CHALLENGES ASSOCIATED WITH MEASURING ATTITUDES USING THE SATS FAMILY OF INSTRUMENTS

DOUGLAS WHITAKER
Mount Saint Vincent University
douglas.whitaker@msvu.ca

ALANA UNFRIED
California State University, Monterey Bay
aunfried@csumb.edu

MARJORIE E. BOND
Monmouth College
mebond@monmouthcollege.edu

ABSTRACT

The Survey of Attitudes Toward Statistics (SATS) is a widely used family of instruments for measuring attitude constructs in statistics education. Since the development of the SATS instruments, there has been an evolution in the understanding of validity in the field of educational measurement emphasizing validation as an on-going process. While a 2012 review of statistics education attitude instruments noted that the SATS family had the most validity evidence, two types of challenges to the use of these instruments have emerged: challenges to the interpretations of scale scores, and challenges using the SATS instruments in populations other than undergraduate students enrolled in introductory statistics courses. A synthesis of the literature and empirical results are used to document these challenges.

Keywords: *Statistics education research; Attitudes toward statistics; Assessing attitudes; Validity evidence*

1. INTRODUCTION

The Survey of Attitudes Toward Statistics (SATS) instruments (Schau, 1992, 2003b) are among the most widely used attitude instruments in statistics education and have been translated into over a dozen languages.[1] As part of the 2012 SERJ special issue on attitudes toward statistics, a systematic review of validity and reliability evidence for instruments measuring constructs related to attitudes in statistics education reported that the SATS family of instruments had the most documented evidence available (Nolan et al., 2012). For nearly three decades, the SATS instruments have been invaluable tools available to researchers and teachers for measuring student attitudes in statistics.

In the decades since the initial release of the SATS instruments, the fields of statistics education and validity/measurement have made considerable progress in deepening our understanding of attitudes toward statistics and how to measure them, and best practices in these areas have been refined and updated. These changes, together with the growing body of literature using the SATS instruments, have implications for researchers and users of the SATS instruments today. This manuscript documents two broad types of challenges to using the SATS instruments. The first challenge is in interpreting scores when the instruments

[1] Arabic (Nasser, 2004), Bahasa Malaysia (Rosli & Maat, 2017), Chinese (Luh et al., 2004; Zhang et al., 2012), Dutch (Vanhoof et al., 2011), Estonian (Hommik & Luik, 2017), French (Carillo et al., 2016), German (Strobl et al., 2010), Greek (Bechrakis et al., 2011), Hebrew (Lipka & Hess, 2016), Italian (Chiesi & Primi, 2009), Portuguese (Vendramini et al., 2011), Russian (Khavenson et al., 2012), Spanish (Estrada, 2002), Thai (Ratanaolarn, 2016), and Turkish (Emmioglu Sarikaya et al., 2018)

Statistics Education Research Journal, 21(1), Article 4. https://doi.org/10.52041/serj.v21i1.88

are used with the intended population of undergraduate introductory statistics students, and the second challenge concerns the growing trend of using the SATS instruments with other populations such as graduate students and teachers.

Some of the challenges to interpretations with undergraduate students have been mentioned previously in the literature but have not been catalogued in one document supporting users and researchers. Instead, they have appeared in many outlets including journal articles focusing on statistics education (e.g., Schau & Emmioğlu, 2012) and other academic areas such as educational psychology (e.g., Cashin & Elmore, 2005), conference presentations and proceedings (e.g., Millar & Schau, 2010; Sorge & Schau, 2002), and as letters to the editor in *SERJ* (e.g., Schau & Millar, 2011). Recommendations for researchers using the SATS have previously been proposed (e.g., Millar & Schau, 2010; Millar et al., 2013; Schau, 2008; Schau & Millar, 2011), but these are not comprehensive and often focus on the statistical analysis of collected data rather than score interpretations. It is likely that researchers and instructors using or considering using the SATS instruments—either by administering it themselves or drawing on research in which others have administered it—are not fully aware of the challenges and recommendations extant in the literature.

This paper aims to document challenges and recommendations for using the SATS instruments. In Section 2, existing validity evidence and challenges previously identified in the literature are summarized. Then, two types of challenges to the use of the SATS instruments are discussed in detail: challenges to the interpretations of scale scores (Section 3) and challenges using the SATS instruments in populations other than undergraduate students enrolled in introductory statistics courses (Section 4). Finally, conclusions are presented in Section 5.

2. EXISTING VALIDITY EVIDENCE

The SATS family of instruments (pre- and post-course versions of the SATS-28 and SATS-36) is used widely to measure attitudes of (primarily undergraduate) students enrolled in first courses in statistics, though researchers are increasingly administering the SATS to populations beyond this traditional group. The first Survey of Attitudes Toward Statistics was a 28-item instrument released in 1992 (SATS-28; Schau, 1992) measuring four constructs:

- Affect (6 items) – students' feelings concerning statistics
- Cognitive Competence (6 items) – students' attitudes about their intellectual knowledge and skills when applied to statistics
- Value (9 items) – students' attitudes about the usefulness, relevance, and worth of statistics in personal and professional life
- Difficulty (7 items) – students' attitudes about the difficulty of statistics as a subject (Schau, 2005a, pp. 1–2)

The SATS-28 was expanded in 2003 by including eight additional items (SATS-36; Schau, 2003b). The SATS-36 measures the four constructs measured by the SATS-28 along with two more constructs:

- Interest (4 items) – students' level of individual interest in statistics
- Effort (4 items) – amount of work the student expends to learn statistics (Schau, 2005b, pp. 2–3)

Pre- and post-course versions of the SATS-28 and SATS-36 instruments are available.[2]

Because the SATS-28 is a subset of the SATS-36 instrument (excluding the effort and interest constructs), research about the SATS-28 continues to be directly relevant to the appropriate subset of the SATS-36. In this article, SATS refers to the SATS-28 and SATS-36 collectively.

[2] Please see the SATS scoring documents and reports for information about specific operational details such as item wording and scoring (Schau, 2005a, 2005b, 2008).

2.1. ORIGINAL SATS VALIDITY EVIDENCE

Results of the initial validation study for the SATS-28 instrument (Schau et al., 1995) included a description of the item and scale development using a modified Nominal Group Technique (NGT; Moore, 1987), a pilot administration that led to deleting items, and an operational administration that led to further refinements and data to be analyzed. The data analysis included calculations of coefficient alpha, a confirmatory factory analysis (CFA) that demonstrated an acceptable four-factor solution, and correlations between the SATS scale scores with scale scores from the Attitudes Toward Statistics instrument (ATS; Wise, 1985). For prior attitude instruments such as the Statistics Attitude Survey (SAS; Roberts & Bilderback, 1980) and ATS instruments (Wise, 1985), validity evidence largely consisted of exploratory factor analysis (EFA) and correlations of scale scores with outcomes such as course grades: little information about the development of the instruments was presented. Drawing on the formal item and scale development procedure and statistical analyses, the validity evidence supporting the use of the SATS-28 presented by Schau et al. (1995) was stronger than the validity evidence for many other available instruments for measuring related constructs (Carmona Márquez, 2004; Nolan et al., 2012). No initial validity study that focused on the development process of the SATS-36 was published, though Ramirez et al. (2012) briefly described the joint SATS development process, and since its release many researchers have contributed to the validity evidence for the SATS-36 in separate studies.

In their 2012 article, Nolan et al. summarized the available validity evidence supporting the use of the SATS-28, SATS-36, and other attitude instruments. Nolan et al. noted the strength of NGT to develop the items relative to the process used by other instruments and an articulated educational theory for the constructs. Nolan et al. also noted numerous studies about the factor structure of the SATS-28 and SATS-36 are consistent with four and six factor solutions, respectively, though the studies were not unanimous in this finding (e.g., Vanhoof et al., 2011) and were the subject of debate within the community at the time (cf. Schau & Millar, 2011). Internal consistency, as measured by coefficient alpha (Cronbach, 1951), was generally high for each of the SATS scales (Nolan et al., 2012). Furthermore, correlations between SATS scales and other mathematics and statistics measures have generally been consistent with hypothesized relationships (Nolan et al., 2012). Lastly, the SATS instruments have been shown to account for some variation in achievement (Nolan et al., 2012). Outside of the evolving discussion about the factor structure of the SATS instruments, we are unaware of any studies published since the 2012 *SERJ* special issue that are inconsistent with the validity evidence claims articulated by Nolan et al. and summarized above.

2.2. USES OF THE SATS INSTRUMENTS

Since their debut, the SATS instruments have been the dominant choice among attitude surveys for statistics educators, both for understanding their own students' attitudes and for conducting research. In the SATS research literature,[3] articles typically belong to one of three categories: articles about the structure or administration of the SATS, articles about translating the SATS into other languages, or articles wherein an SATS instrument is used to measure change or group differences such as with an intervention or demographic comparison. We will first briefly review some research uses of the SATS instruments, and then we will re-examine the validity evidence supporting the use of the SATS instruments for the remainder of the paper.

A key use of SATS instruments has been measuring differences between educational approaches, particularly the effect of active learning strategies on student attitudes. For these uses, results have been mixed: some studies have reported statistically significant changes or differences in at least some of the SATS components while others have reported no changes. In an analysis of approximately 2200 students from 101 sections of statistics courses, Schau and Emmioğlu (2012) demonstrated that students tend to

[3] A list of published articles that use an SATS instrument is available on Dr. Schau's website (https://www.evaluationandstatistics.com/references). Given the popularity of the instrument, however, the list is incomplete.

begin a semester with neutral attitudes on four of the six components (Affect, Cognitive Competence, Difficulty, and Interest), value statistics somewhat positively, and intend to put forth a lot of effort in the course. Schau and Emmioğlu (2012) also found, however, that the average attitudes for sections of statistics courses tended to either remain the same (Affect, Cognitive Competence, and Difficulty) or decrease (Value, Interest, and Effort).

