Learning and Individual Differences 47 (2016) 129–135

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif

Bayesian analysis in educational psychology research: An example of gender differences in achievement goals

Steven K. Peterson *, Avi Kaplan

Temple University, 1801 N Broad St, Philadelphia, PA 19122, USA

ARTICLE INFO

Article history:
Received 29 October 2014
Received in revised form 9 November 2015
Accepted 10 January 2016

Keywords:
Achievement goals
Motivation
Bayesian statistics
Junior high
Gender

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 7th-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.

1. Introduction

1.1. Uncertainty regarding gender differences

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, Schrager, Bodmann, & Harackiewicz, 2010; Linnenbrink-Garcia, Tyson, & Patall, 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.

1.2. Achievement goal theory

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

* Corresponding author at 1717 Earlington Rd, Havertown, PA 19083, USA.
E-mail address: peterson.steve@gmail.com (S.K. Peterson).

[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, Kaplan, & Weinstock, 2011), and it was not included in the current study. For brevity, mastery-approach goals in the current manuscript are labeled simply mastery goals.

http://dx.doi.org/10.1016/j.lindif.2016.01.005
1041-6080/

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 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.Cohen, 1994, Dienes, 2011, McLean & Ernest, 1998, Rozeboom, 1960) with a lamentable influence on normative practice (Falk & Greenbaum, 1995; Lecoutre, 2006; Sedlmeier & Gigerenzer, 1989). 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.

2. Theoretical underpinnings

2.1 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).

2.2. 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; Gelman, Hill, & Yajima, 2012; Gelman & Tuerlinckx, 2000; Wagenmakers, 2007). 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 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.

2.3. Critiques of NHST

NHST has been criticized repeatedly on several grounds (see Christensen, 2005; Cohen, 1994; Falk & 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

[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. Kruschke (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.