likely to contain the true differences in population means. In general, although the CI estimates of differences in means for mastery, performance-approach, and performance-avoid goals are consistent with those of the Bayesian analysis, as will be the case when presuming utter ignorance prior to seeing the data, the naïve Bayesian interpretation that is routinely misapplied to NHSTs and CIs is still fallacious (Falk & Greenbaum, 1995).

In the current study as in many other practical cases, Bayesian prior distributions are vague and uninformed or only mildly informed and therefore have little influence on the posterior distribution (Kruschke, 2010). The current case is based on the uncertainty in the literature regarding gender differences in achievement goals. Yet, the findings of the current study can join findings of other studies about the direction and size of gender difference in achievement goals (even if not statistically significant according to NHST) in establishing an increasingly credible prior knowledge that can be integrated into future estimation of such differences. Such incorporation of accumulating knowledge is not part of the NHST practice. Moreover, the ability to build on prior knowledge about variables of interest such as boys and girls adoption of different achievement goals in Bayesian analyses may make it particularly advantageous for mixed-methods research, since qualitative findings can be used to inform Bayesian priors (Gorard et al., 2004). The incorporation of prior knowledge that is both

cumulative and consensual (as in peer review) reflects scientific progress as the prior knowledge becomes better founded with each study (Kruschke, 2012). Further research on estimating the sizes of gender differences in achievement goals for the population investigated in the current study (or populations believed to be similar) could theoretically be informed by the results of the current study to achieve better precision than what could be achieved with traditional *t* intervals or Bayesian estimation with an uninformative prior. It may therefore be regarded as a strength rather than a weakness that Bayesian methods can make proper use of prior knowledge.

When we do have strong beliefs prior to collecting the data, Bayesian methods can account for the skepticism we have toward wild deviations from our pre-existing beliefs while letting the data have its say. Rather than being invisibly manipulated to produce a forgone conclusion, prior distributions are explicitly reported for scientific publications and must be chosen to be acceptable to the audience of the reported finding. Further, alternative analyses can be run based on alternative priors to demonstrate robustness of findings to other plausible prior beliefs that may be held by different audience members.

The current study identified gender differences with boys more likely to endorse performance approach and avoidance goals than are girls, and girls more likely to endorse mastery goals than are boys—findings that are similar to those of Anderman and Young (1994). In the current study, boys' higher endorsement of performance goals relative to girls had a moderate to strong effect (d = .503 to .900) for performance-approach goals and a moderate effect (d = .303 to .691) for performance-avoid goals. Girls' higher endorsement of mastery goals relative to boys had a small to moderate effect (d = -.566 to -.148). Since performance-approach and avoid goals, primarily those focusing on demonstrating ability, have been associated with maladaptive patterns of learning such as disruptive behavior, unwillingness to

cooperate, low efficacy, negative emotions, self-handicapping strategies, and low performance (Kaplan & Maehr, 2007), these differences in achievement goal endorsement could explain gender differences in a school adjustment by boys and girls. Increasing evidence of girls doing better in school than do boys is compatible with the current motivational findings. The interpretation should take account that the HDI estimates for the differences in standard deviations for the measures of degrees of endorsement of both performance-approach and performance-avoid included zero as a credible difference. As discussed, such a case should not lead us to conclude that there is no difference in the variability of levels of endorsement for both types of performance goals. On the contrary, this result means that we remain uncertain as to whether boys are girls exhibit more variability. Regarding mastery goals, there is credibly more variability in the degree of endorsement for boys than for girls (HDI for difference in standard deviations .048 to .203). Past research has not addressed differences in variability, and though the BEST software produces estimates of these differences, it is not clear how they could contribute to application or development of achievement goal theory.

The findings of the current study are consistent with those of Anderman and Young (1994). However, it is not obvious whether or not these results about the direction and size of gender differences in achievement goals contradict or can be considered consistent with those of Meece and Jones (1996) who reported no main effect or interval estimates for the sizes of gender differences on achievement goals. The important point is that the absence of evidence is not the same as evidence of absence. Detecting no statistically significant difference in a study of gender effects for achievement goals using NHST is not a contribution to knowledge. It is admitting the failure to generate new knowledge. By contrast, Bayesian estimation methods can reveal an effect to be so small as to be practically insubstantial, e.g. a 95% HDI for effect size ranging

from .01 to .05. In such a case, a Bayesian analysis, but not an NHST test, could advance achievement goal theory by supporting a finding of *no meaningful difference* between genders. We could only know if the finding of no main effect for gender in Meece and Jones (1996) does indeed conflict with Anderman and Young (1994) and with the current findings if the researchers had estimated the *size* of the difference rather than merely applying NHST to test for the *existence* of a difference. A state of conflict in results would obtain only if there were no overlap in the interval estimates.