Schau and Emmioğlu's work has served as an important point of comparison by providing pre- and post-course average attitude scores for sections. Other studies have also demonstrated essentially no change or negative change among the attitude components as measured by the SATS-36 (e.g., Bateiha et al., 2020; Carnell, 2008), but many uses of the SATS instruments occur in studies with small sample sizes and interventions that are limited in scope—which may explain why differences are sometimes not observed. Other researchers have used the SATS-36 to show improvement of attitudes in active learning environments, with Carlson and Winquist (2011) and Posner (2011) reporting statistically significant mean changes in four and six of the attitude components, respectively.

The utility of the SATS instruments and the contributions to statistics education they enable are most apparent in studies that utilize statistical methods beyond inferential comparisons of two groups or incorporate other complementary data sources. For example, Gundlach et al. (2015) used ANOVA to compare pre- and post-course (time) attitude scores for students in traditional, online, and flipped classes (section). They found that, across all section types, students tended to have decreases in scores from pre- to post-course on the Value, Interest, and Effort scales (Gundlach et al., 2015). Gundlach et al. also identified statistically significant interaction effects between time and section for the remaining scales: there were statistically significant changes for Affect (positive change for traditional and flipped), Cognitive Competence (positive change for traditional), and Difficulty (easier for traditional and online). Using Structural Equation Modeling (SEM) with the pre- and post-course scales as latent variables to construct pre- and post- attitude variables in their SEM analyses, Chiesi and Primi (2010) found that attitudes at the beginning of the course have an effect on attitudes at the end of the course. Combining results on a concept inventory and the SATS-36, Chance et al. (2016) used hierarchical models to show that pre-Cognitive Competence and pre-Difficulty scores were better predictors of gains on the concept inventory than other attitude components and identified an interaction between the experience of instructor and student Difficulty scores. Chance et al. also used SATS-36 pre-course scores as part of a cluster analysis to identify student groups. Whether researchers use advanced analysis techniques or make straightforward pre/post comparisons, there are numerous studies that have expanded the statistics education literature about attitudes and their relationship to interventions and other constructs in which the SATS family of instruments has served a foundational role.

2.3. CURRENT UNDERSTANDING OF VALIDITY EVIDENCE

It is important to recognize that neither the body of validity evidence supporting an instrument nor the discipline's understanding of validity evidence are static. Since the release of the SATS instruments and publication of seminal studies in the SATS literature (e.g., Nolan et al., 2012; Schau & Emmioğlu, 2012; Schau et al., 1995), the understanding of validity in educational measurement has evolved. A cornerstone of the contemporary understanding of validation is the 2014 *Standards for Educational and Psychological Testing*, a joint publication of the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME), the latest release in a series dating back decades (2014). The conceptualization of validity in the *Standards* is an evolution of prior conceptualizations such as that of Messick (1989, 1995) and the tripartite view of criterion, content, and construct validity (Cronbach & Meehl, 1955) and focuses on interpretations and the unitary nature of validity. The *Standards* make clear the connections among interpretations and validity:

> It is the interpretations of test scores for proposed uses that are evaluated, not the test itself …. Statements about validity should refer to particular interpretations for specified uses. It is incorrect to use the unqualified phrase "the validity of the test." (AERA et al., 2014, p. 11)

For an instrument or test, how scores are interpreted and used is the central focus of validation studies: without sufficient evidence supporting the proposed interpretation or use of a score from a scale, instrument, or test, the interpretation or use should be made cautiously or not at all. Validity is also better understood as referring to the appropriateness of specific uses of an instrument with specific groups:

> Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use. Like the 1999 *Standards*, this edition refers to types of validity evidence, rather than distinct types of validity. (AERA et al., 2014, p. 14)

The process of establishing validity evidence is not finished once an initial validation study has been completed, and the *Standards* makes explicit the "joint responsibility of the test developer and the test user" (AERA et al., 2014, p. 13) for documenting appropriate validity evidence for intended uses. This contemporary view of validity evidence emphasizing shared responsibility means that end users of the instrument should also document confirming and disconfirming validity evidence for an instrument's use in specific contexts. It is in this context of a more integrated view of validity evidence that we articulate further validity evidence for the SATS family of instruments beyond that previously considered (e.g., Carmona Márquez, 2004; Nolan et al., 2012). A more robust understanding of the validity evidence related to the SATS-36 instruments will clarify the interpretations and uses that are and are not supported.

2.4. CHALLENGES WITH UNDERGRADUATE INTRODUCTORY STUDENTS

While the SATS family of instruments has the most validity supporting their use (Nolan et al., 2012) among existing instruments used to measure attitudes about statistics with students enrolled in introductory statistics courses, evidence of five challenges to the use of the SATS instruments with students in introductory statistics courses has been accumulating. These challenges concern the SATS instruments' theoretical framework, empirical factor structure, development for use in traditional courses, rigid pre/post structure, and interpretability of scores from the Difficulty scale. Each of these challenges complicates or threatens the interpretability of SATS scores and, thus, the use of these instruments with the intended population of introductory statistics students.

Alignment to theoretical framework. Alignment to a theoretical framework or conceptual domain is a critical part of the instrument development process (Bandalos, 2018; Wilson, 2005). The SATS instruments are among the few attitude measures in statistics education aligned explicitly with an educational theory (Nolan et al., 2012). The educational theory most germane for the SATS family is Eccles and colleagues' Expectancy-value Theory (EVT; Eccles, 1983, 2014; Eccles & Wigfield, 2002; Wigfield & Eccles, 2000). In the EVT theory of learning, one's achievement-related choices and outcomes are based on what one values and what one expects to happen: all other constructs are mediated through these expectancies and values. The conceptual model underlying the SATS instruments, the SATS-M, is based primarily on Eccles's EVT and has been described as an application of EVT to statistics education (Schau, 2003a, 2003c; Sorge & Schau, 2002). The most recent SATS-M also draws from self-determination theory, self-efficacy theory, and achievement goal theory (Ramirez et al., 2012). While there are compatibilities between EVT and the attitudinal components measured by the SATS instruments (Schau, 2003c), EVT did not guide the development of the SATS scales (Xu & Schau, 2019). Instead, the initial four-construct structure of the SATS-28 was a result of the NGT process used in its development (Schau et al., 1995). This has resulted in an incomplete, imperfect alignment between the constructs measured by the SATS instruments and the widely used EVT framework.

The alignment of the SATS instruments to the underlying theoretical framework is related to several challenges when using the instruments. As previously noted, the theoretical framework was not an a priori guide for the development of the SATS-28 instrument: the four-factor structure was proposed by the participants in the NGT process (Schau et al., 1995). A posteriori alignment of SATS instruments with EVT through the SATS-M framework results in important oversimplifications of Eccles's EVT (Ramirez et al., 2012). For example, the EVT model conceptualizes Value as having four components: Interest, Attainment,

and Utility Value, and Cost (Eccles & Wigfield, 2002). The SATS Value construct focuses on Utility Value, and Attainment Value is largely unmeasured (Whitaker & Gorney, 2017). The SATS-36 Effort construct only partially measures the EVT Cost construct (Whitaker et al., 2018). Reflecting these differences, Tempelaar, Gijselaers et al. (2007) refer to the theoretical framework for the SATS instruments as "Schau's … expectancy-value model" (p. 107). Although the SATS-M has been used to describe relationships among statistics attitudes and statistics achievement (Ramirez et al., 2012), the alignment challenges preclude the use of the SATS instruments for measuring attitudes consistent with more widely known and researched educational theories such as Eccles's EVT. These alignment challenges complicate the interpretations of the SATS scale scores: it is not clear to what extent each of the SATS scales measures the constructs they are nominally aligned to from either the SATS-M or EVT frameworks.

Factor structure of the SATS. The factor structure of the SATS-28 was well-documented after its initial creation (e.g., Dauphinee et al., 1997; Hilton et al., 2004; Schau et al., 1995). Determining the mapping of items to scales is an important part of instrument development, and the empirical factor structure of an instrument plays a critical role in this determination. The final item-scale mapping is used for creating scores for each construct and ensures comparability across studies that use the same instrument. CFA results showed a four-factor structure as expected (Schau et al., 1995), but item parceling was used, which is no longer strongly recommended (VanHoof et al., 2011). Using an EFA, Cashin and Elmore (2005) found only two factors on the post-course SATS-28: Difficulty, all but one Cognitive Competence item, and most Affect items formed one factor while the Value items and the three remaining items formed the other.

In more recent literature using the SATS-36, studies have shown that both four-factor (Cashin & Elmore, 2005; Vanhoof et al., 2011) and six-factor (Carillo et al., 2016; Persson et al., 2019; Tempelaar et al., 2007; Vanhoof et al., 2011; Xu & Schau, 2019) structures are supported. Studies showing support for a six-factor structure, however, tend to find better model fit for data from the post-course survey than from the pre-course survey, and researchers tend to identify several items that do not fit well (sometimes proposing them for deletion). Schau and colleagues propose that a six-component structure be preferred even if other factor structures give a similar fit statistically (Schau & Millar, 2011) because it is congruent with EVT (Ramirez et al., 2012). These studies about the factor structure of the SATS instruments might be used to justify selecting subsets of SATS items to administer or changing how the responses are summarized into scale scores (i.e., grouping the items differently to form new scales), but any deviations from the standards proposed by Schau and colleagues must be justified. Such changes, however, are generally inadvisable: deviations from the six-factor structure using all 36 items and the standard scoring algorithms make interpreting SATS scale scores and comparing them across studies challenging.

New course formats. Because statistics education reform was in its infancy, the students for whom the SATS instruments were developed were enrolled in what would now be considered traditional, lecture-based, face-to-face statistics courses (Schau, 2003c). Subsequent studies that contributed to the body of validity evidence largely omit descriptions of course formats; consequently, the validity evidence supporting the SATS instruments is strongest for more traditional course formats. Researchers seeking to use the SATS instruments in newer course formats (e.g., online, flipped, hybrid, or face-to-face transitioned to online due to substantial disruptions) should document validity evidence supporting their instrument choice. Validity evidence supporting such use may be similar to other published studies (e.g., analysis of internal consistency and factor structure), but may also use more sophisticated analyses to compare how students learning in these different formats interpret items, such as focus groups or checking for differential item functioning (e.g., Wilson, 2005).

Rigid pre/post structure. The SATS instruments are each available in two forms: a pre-course form to be administered prior to an intervention (such as a semester of statistics), and a post-course form to be administered after. Some items (Schau, 2003b) are the same on both forms (e.g., "Statistics is a complicated subject") while others are modified to change the tense (cf. "I plan to work hard in my statistics course" and "I worked hard in my statistics course"). This pre/post format reflects a reality of how education

research measuring change has been conducted for decades (Gal et al., 1997) but raises both logistical challenges and is not compatible with trends in education research toward more sophisticated longitudinal modeling (Sloane & Wilkins, 2017).

There are three waysthe SATS has been used in the literature where the rigid pre/post structure is problematic. While the term "introductory statistics" often implies one first course in statistics (e.g., Blair et al., 2018; GAISE College Report ASA Revision Committee, 2016; Schau & Emmioğlu, 2012), Millar and White (2014) describe the use of the SATS-36 in a two-course introductory statistics sequence. The pre-course and post-course versions were used in each course of the sequence (Millar & White, 2014), illustrating complications from using the SATS-36 longitudinally. Kerby and Wroughton (2017) were similarly challenged in their study, which measured students' attitudes at three time points in one semester (beginning, middle, and end). The post-course survey was used at the mid-semester administration (Kerby & Wroughton, 2017), but the language of the SATS-36 items made interpreting scores from that time point clumsy. Others have used the SATS-36 to provide a snapshot of students' attitudes using a single administration (e.g., Hood et al., 2012; Posner, 2014). Interpreting the SATS scores to provide a snapshot of students' attitudes without attending to a change, however, was not a use described in the development of the SATS (Schau et al., 1995), and it is not clear which form would be better suited to this purpose. A careful justification of the appropriateness of the form and the intended interpretations would be needed because this use is outside of the intended uses for the instruments.

Difficulty scale score interpretations. The 2014 *Standards* makes clear that validity is a property of the interpretations of scores from an instrument, not an intrinsic property of the instrument itself (AERA et al., 2014). For the SATS, scale scores are interpreted on a continuum of negative to neutral to positive: "Using the 7-point response scale, higher scores then correspond to more positive attitudes" (Schau, 2005b, p. 1). This interpretation is problematic for scores from the Difficulty scale, though it has also been noted in the literature that the name of the Difficulty component "is not ideal" (Ramirez et al., 2012, p. 60).

Interpreting higher scores and responses as indicating a more positive attitude toward statistics (after accounting for reverse-coded items) was first proposed by Schau et al. (1995) and seems to have been clarified in a presentation by Schau (2003c, p. 14) who noted, "Higher scores on the Difficulty component mean that students think that statistics is easier while lower scores mean that they think it is harder." Schau and Emmioğlu (2012) further clarified, for Difficulty, scores closer to 4 on the 7-point response are desirable: "we don't want students to believe that statistics is either too easy or too hard" (p. 93). Difficulty is thus conceptualized as a construct where higher scores from respondents are not necessarily indicative of better attitudes toward statistics. Despite this clarification, Schau and Emmioğlu at times refer to Difficulty scale scores using the same language as for other scales, characterizing pre-course scores for Affect and Difficulty together, saying "results showed that, on average, students entered these courses with neutral [attitudes] (Affect, Difficulty)" (2012, p. 86).

This clarification of the interpretation of the Difficulty scale scores as indicating the perceived difficulty of statistics on an easy to difficult continuum makes sense at face value but is ultimately complicated by inconsistent use of language in the literature and scoring guidelines. The result of these inconsistencies are scenarios that may lack a clear interpretation. For example, if the average Difficulty score for a class is a 4.0 pre-course and 5.0 post-course, this indicates that, on average, students view statistics as easier. Does this indicate an improvement in attitudes as it is more positive, or instead does it indicate a decline in attitudes because it is farther from neutral? Because validity is a property of the interpretations of scores, these challenges to the interpretability of the difficulty scales are threats to the use of the SATS instruments.

3. FURTHER COMPLICATIONS TO INTERPRETABILITY

The previous section described the five documented challenges to the interpretability of SATS scores when used with introductory statistics students. In addition to these challenges, two more should be recognized by researchers and users of the SATS instruments: the skew of the Effort scale score distribution and the resistance of scale scores to change. The dataset used for all analyses in this section is an

undergraduate SATS data warehouse. All data were analyzed in R version 3.6.0; notable packages used were *effsize* (Torchiano, 2019) for calculating effect sizes and *ggplot2* (Wickham, 2009) for some of the graphics.

Data were collected for the SATS data warehouse project between 2007 and 2010 under the direction of Schau and Bond (e.g., Bond et al., 2012; Schau & Emmioğlu, 2012). Pairwise complete data were available for the analyses presented here from approximately 2300 students enrolled in 120 sections of introductory statistics from 33 instructors: of these, approximately 90% were undergraduate students. Response rates were calculated using instructor-reported values for the number of students who could have started and ended the course. Instructors provided response rate information for 102 sections for pre-course administrations and 104 post-sections for post-course. Sections for which response rate information was not available were eliminated from these calculations; without further insight into why instructors did not provide response rate information, the effect on the reported response rates is not known. For the pre-course survey, the mean response rate was 0.84 with first and third quartile values of 0.76 to 0.95, and for the post-survey, the mean response rate was 0.77 with first and third quartile values of 0.62 to 0.93. These data are not a nationally representative sample; instead, participants were the students of statistics instructors who agreed to participate in the SATS data collection and might, therefore, be more attuned to statistics education, in general, and student attitudes, in particular.

3.1. EFFORT SCALE SKEW

Histograms illustrating the pre-course and post-course scores for each of the six SATS-36 scales for the SATS data warehouse sample are shown below in Figure 1.[4] A clear left skew is visible in both histograms for Effort and is more extreme for the pre-course scores than the post-course scores, reflecting truncation at the upper end of the response scale (a ceiling effect). The histograms for the scores of other scales do not exhibit a similar truncation and are instead reasonably described as approximately bell-shaped. The extreme truncation of the Effort scale scores has been reported previously in recommendations (e.g., Millar & Schau, 2010; Schau, 2008) and been alluded to in other studies (e.g., Millar & White, 2014; Schau & Emmioğlu, 2012). Some published studies have not included histograms of scale scores, instead opting to include boxplots of pre, post, and/or change scores, which do not illustrate the phenomenon as powerfully.

[4] The same colors are used for the constructs in all the figures and tables in this manuscript to facilitate following the results for a single construct across the different analyses.

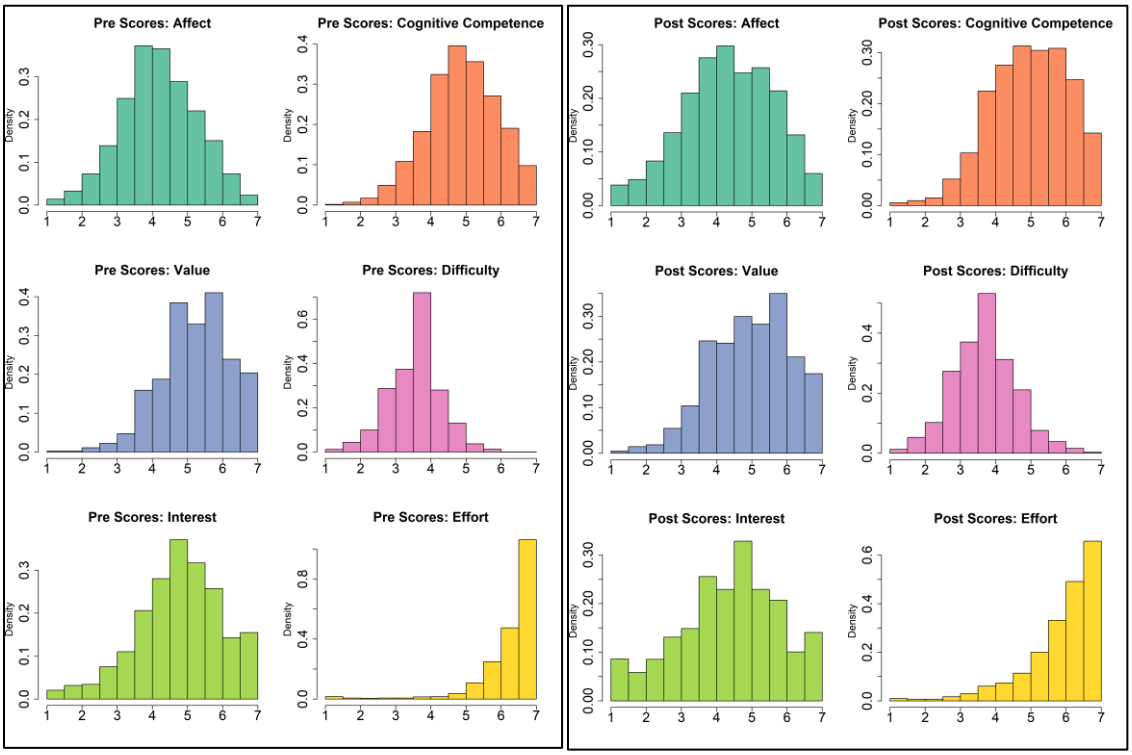

Figure 1. Histograms showing the pre-course (left) and post-course (right) scores for the SATS-36 constructs. The Effort construct is in the bottom right for both pre and post.