Just as some journal guidelines require authors to distinguish statistical significance from practical significance—to distinguish between a result that matters from a result that is of a size that would be unlikely to occur by chance alone under certain assumptions—guidelines might also require language in reporting of findings that distinguish between finding the absence of a meaningful difference (something that could be accomplished with Bayesian estimation) and failing to determine the direction of the difference (i.e., $p > .05$).

Conclusion

While conclusions regarding group differences in boys' and girls' motivational processes have theoretical and practical implications, the current state of the literature concerning gender differences in achievement goals includes inconsistent findings and uncertain conclusions. To some degree, this state of knowledge stems from researchers' reliance on NHST and its structural and interpretive problems. Researchers using statistics routinely misinterpret *p*-values in tests of significance as indicating the probability that the null hypothesis is true and with 1-*p* indicating the probability that the alternative hypothesis is true. Similarly, 95% CIs are routinely misinterpreted as having a 95% chance of containing the parameter being estimated. These misinterpretations are made in part because statements about the probability that a hypothesis is

true or that a parameter is in a CI—probability statements that are forbidden by the frequentist paradigm underlying CIs and NHST—are answers to questions that researchers want answered. It is the Bayesian paradigm that can directly address these questions.

Though Bayesian statistics has always had many logical and philosophical advantages over the dominant frequentist methods, in large part because of the opposition of Ronald Fisher who advocated for his approach of significance testing over use of Bayes' Rule, frequentism was widely thought to have the high-ground of objectivity since the early nineteen hundreds (McGrayne, 2011). However, researchers have become more and more troubled with problems with traditional null hypothesis significance testing and are looking for alternatives (Wagenmakers, 2007). At the same time, modern advances in computing capabilities have made the Bayesian approach far more practical than it had been in the days of Ronald Fisher (Gelman, 2004; Wagenmakers, 2007; Krushke, 2011). The Bayesian approach has therefore reemerged in recent decades and seems to be poised to become the dominant one in the 21st century (e.g., Gelman, 2004; Lee and Wagenmakers 2005; Wagenmakers, 2007; Krushke, 2011). Wagenmakers (2007) reports that the proportion of Bayesian articles in the Journal of the American Statistical Association rose to over 25% in the 2000s (Wagenmakers, 2007). There is no reason to believe that the field of educational research will be left behind in the Bayesian revolution.

Research on gender differences in motivational processes—and statistical research in motivation more generally—might benefit from adopting Bayesian statistics as its analytic paradigm. Regardless of whether education researchers choose approaches similar to that used in this investigation of gender differences in achievement goals, consideration of a comparison of the Bayesian framework and traditional methodology may shed light on the meaning of NHST

and its limitations and may improve interpretations of findings based on NHST. We follow Lecoutre (2006) in recommending consideration of Bayesian methods "as a *therapy* against the misuses and abuses of NHST" (p. 208).

References

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*(3), 261.

Anderman, E. M., & Young, A. J. (1994). Motivation and strategy use in science: Individual differences and classroom effects. *Journal of Research in Science Teaching, 31*(8), 811-831.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science, 18*(1), 1-32.

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician, 59*(2), 121-126.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*(3), 274-290.

Eccles, J. (1983). Female achievement patterns: Attributions, expectancies, values, and choice. *Journal of Social Issues, 1*, 1-22.

Elliot, A. J. (2005). A conceptual history of the achievement goal construct. *Handbook of Competence and Motivation, 16*, 52-72.

Falk, R., & Greenbaum, C. W. (1995). Significance Tests Die Hard The Amazing Persistence of a Probabilistic Misconception. *Theory & Psychology*,*5*(1), 75-98.

Fisher, R. A. (1935). The Design of Experiments.

Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*,*15*(3), 373-390.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*,*5*(2), 189-211.

Gorard, S., Roberts, K., & Taylor, C. (2004). What kind of creature is a design experiment? *British Educational Research Journal, 30*(4), 577-590.

Greene, B. A., DeBacker, T. K., Ravindran, B., & Krows, A. J. (1999). Goals, values, and beliefs as predictors of achievement and effort in high school mathematics classes. *Sex Roles, 40*(5-6), 421-458.

Huck, S. W. (2007). Reading statistics and research. *Reading Statistics and Research, 5*, 226-284.