While research supporting the exact cause of this phenomenon with the SATS-36 Effort scale has not been conducted, it is plausible that this is an example of social desirability bias, wherein participants will self-report answers they think are aligned with the research goals (e.g., Dillman et al., 2014). Each of the items on the Effort scale appears is a statement that students may perceive as having a response that their instructor or the researchers view more favorably (e.g., "I plan to attend every statistics class session" [Schau, 2005, p. 3]). It is also plausible that students simply systematically over-estimate the amount of effort they will expend in statistics courses before the semester begins, which may be related to instructor emphasis on the first day of class (e.g., Posner, 2014). The ceiling effect in the pre-course Effort scores may account for the often noted decline in Effort scores (e.g., Bateiha et al., 2020; Ramirez & Bond, 2014; Schau & Emmioğlu, 2012).

Regardless of the cause, users of the SATS-36 instrument should be aware of a dramatic ceiling effect for the Effort scale and decreases in Effort scores from pre-course to post-course present in multiple studies with different samples of students. Interpretation of Effort scores for describing a change in attitudes toward statistics (either positive or negative) may not be possible in the scale's current form. Many types of statistical analyses used by practitioners of the SATS include assumptions of normality, and the Effort scale will often cause a violation of these assumptions. Researchers who are concerned about this phenomenon in their sample but who also wish to administer the complete SATS-36 instrument may consider reporting only descriptive statistics for the Effort scale and foregoing traditional inferential procedures or using nonparametric methods.

3.2. RESISTANCE OF SCALE SCORES TO CHANGE

Attitudes are recognized as an important outcome for introductory statistics courses both because they relate to student achievement and because they affect statistical behavior and choices outside the

introductory statistics classroom (Gal et al., 1997; Ramirez et al., 2012). Assessing changes in these attitudes is therefore a natural goal of research, and numerous instruments have been created to measure these changes (Nolan et al., 2012). The pre/post design of the SATS instruments makes this focus on measuring changes in attitude components explicit, and a ½-point change in scale scores for individuals has been recommended as a criterion for noting the importance of a change (Schau & Emmioğlu, 2012). Results exceeding a ½-point change in section means are not commonplace, though such change in individuals may be more common.

Two similar but distinct ½-point change thresholds have been used in the literature: thresholds for individuals and thresholds for group means. The rationale for a ½-point change in a scale score being deemed important for *individuals* is articulated by Schau and Emmioğlu:

> … we considered differences of about ½ point or more as important. That value represents a change of about 8% of the possible range in the Likert scale for each item. As examples, we describe how students could change their scores by ½ point on the Interest component, one of the components with 4 items (the fewest number of items in a component), and on the Value component (the component with 9 items, the greatest number of items). For Interest, students' scores would change by ½ point if they changed their Likert scale responses by 1 point on two items (half of the items in the component) or by 2 points on one item. Students could change their scores by slightly over ½ point on the Value component by changing their scale responses by 1 point on 5 items (about half of the items) or in several other ways. We believe that this degree of change is important. (2012, p. 88)

This explanation of the ½-point change in scale scores being important is clearly framed by Schau and Emmioğlu as pertaining to individuals' scores, not ½-point changes in class averages. This ½-point threshold for individuals has been used to classify individual students' changes in attitudes in recent studies (e.g., Chiesi & Primi, 2018; Kerby & Wroughton, 2017).

A ½-point change in *mean* scale score changes for statistics courses has also been discussed in the literature. The recommendation to use a ½-point change in mean scores as the threshold for practical significance predates the recommendation for individuals: "we do not consider mean gains of less than 0.5 to be of practical significance" (Millar & Schau, 2010, p. 1137). The existence of these two similar thresholds complicates interpretations. After articulating the rationale for the ½-point change threshold for individuals, Schau and Emmioğlu (2012) use a ½-point change threshold to examine the mean change in statistics attitudes from the pre-course to post-course administrations; they state, for example, "Most of the *Difficulty* section means did not change (within the range from -½ to +½ point)" (p. 90) and include indicator lines demarcating this region on their boxplots.

Using a ½-point change threshold for practical significance of section mean changes tends to result in few, if any, constructs exhibiting a change of practical significance (e.g., Schau & Emmioğlu, 2012). Bond (2008) noted that small or negligible changes in section mean scale scores from pre-course to post-course administrations may obscure an important finding: there are students within a statistics section that do change their attitudes. Examining the individual students may be the key to explaining why those changes occur and, consequently, why section mean changes are rather small. A similar phenomenon is noted by Kerby and Wroughton (2017). Among published pre/post SATS studies that report mean changes for each scale (i.e., the recommended use of the SATS), some have noted no changes for any components that exceed ½ point (e.g., Schau & Emmioğlu, 2012) while others have noted some changes. Chiesi and Primi (2009) found statistically significant differences for each of the four components in an Italian translation of the SATS-28, but only for the Value component did the change exceed ½ point. Other studies have noted changes that exceed ½ point, but only with negative changes for the Interest or Effort component (e.g., Bond et al., 2012; Ramirez & Bond, 2014). The mean differences from the SATS data warehouse are shown in Table 1 and illustrate largely the same phenomenon: none of the scales had changes that exceeded ½ point, but Interest and Effort have effect sizes that can be characterized as medium and large, respectively. Boxplots are shown in Figure 2 to illustrate the mean change scores for each of the 120 introductory statistics sections represented in the SATS data warehouse: this figure illustrates largely the same pattern in section mean scores noted by Schau and Emmioğlu (2012).

Table 1. The section mean differences (post-pre) and effect sizes for the N=120 sections in the SATS data warehouse sample.

Color	Component	Mean Difference	Cohen's *d*	Interpretation
	Affect	0.15	0.29	Small
	Cognitive Competence	0.11	0.23	Small
	Value	-0.20	-0.40	Small
	Difficulty	0.12	0.35	Small
	Interest	-0.36	-0.57	Medium
	Effort	-0.39	-1.23	Large

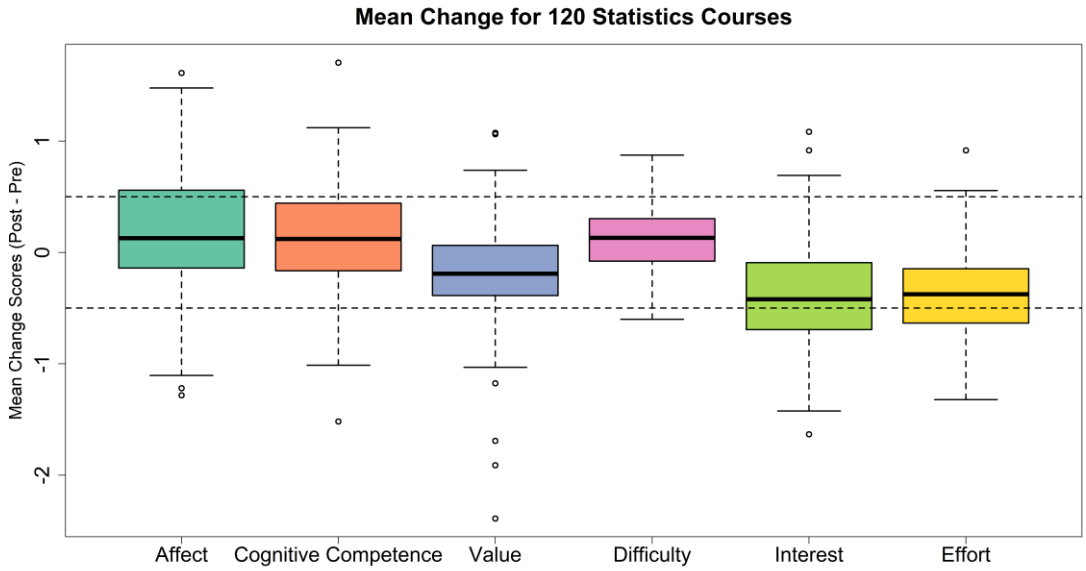

Figure 2. Overall mean change scores (post-pre) for each of the 120 sections of introductory statistics courses in the SATS data warehouse. This figure is modeled on Schau and Emmioğlu's (2012) Figure 2 and exhibits the same pattern. Horizontal dashed lines indicate the ½-point threshold for an important change.

Half-point differences have been noted in other studies that are beyond the recommended use for the SATS instruments. In their work with Irish pre-service teachers using the SATS-36 in a single administration snapshot design, Hannigan et al. (2013) found differences greater than ½ point in every component in at least one pair of years. Kerby and Wroughton administered the SATS-36 three times and identified a ½-point decrease in the effort scale from the pre-course administration to the mid-course administration for the Effort scale, but no other changes exceeded ½ point. Analyzing the change in SATS scores between semesters, Millar and White (2014) found no mean differences exceeding ½ point; however, as part of searching for ways to better capture changes,they examined the means of the absolute values of the differences and found values all exceeding ½ point. Across these and numerous other studies with students who have taken the SATS within the course of a semester or academic year, section mean changes exceeding the ½-point threshold identified as important have not been commonly identified. When such changes are identified, they are often in the Effort scale scores, which itself has challenges to interpretation that have been noted previously.

Although ½-point mean changes by *sections* are not observed in the SATS warehouse data (Figure 2 and Table 1), applying the ½-point change threshold to *individuals* reveals a notable finding. As shown in Table 2, for each of the SATS-36 scales at least half of the respondents had a change score that would be of practical significance, either less than -½ point or greater than +½ point. The distributions of these change scores are roughly symmetric (not shown) and suggest that more sophisticated analyses should be pursued to investigate changes rather than relying on section means.

Table 2. The percentage of all respondents in the SATS data warehouse with a change score that is less than -½ point or greater than +½ point. (n = 2300)

Color	Component	% with Change Score Less than -½ Point	% with Change Score Greater than +½ Point
	Affect	26%	34%
	Cognitive Competence	25%	31%
	Value	36%	20%
	Difficulty	20%	30%
	Interest	39%	17%
	Effort	31%	8%

The lack of ½-point changes in means might have several consequences. First, it might indicate a lack of ability of the SATS to detect changes in student attitudes properly. On the other hand it might reflect average change scores properly, indicating that we should not expect to see such changes after one semester of introductory statistics for some or all attitude components (e.g., Gal et al., 1997) or that individual positive and negative changes cancel each other out at the class level. It is not possible for us to tell if this is measurement error on behalf of the SATS or a true indication of only minor changes in average student attitudes toward statistics. This indicates that failing to observe a ½-point average change during a semester should not reflect negatively on the instructor or intervention used (if any) because such change may not be possible in only one semester or the instrument may not be able to detect it if it does occur. The creation of a ½-point rule by the SATS authors, however, indicates they believe the instruments should be able to detect such changes.

Further investigation into the ½-point change recommendations is needed. For individuals, no studies have examined the stability of individuals' scores in a test-retest setting (e.g., Crocker & Algina, 1986). For a ½-point change to be meaningful for individuals, the stability of scores for the SATS instruments administered in short succession (i.e., not at the beginning and end of the course) should be evaluated. For both ½-point change recommendations, a quirk in the SATS-36 instrument instructions may partially explain a lack of observed changes: "If you have no opinion, choose response 4" (Schau, 2003b, p. 1). This instruction to respondents may have the effect of biasing pre-course scores toward neutral because students may not yet have a firm opinion about statistics. For example, students may have a weak negative attitude at the beginning of a course due to lack of exposure and develop a strong neutral attitude by the end of the course but, due to the instructions, respond to items with a 4 each time, resulting in little to no observed change. The effect of this instruction, if any, has not been empirically investigated with the SATS.

In summary, it is unclear what type of ½-point change is considered meaningful for SATS data (individual change or group change) or if this ½-point threshold is appropriate. There are several plausible reasons for why ½-point section mean changes may be uncommon: such change may be difficult to elicit in a single semester, the changes may occur but be undetectable by the SATS instruments, or a more subtle reason such as more students responding "4" only in the pre-course surveys due to the instructions because they have little prior exposure to statistics.

4. USE WITH OTHER POPULATIONS

The previous discussion of challenges to interpretation of the SATS instrument scores were in the context of their use with undergraduate students in introductory statistics courses—the population for which they were developed (Schau et al., 1995) and for which there is evidence of validity (e.g., Nolan et al., 2012). A broad challenge to the use of the SATS instruments occurs when users seek to administer it to other populations. Interpreting scores from an instrument administered to a different population requires the collection of additional validity evidence (AERA et al., 2014). For example, the use of the SATS with introductory students in a large, public university in the American southeast would likely be similar to its use with students in a similar institution in the American southwest. As the studied population becomes more dissimilar from the original population; however, users of the instrument must take more care to support the intended interpretation of scores. For example, using the SATS with pre-service teachers in education courses might represent a substantial departure from the original intended population. In this section we describe uses of the SATS instruments with graduate students, with pre-service and in-service teachers, and with students who are not enrolled in introductory statistics courses.

4.1. USE WITH GRADUATE STUDENTS

The development of the SATS-28 was focused explicitly on students enrolled in introductory statistics courses at both the undergraduate and graduate level (Schau et al., 1995). The participants in the NGT process used to generate items were two graduate students, two undergraduate students, and two instructors of introductory statistics courses (Schau et al., 1995). Additionally, data were collected from undergraduate and graduate students in the pilot and operational administrations of the SATS-28 instrument (Schau et al., 1995). When the factor structure was examined, however, Schau et al. (1995) excluded graduate students, undergraduate students who were not seeking a degree, and international students because of differences in course characteristics and students' prior exposure to statistics. Taken together, the original development paper (Schau et al., 1995) provides some validity evidence supporting the use of the SATS with introductory statistics students broadly, but stronger evidence is provided for degree-seeking undergraduate students only.

Most studies that examined the factor structure of the SATS-28 and SATS-36 instruments included only undergraduate students and generally identified factor structures consistent with the proposed four and six factor solutions, respectively (e.g., Dauphinee et al., 1997; Hilton et al., 2004; Vanhoof et al., 2011). Few studies included graduate students in analyses of the structure of SATS instruments (e.g., Coetzee & Van der Merwe, 2010): one study with a sample of 342 students had approximately 80% graduate students, and an EFA yielded a two-factor solution for the SATS-28 (Cashin & Elmore, 2005). Although the use of EFA rather than CFA was a limitation of this study, the result is notable because of the sample of primarily graduate students. Because graduate students were excluded by the developers when performing CFA, the finding of a different factor structure with a group of primarily graduate students challenges the use of the SATS instruments with groups beyond undergraduate introductory statistics courses without additional validity evidence.

4.2. USE WITH TEACHERS

There has been a trend toward using the SATS instruments with both pre-service teachers (e.g., Batanero et al., 2005; Hannigan et al., 2013; Lipka & Hess, 2016; Nasser, 2004) and in-service teachers (e.g., Harshe & Abraham, 2017; Zapata Cardona & Rocha Salamanca, 2011). This is immediately problematic when the pre- or in-service teacher is not currently enrolled in a statistics class, due to the wording of several SATS items (as noted in section 4.3), but there are also other concerns.

Although pre-service teachers are often undergraduate students, they are a special sub-population of undergraduates that cannot be conflated with undergraduates in general. One key difference is that pre-service teachers are on a trajectory toward becoming instructors—and possibly instructors of statistics. As

pre-service teachers progress through their programs, they take on additional roles: in some situations, they are a student, and in other situations they are a teacher. While a first-year pre-service teacher enrolled in a general introductory statistics course may be quite similar to other undergraduate students, a pre-service teacher enrolled in their last-semester student teaching course may be more similar to in-service teachers in their responses.

To collect validity evidence for the use of the SATS instruments with teachers, Whitaker (2020) administered the SATS-36 instrument to a group of highly-experienced statistics teachers—the logical extreme of the trend toward using the SATS instruments with pre-service and in-service teachers. A critical case sample (Patton, 2002) of 12 exemplary secondary school teachers of statistics were recruited and asked to complete the SATS-36. The participants were primarily mid-career and late-career teachers, having an average of 30 years of teaching experience; for further details on this sample of statistics teachers see Whitaker (2020).

Informal discussions with the exemplary statistics teacher participants in this study raised concerns about the validity of interpretations of the SATS-36: for Difficulty, it was unclear if participants were to answer the items based on their own personal experiences as learners of statistics, their experiences as teachers of statistics, or how they thought their students would characterize the difficulty of statistics. This raises concerns about what instructions, if any, should be given to the respondents and about how to interpret the results of already-administered surveys with respondents who have an instructional role (whether pre-service or in-service teachers).

Another potential concern with the use of the SATS-36 instrument with teachers comes from comparing the scale scores from the exemplary statistics teachers in this study with scale scores from teachers in other studies. Specifically, Hannigan et al. (2013) report SATS-36 results from Irish pre-service teachers, disaggregated by year in school (i.e., Year 1 to Postgraduate). These results, together with the scores from the exemplary statistics teachers in this study, are shown in Figure 3. There are several inferences that may be made from the comparison of these scores that have implications for the validity of interpretations from the SATS-36 instrument.