Hulleman, C. S., Schrager, S. M., Bodmann, S. M., & Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin*, *136*(3), 422-449.

Kaplan, A., & Maehr, M. L. (2007). The contributions and prospects of goal orientation theory. *Educational Psychology Review, 19*(2), 141-184.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS.* Academic Press / Elsevier.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14*(7), 293-300.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General, 142*(2), 573.

Lecoutre, B. (2006). Training students and researchers in Bayesian methods. *Journal of Data Science*, *4*, 207-232.

Maehr, M. L., & Zusho, A. (2009). Achievement goal theory: The past, present, and future. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 77-104). New York: Routledge.

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools, 5*(2), 15-22.

Meece, J. L., Glienke, B. B., & Burg, S. (2006). Gender and motivation. *Journal of School Psychology, 44*(5), 351-373.

Meece, J. L., & Jones, M. G. (1996). Gender differences in motivation and strategy use in science: Are girls rote learners? *Journal of Research in Science Teaching, 33*(4), 393-406.

Middleton, M. J., & Midgley, C. (1997). Avoiding the demonstration of lack of ability: An underexplored aspect of goal theory. *Journal of Educational Psychology, 89*(4), 710.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., . . . Middleton, M. J. (2000). Manual for the patterns of adaptive learning scales. *Ann Arbor, 1001*, 48109-41259.

Neyman, J. (1937) Outline of a theory of statistical estimation based on the classical theory of probability." *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 236.767 (1937): 333-80.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London.Series A, Containing Papers of a Mathematical or Physical Character, 231*, 289-337.

Nicholls, J. G. (1989). *The competitive ethos and democratic education.* Harvard University Press.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*(5), 416.

Senko, C., Hulleman, C. S., & Harackiewicz, J. M. (2011). Achievement goal theory at the crossroads: Old controversies, current challenges, and new directions. *Educational Psychologist*, *46*(1), 26-47.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*(2), 309.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*(1), 100-116.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779-804.

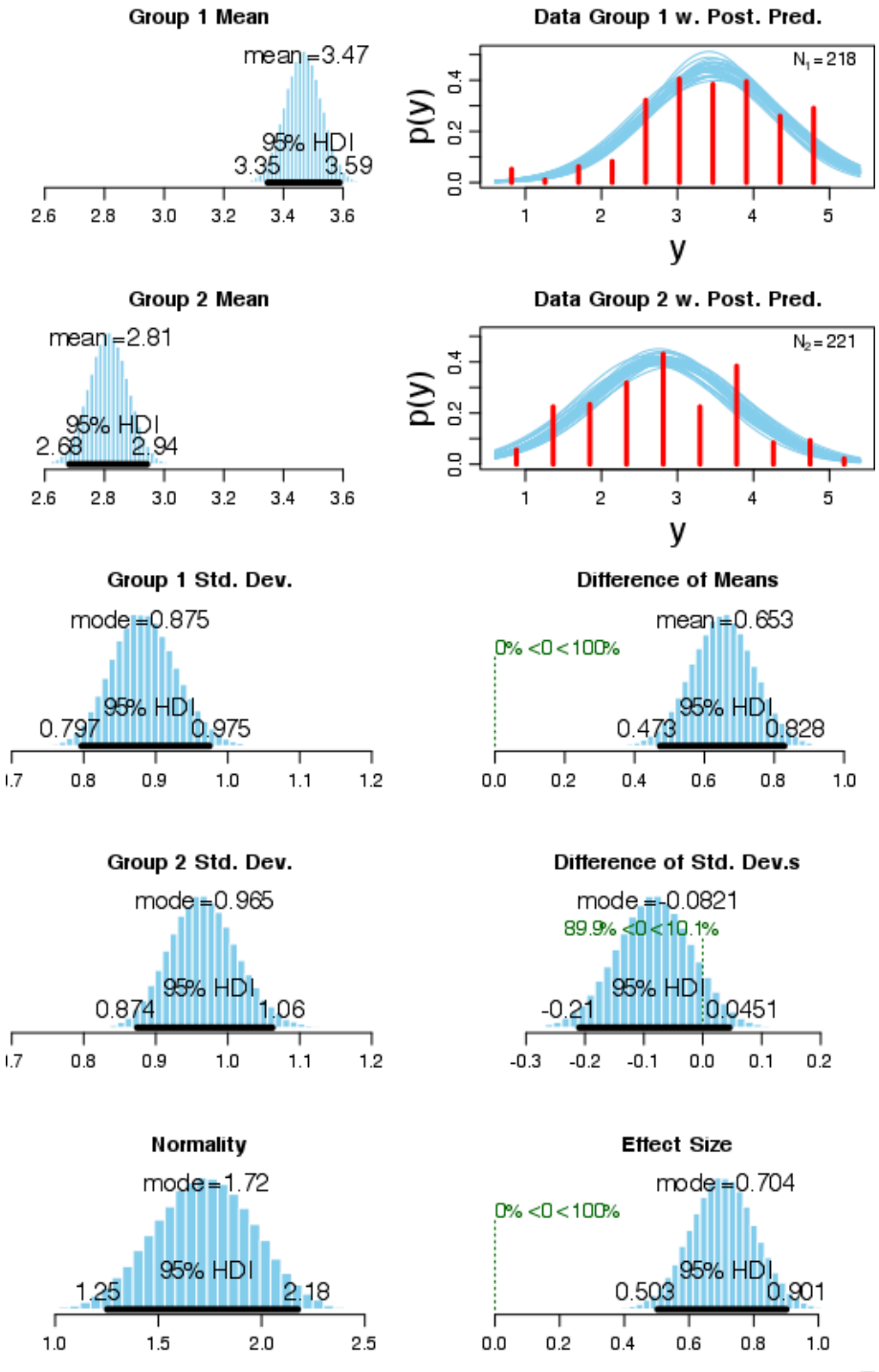
Wainer, H., & Robinson, D. H. (2003). Shaping up the practice of null hypothesis significance testing. *Educational Researcher, 32*(7), 22-30.