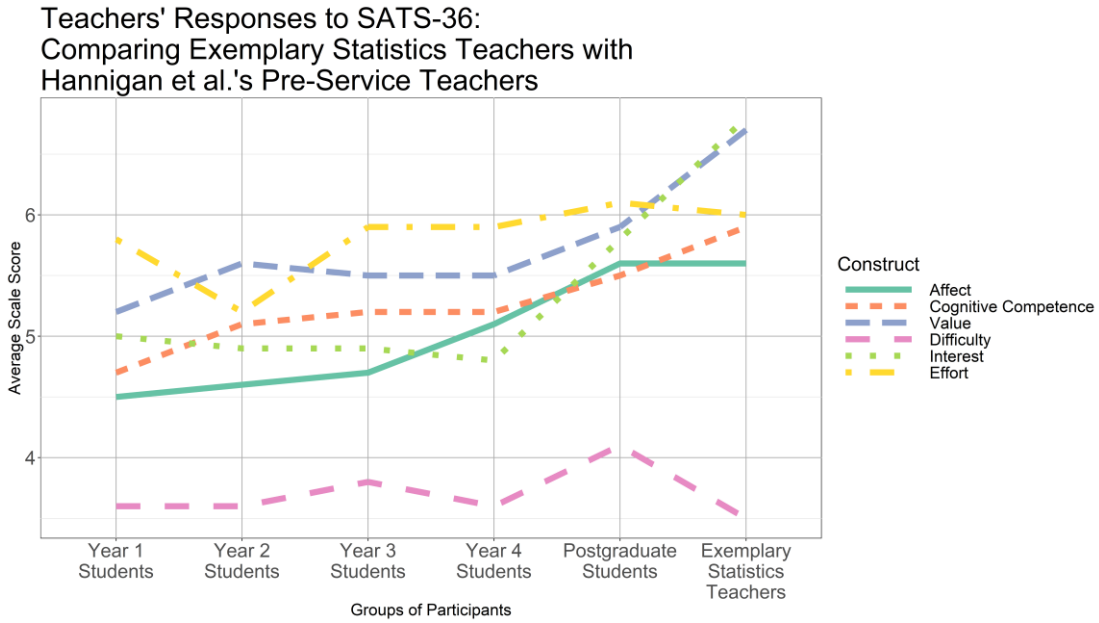

Figure 3. Pre-course SATS-36 average construct scores from Irish pre-service teachers, disaggregated by year in school as reported by Hannigan et al. (2013), together with results from the exemplary in-service teachers in this study.

A continuum of scores is visible in Figure 3 for most constructs, with Year 1 students generally having lower scores and Exemplary In-Service Teachers having higher scores. That the exemplary statistics teachers had the highest scores is consistent with other Hannigan et al. (2013) results and other studies that showed an increase in positive attitudes with more exposure to statistics (e.g., Gárcia-Martínez et al., 2015; Zapata Cardona & Rocha Salamanca, 2011). This trajectory of increasingly positive attitudes is a type of internal structure validity evidence and supports the validity of interpretations from the SATS-36 instrument with students and teachers.

Although a trajectory of increasingly positive attitudes from Year 1 students to exemplary in-service teachers can be seen in Figure 3, there are still unknowns about what this trajectory would look like if early and mid-career in-service teachers were examined. Moreover, the rather high scores from the exemplary statistics teachers raise concerns about possible ceiling effects when using the SATS-36 instrument, and the similarity of the pre-service and in-service teachers' scores raises concerns about whether the SATS-36 fails to capture real differences among these groups or if the groups really are quite similar.

4.3. OTHER USES OUTSIDE OF INTRODUCTORY STATISTICS COURSES

The growing interest in using the SATS instruments has also resulted in their use in courses beyond introductory statistics, although the development of the SATS instruments was firmly grounded in the context of introductory statistics courses (Schau et al., 1995), and most validity evidence supports the use of the instruments in only this context. Uses beyond introductory statistics courses fall into three categories: use of adapted versions of the SATS administered in other subjects for research purposes, use with students in discipline-specific courses containing statistics content (e.g., "Business Strategy"), and use with respondents who are not enrolled in any course containing appreciable statistics content. Adaptations of the SATS instruments result in new instruments that are not the SATS instruments for which documented validity evidence is available. Even though such adaptations may be reasonable, new validity evidence must be collected to support their intended use and interpretations of scores from the instruments.

As part of a study into the resistance of scale scores to change (see Section 3.2), the SATS-36 was adapted for use in other STEM subjects by Bond (2008) and undergraduate research assistants participating in Monmouth College's Summer Opportunities for Intellectual Activities (SOfIA) program. Items were modified by replacing the term "statistics" with another STEM discipline (i.e., biology, chemistry, computer science, mathematics, physics, astronomy, and psychology, based on the courses the adapted versions were to be administered in). The goal of this study was to collect validity evidence about the SATS-36 for use in statistics contexts and not to develop new instruments for widespread use in STEM classes. Based on data collected from 115 students from 15 different STEM course sections who completed both the pre- and post-course surveys, the results were similar to those discussed in Section 3.2. That is, many individual students exceeded the ½-point threshold on each construct—reflecting both positive and negative changes at the individual level—but few sections exhibited mean changes greater than ½ point. This is further evidence that either some components measured by the SATS-36 may not be changeable within a semester or that, if such change occurs, it may not be detectable by the instrument.

Out of a desire to use the SATS-36 with students enrolled in courses such as Business Strategy or Finance and Accounting, which may be related to introductory statistics in some way (e.g., Tempelaar, Gijselaers et al., 2007), some users have developed modified versions of SATS instruments. As with the STEM study, the primary changes were replacing the term "statistics" in items with the relevant course name. Unlike the STEM study, however, the goal was to interpret the students' scores in the context of these other subjects (rather than studying the properties of the SATS instruments in general). Such modifications represent a significant departure from an instrument, and additional validity evidence supporting the interpretations of scores from the new instrument is needed (Schau, 2008).

Lastly, researchers have used the SATS-28 (e.g., García-Martínez et al., 2015) and SATS-36 (e.g., Hannigan et al., 2013) to compare attitudes toward statistics across students in different years of a program. In these studies, however, only students in one year received explicit instruction in statistics, and the appropriateness of using the SATS-36 with the students who were not receiving instruction in statistics is

not immediately clear. While five of the constructs measured by the SATS-36 are defined in a way that is ostensibly appropriate for people who are not learners of statistics, the Effort construct positions the respondent as a learner of statistics (Schau, 2005b). Moreover, ten of the 36 items on the SATS-36 position the respondent as either a learner of statistics or enrolled in a statistics course. For example, item 11 states "I will have no idea of what's going on in this statistics course" and item 14 states "I plan to study hard for every statistics test" (Schau, 2005b, pp. 1–3). The use of these items and their associated scales with respondents who are not enrolled in a statistics course compromises the validity of the interpretations of the score because it is unclear as to how responses to items about planning to study for a statistics exam could be meaningful when the respondent knows such an exam will never occur.

5. CONCLUSION

The SATS-28 (Schau, 1992) and SATS-36 (Schau, 2003b) instruments were developed to measure changes in attitudes toward statistics in introductory statistics courses (primarily with undergraduate students) using a development process (Schau et al., 1995) described in the *SERJ* special issue on attitudes in statistics education (Nolan et al., 2012). At the time of their release, the development process for the SATS instruments was rigorous, and this has led to their widespread, sustained use for nearly three decades. Many contributions to the statistics education literature have drawn upon the SATS instruments in some way. Researchers, however, have increasingly sought to use them in novel contexts outside of the original intended use. In addition, the understanding of validity evidence and acceptable use of instruments has been updated and refined. The 2014 *Standards* note that documenting appropriate validity evidence for intended uses is "the joint responsibility of the test developer and the test user" (AERA et al., 2014, p. 13). Previous recommendations and cautions regarding the use of the SATS instruments have been made (e.g., Millar & Schau, 2010; Schau, 2008), but have not addressed all of the challenges discussed in this paper. These challenges reflect two realities: the use of the SATS in practice leads to problematic interpretations, and researchers are implementing the SATS in ways for which it was not originally intended, without collecting and documenting appropriate validity evidence.

For many researchers, the SATS family of instruments continues to be suitable for common uses. The authors of this manuscript intend to continue using the SATS-36 instrument for both research and program evaluation purposes in the near future. The biggest challenge for using the SATS instruments in their intended context of undergraduate introductory statistics courses relates to appropriate ways to interpret the scale scores. There are important differences between the constructs as operationalized in Eccles's EVT and the SATS instruments (Whitaker et al., 2018), and these limit the ability of the SATS instruments to draw on the extant body of EVT literature. Misalignment between the SATS family of instruments, SATS-M framework, and EVT framework may account for some of observed inconsistencies in the empirical factor structure (e.g., Cashin & Elmore, 2005; Vanhoof et al., 2011) and complicated score interpretations. Similarly, the constructs of Difficulty (Schau & Emmioğlu, 2012) and Effort may not be interpretable on the same positive to negative attitude continuum that is documented in the scoring guides (Schau, 2005a, 2005b) and appropriate for the other constructs. The instructions "If you have no opinion, choose response 4" (Schau, 2003b, p. 1) result in a 1-7 scale that cannot truly be interpreted as a continuum because students with no opinion at all and a truly neutral opinion cannot be distinguished based on their responses. Further research into distinguishing among students with no opinion, a neutral attitude, and contradictory attitudes may clarify future interpretations and lead to better identification of attitude changes within a semester.

Because of the complexity of interpreting changes in some constructs, users of the SATS instruments might choose to focus on performing inferential techniques for only a subset of the constructs—for example, the four constructs on the SATS-28 (e.g., Cladera et al., 2019). Focusing on an a priori subset would help mitigate the challenge of making multiple comparisons (e.g., Millar & Schau, 2010). Effort is a strong candidate for reporting only descriptive statistics, and similar arguments might be made for the Interest and Difficulty scales as well. Schau and Emmioğlu (2011) view Value, Cognitive Competence, and Interest as the most important of the six components, and these might be another reasonable a priori subset

to analyze. Of course, users should report sufficient information about the context of their study, the observed properties of the instrument for their sample, and follow the recommendations previously articulated for reporting (e.g., Millar & Schau, 2010; Millar et al., 2013; Schau, 2008; Schau & Millar, 2011). Decisions about analytic methods are often preferable to modifying an instrument by changing items, scoring algorithms, or scale composition. Great care must be taken when modifying an instrument or when interpreting the results of a study that has done so, and validity evidence should be documented for the modified instrument and alternate audience.

Another major challenge to the use of the SATS instruments occurs when researchers use the instruments outside of the intended context of undergraduate introductory statistics courses in a pre/post design. Owing in part to its popularity and the validity evidence available (Nolan et al., 2012; Ramirez et al., 2012), the SATS has been used in studies with different populations such as graduate students (e.g., Cashin & Elmore, 2005), students not enrolled in a course with statistics content (e.g., Hannigan et al., 2013), students enrolled in discipline-specific courses with some statistics content (e.g., Tempelaar, Gijselaers et al., 2007), pre-service teachers (e.g., Batanero et al., 2005; Hannigan et al., 2013), and in-service teachers (e.g., Zapata Cardona & Rocha Salamanca, 2011). Moreover, the SATS instruments have been used with research designs other than the intended pre/post design such as snapshots of attitudes (e.g., Hannigan et al., 2013; Hood et al., 2012) and longitudinal designs with more than two time periods (e.g., Kerby & Wroughton, 2017; Millar & White, 2014). Each of these uses requires additional validity evidence to be collected to document the appropriateness of using the SATS instruments in a new context: the results of the studies themselves contribute to the validity evidence, but they are incomplete. A thorough validity argument supporting each of these uses drawing from various sources of evidence (e.g., AERA et al., 2014) or another validity framework (cf. Krupa et al., 2019) should be conducted before interpretations from the SATS in these new contexts are accepted as appropriate. The use of SATS instruments in some contexts may also be inappropriate, and documenting sufficient validity evidence for their use in such contexts may not be possible.

Many instruments measuring attitudes toward statistics exist and have been described by Nolan et al. (2012): these instruments represent different conceptualizations and aspects of attitudes. Instruments to measure affective constructs in teachers are increasingly being developed, though none measure attitudes toward statistics in the same broad manner that the SATS instruments do with students. Zumbrun (2016) developed a scale that augments the SATS-36 (Schau, 2003b) with items aligned with the GAISE Framework (Franklin et al., 2007). The Self-efficacy for Teaching Statistics (SETS) instruments (Harrell-Williams et al., 2014, 2015) are designed to measure teachers' self-efficacy (e.g., Bandura, 1977, 1986) for teaching specific statistical concepts. The Statistics Teaching Inventory (STI) measures statistics teachers' beliefs and practices (Zieffler et al., 2012). Instruments developed for use with teachers might be substantially better-suited than the SATS instruments to answering research questions where participants are teachers. Work has also begun on a new family of instruments based on Expectancy Value Theory, designed to measure the attitudes toward statistics of students and instructors that do not require a pre/post class-focused structure (e.g., Batakci et al., 2018; Whitaker et al., 2019, 2018); data collection is on-going Unfried et al., 2018; Whitaker, 2021). Appropriate use of instruments, on-going collection and documentation of validity evidence, and the development of new instruments for measuring either new constructs or for use in new contexts will strengthen statistics education research.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/open-access-files.html

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Press.

Bandura, A. (1977). *Social learning theory*. Prentice Hall.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Batakci, L., Bolon, W., & Bond, M. E. (2018). A framework and survey for measuring instructors' motivational attitudes toward statistics. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics* (ICOTS10), Kyoto, Japan, July 8–14. http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_4J3.pdf

Batanero, C., Estrada, A., Díaz, C., & Fortuny, J. M. (2005). A structural study of future teachers' attitudes towards statistics. In M. Bosch (Ed.), *Proceedings of the Fourth Congress of the European Society for Research in Mathematics Education* (pp. 508–517). http://www.mathematik.uni-dortmund.de/~erme/CERME4/CERME4_WG5.pdf

Bateiha, S., Marchionda, H., & Autin, M. (2020). Teaching style and attitudes: A comparison of two collegiate introductory statistics classes. *Journal of Statistics Education, 28*(2), 154–164. https://doi.org/10.1080/10691898.2020.1765710

Bechrakis, T., Gialamas, V., & Barkatsas, A. N. (2011). Survey of Attitudes Toward Statistics (SATS): An investigation of its construct validity and its factor structure invariance by gender. *International Journal of Theoretical Educational Practice*, *1*(1), 1–15.

Blair, R. M., Kirkman, E. E., & Maxwell, J. W. (2018). *Statistical abstract undergraduate programs in the mathematical sciences in the united states: 2015 CBMS survey*. American Mathematical Society. http://www.ams.org/cbms-survey/cbms2015-Report.pdf

Bond, M. E. (2008). Survey of Attitudes Toward Statistics: An exploratory look. *2008 JSM Proceedings*, (pp. 3758-3760). http://www.statlit.org/pdf/2008BondASA.pdf

Bond, M. E., Perkins, S. N., & Ramirez, C. (2012). Students' perceptions of statistics: An exploration of attitudes, conceptualizations, and content knowledge of statistics. *Statistics Education Research Journal*, *11*(2), 6–25. https://doi.org/10.52041/serj.v11i2.325

Carillo, K., Galy, N., Guthrie, C., & Vanhems, A. (2016). « J'aime pas les stats! » Mesure et analyse de l'attitude à l'égard du cours de statistique dans un école de management ["I hate statistics!" — The measure and analysis of attitudes towards statistics in a management school]. *Statistique et Enseignement*, *7*(1), 3–31. http://statistique-et-enseignement.fr/article/view/537

Carlson, K. A., & Winquist, J. R. (2011). Evaluating an active learning approach to teaching introductory statistics: A classroom workbook approach. *Journal of Statistics Education*, *19*(1), 1–23. https://doi.org/10.1080/10691898.2011.11889596

Carmona Márquez, J. (2004). Una revisión de las evidencias de fiabilidad y validez de los cuestionarios de actitudes y ansiedad hacia la estadística. *Statistics Education Research Journal, 3*(1), 5–28. http://iase-web.org/documents/SERJ/SERJ3(1)_marquez.pdf

Carnell, L. J. (2008). The effect of a student-designed data collection project on attitudes toward statistics. *Journal of Statistics Education*, *16*(1). http://jse.amstat.org/v16n1/carnell.html

Cashin, S. E., & Elmore, P. B. (2005). The survey of attitudes toward statistics acale: A construct validity study. *Educational and Psychological Measurement*, *65*(3), 509–524. https://doi.org/10.1177/0013164404272488

Chance, B., Wong, J., & Tintle, N. (2016). Student performance in curricula centered on simulation-based inference: A preliminary report. *Journal of Statistics Education*, *24*(3), 114–126. https://doi.org/10.1080/10691898.2016.1223529

Chiesi, F., & Primi, C. (2009). Assessing statistics attitudes among college students: Psychometric properties of the Italian version of the Survey of Attitudes Toward Statistics (SATS). *Learning and Individual Differences*, *19*(2), 309–313. https://doi.org/10.1016/j.lindif.2008.10.008

Chiesi, F., & Primi, C. (2010). Cognitive and non-cognitive factors related to students' statistics achievement. *Statistics Education Research Journal*, *9*(1), 6–26. https://doi.org/10.52041/serj.v9i1.385

Chiesi, F., & Primi, C. (2018). What happens when attitudes toward statisitcs change (increases vs decrease) during the course? In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics* (ICOTS10), Kyoto, Japan, July 8–14. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_7D1.pdf

Cladera, M., Rejón-Guardia, F., Vich-i-Martorell, G. À., & Juaneda, C. (2019). Tourism students' attitudes toward statistics. *Journal of Hospitality, Leisure, Sport & Tourism Education*, *24*, 202–210. https://doi.org/10.1016/j.jhlste.2019.03.002

Coetzee, S., & Van der Merwe, P. (2010). Industrial psychology students' attitudes towards statistics. *SA Journal of Industrial Psychology*, *36*(1), 8 pages. https://doi.org/10.4102/sajip.v36i1.843

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://link.springer.com/article/10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. https://doi.org/10.1037/h0040957

Dauphinee, T. L., Schau, C., & Stevens, J. J. (1997). Survey of attitudes toward statistics: Factor structure and factorial invariance for women and men. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(2), 129–141. https://doi.org/10.1080/10705519709540066

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method* (4th edition). Wiley.

Eccles, J. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 75–145). W.H. Freeman.