Wigfield, A., & Eccles, J. S. (2002). Students' motivation during the middle school years. In J. Aronson, (Ed), Improving academic achievement: Impact of psychological factors on education (pp. 159-184). San Diego, CA: Academic Press.

Appendix A

BEST graphical output comparing boys (Group 1) with girls (Group 2)

with respect to performance-approach goals.



Appendix B

The model for the data is defined in the program as

```
model {
  for ( i in 1:Ntotal ) {
  y[i] ~ dt( mu[x[i]] , tau[x[i]] , nu )
 }
 for ( j in 1:2 ) {
  mu[j] ~ dnorm( muM , muP )
  tau[j] <- 1/pow( sigma[j] , 2 )
  sigma[j] ~ dunif( sigmaLow , sigmaHigh )
 }
 nu <- nuMinusOne+1
 nuMinusOne ~ dexp(1/29)
}
```

The first four lines specify the model for the data as having *t* distributions with parameters mu, tau, and nu. These parameters refer to the mean, the precision (the reciprocal of the variance), and the degrees of freedom) that are to be estimated from the data based on their prior distributions. The priors are specified in the remaining lines of code above with certain constants specified later in the program with the lines

```
muM = mean(y) ,
muP = 0.000001 * 1/sd(y)^2 ,
sigmaLow = sd(y) / 1000 ,
sigmaHigh = sd(y) * 1000
```

Broad uncertainty prior to seeing the data is assumed in the original program, therefore the prior distributions for the possible values of the means of the *t* distributions that model the data are assigned means equal to the pooled mean of the data with a standard deviation that may be as small as 1/1000 the pooled standard deviation of the data to 1000 times the pooled standard deviation of the data. Specifying an informative prior requires programming knowledge of R and JAGS for modifying the model specification by changing some of the code.

Table 1: False Alarm Rates Based on 1000 Hypothetical Low-Powered NHSTs

	Difference in population means is non-zero	No difference in population means	Total
H_0 is rejected	40	45	85*
H_0 is not rejected	60	855	915
Total	100	900	1000

*Out of the 85 hypothetical NHSTs where H_0 is rejected, there is a difference in the population means in fewer than half of them.

Table 2: Traditional Estimates

	Mean Difference for Boys - Girls (Lower Bound, Upper Bound)
Mastery goals	-.186 (-.294, -.078)
Performance – Approach	.643 (.468, .819)
Performance – Avoid	.464 (.286, .643)

Table 3: Bayesian Estimates

	Difference in Means for Boys - Girls	Difference in Std. Dev. for Boys - Girls	Effect Size
Mastery goals	-.180 (-.284, -.075)	.124 (.048, .203)	-.359 (-.566, -.148)
Performance - Approach	.653 (.473, .828)	-.082 (-.210, .045)	.705 (.503, .900)
Performance – Avoid	.467 (.289, .650)	-.079 (-.207, .051)	.497 (.303, .691)

View publication stats