Eccles, J. S. (2014). Expectancy-value theory. In R. Eklund & G. Tenenbaum (Eds.), *Encyclopedia of Sport and Exercise Psychology*. https://doi.org/10.4135/9781483332222.n110

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*, 109–132. https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.53.100901.135153

Emmioglu Sarikaya, E., Ok, A., Capa Aydin, Y., & Schau, C. (2018). Turkish version of the Survey of Attitudes Toward Statistics: Factorial structure invariance by gender. *International Journal of Higher Education*, *7*(2), 121–127. https://doi.org/10.5430/ijhe.v7n2p121

Estrada, A. (2002). *Análisis de las actitudes y conocimientos estadísticos elementales en la formación del profesorado. [Analysis of attitudes and elementary statistical knowledge in training teachers]* (Universidad Autónoma de Barcelona). https://www.tdx.cat/handle/10803/4697

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 curriculum framework*. www.amstat.org/education/gaise

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

Gal, I., Ginsburg, L., & Schau, C. (1997). Monitoring attitudes and beliefs in statistics education. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 37–51). IOS Press. http://iase-web.org/Books.php?p=book1

García-Martínez, J. A., Fallas-Vargas, M. A., & Romero-Hernández, A. (2015). Las actitudes hacia la estadística del estudiantado de orientación. *Revista Electrónica Educare*, *19*(1), 25–41. https://doi.org/10.15359/ree.19-1.2

Gundlach, E., Richards, K. A. R., Nelson, D., & Levesque-Bristol, C. (2015). A comparison of student attitudes, statistical reasoning, performance, and perceptions for web-augmented traditional, fully online, and flipped sections of a statistical literacy class. *Journal of Statistics Education*, *23*(1). http://jse.amstat.org/v23n1/gundlach.pdf

Hannigan, A., Gill, O., & Leavy, A. M. (2013). An investigation of prospective secondary mathematics teachers' conceptual knowledge of and attitudes towards statistics. *Journal of Mathematics Teacher Education*, *16*(6), 427–449. https://doi.org/10.1007/s10857-013-9246-3

Harrell-Williams, L. M., Sorto, M. A., Pierce, R. L., Lesser, L. M., & Murphy, T. J. (2014). Validation of scores from a new measure of preservice teachers' self-efficacy to teach statistics in the middle grades. *Journal of Psychoeducational Assessment*, *32*(1), 40–50. https://doi.org/10.1177/0734282913486256

Harrell-Williams, L. M., Sorto, M. A., Pierce, R., Lesser, L. M., & Murphy, T. J. (2015). Identifying statistical concepts associated with high and low levels of self-efficacy to teach statistics in middle grades. *Journal of Statistics Education*, *23*(1). http://jse.amstat.org/v23n1/harrell-williams.pdf

Harshe, D., & Abraham, D. (2017). A study of attitudes of teaching faculty and postgraduate residents at a tertiary care teaching hospital toward biostatistics. *Muller Journal of Medical Sciences and Research*, *8*(1), 10–14. https://doi.org/10.4103/0975-9727.199362

Hilton, S. C., Schau, C., & Olsen, J. A. (2004). Survey of Attitudes Toward Statistics: Factor structure invariance by gender and by administration time. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(1), 92–109. https://doi.org/10.1207/S15328007SEM1101_7

Hommik, C., & Luik, P. (2017). Adapting the Survey of Attitudes Towards Statistics (SATS-36) for Estonian secondary school students. *Statistics Education Research Journal*, *16*(1), 228–239. https://doi.org/10.52041/serj.v16i1.229

Hood, M., Creed, P. A., & Neumann, D. L. (2012). Using the expectancy value model of motivation to understand the relationship between student attitudes and achievement in statistics. *Statistics Education Research Journal*, *11*(2), 72–85. https://doi.org/10.52041/serj.v11i2.330

Kerby, A. T., & Wroughton, J. R. (2017). When do students' attitudes change? Investigating student attitudes at midterm. *Statistics Education Research Journal*, *16*(2), 476–486. https://doi.org/10.52041/serj.v16i2.202

Khavenson, T., Orel, E., & Tryakshina, M. (2012). Adaptation of Survey of Attitudes Towards Statistics (SATS 36) for Russian Sample. *Procedia - Social and Behavioral Sciences*, *46*, 2126–2129. https://doi.org/10.1016/j.sbspro.2012.05.440

Krupa, E. E., Carney, M., & Bostic, J. (2019). Argument-based validation in practice: Examples from mathematics education. *Applied Measurement in Education*, *32*(1), 1–9. https://doi.org/10.1080/08957347.2018.1544139

Lipka, O., & Hess, I. (2016). Attitudes toward statistics studies among students with learning disabilities. *Numeracy*, *9*(2) Article 7. https://doi.org/10.5038/1936-4660.9.2.7

Luh, W. M., Guo, J. H., & Wisenbaker, J. M. (2004). Difference of attitudes toward statistics between cadets and college students. *Paper Presented at the 10th International Congress on Mathematical Education*, *July 4–11, Denmark*. http://iase-web.org/documents/papers/icme10/Luh.pdf

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). American Council on Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Millar, A. M., & Schau, C. (2010). Assessing students' attitudes: The good, the bad, and the ugly. *2010 JSM Proceedings* (pp. 1133–1143). http://www.statlit.org/pdf/2010MillarSchauASA.pdf

Millar, A. M., & White, B. J. G. (2014). How do attitudes change from one stats course to the next? In K. Makar & R. Gould (Eds.), *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics* (ICOTS9) Flagstaff, Arizona, July 13–18. https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_1F2_MILLAR.pdf

Millar, A. M., White, B. J., & Romo, R. (2013). Comparing apples with apples: Assessing student attitudes in the presence of regression to the mean. *2013 JSM Proceedings*. Presented at the Joint Statistics Meetings, San Francisco, CA. http://ec.msvu.ca:8080/xmlui/handle/10587/1284

Moore, C. M. (1987). *Group techniques for idea building*. SAGE Publications.

Nasser, F. M. (2004). Structural model of the effects of cognitive and affective factors on the achievement of Arabic-speaking pre-service teachers in introductory statistics. *Journal of Statistics Education*, *12*(1). http://ww2.amstat.org/publications/jse/v12n1/nasser.html

Nolan, M. M., Beran, T., & Hecker, K. G. (2012). Surveys assessing students' attitudes toward statistics: A systematic review of validity and reliability. *Statistics Education Research Journal*, *11*(2), 103–123. https://doi.org/10.52041/serj.v11i2.333

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3 ed). SAGE Publications.

Persson, I., Kraus, K., Hansson, L., & Wallentin, F. Y. (2019). Confirming the structure of the Survey of Attitudes Toward Statitics (SATS-36) by Swedish Students. *Statistics Education Research Journal*, *18*(1), 83–93. https://doi.org/10.52041/serj.v18i1.151

Posner, M. A. (2011). The impact of a proficiency-based assessment and reassessment of learning outcomes system on student achievement and attitudes. *Statistics Education Research Journal*, *10*(1), 3–14. https://doi.org/10.52041/serj.v10i1.352

Posner, M. A. (2014). A fallacy in student attitude research: The impact of the first class. In K. Makar & R. Gould (Eds.), *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics* (ICOTS9), Flagstaff, Arizona, July 13–18. https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_1F3_POSNER.pdf

Ramirez, C., & Bond, M. (2014). Comparing attitudes toward statistics among students enrolled in project-based and hybrid statistics courses. In K. Makar & R. Gould (Eds.), *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics* (ICOTS9), Flagstaff, Arizona, July 13-18. https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_1F4_RAMIREZ.pdf?1405041565

Ramirez, C., Schau, C., & Emmioğlu, E. (2012). The importance of attitudes in statistics education. *Statistics Education Research Journal*, *11*(2), 57–71. https://doi.org/10.52041/serj.v11i2.329

Ratanaolarn, T. (2016). The development of a structural equation model of graduate students' statistics achievement. *The Journal of Behavioral Science*, *11*(2), 153–168. https://so06.tci-thaijo.org/index.php/IJBS/article/view/63288

Rosli, M. K., & Maat, S. M. (2017). Attitude towards statistics and performance among post-graduate students. In M. Puteh, N. Z. A. Hamid, & N. H. Adenan (Eds.), *AIP Conference Proceedings, 1847*(1) 030004. https://doi.org/10.1063/1.4983881

Schau, C. (1992). *Survey of Attitudes Toward Statistics (SATS-28)*. http://evaluationandstatistics.com/

Schau, C. (2003a). Students' attitudes: The "other" important outcome in statistics education. In proceedings of the *2003 Joint Statisitcal Meetings,* 3 (pp. 3673–3683). http://www.statlit.org/pdf/2003SchauASA.pdf

Schau, C. (2003b). *Survey of Attitudes Toward Statistics (SATS-36)*. http://evaluationandstatistics.com/

Schau, C. (2003c, August). *Students' attitudes: The "other" important outcome in statistics education*. Paper presented at the Joint Statistical Meetings, San Francisco, CA. https://irp-cdn.multiscreensite.com/281322c3/files/uploaded/JSM2003.pdf

Schau, C. (2005a). *Scoring the SATS-28©*. https://irp-cdn.multiscreensite.com/281322c3/files/uploaded/SATS28Scoring.pdf

Schau, C. (2005b). *Scoring the SATS-36©*. https://irp-cdn.multiscreensite.com/281322c3/files/uploaded/Final36scoring.pdf

Schau, C. (2008). Common issues in SATS© research. *2008 JSM Proceedings* (pp. 3761–3763). http://www.statlit.org/pdf/2008SchauASA.pdf

Schau, C., & Emmioğlu, E. (2011). Changes in US students' attitudes toward statistics across introductory statistics courses. *Proceedings of the 58th World Statistics Congress of the International Statistical Institute* (pp. 2802–2808). http://www.2011.isiproceedings.org/papers/650335.pdf

Schau, C., & Emmioğlu, E. (2012). Do introductory statistics courses in the United States improve students' attitudes? *Statistics Education Research Journal*, *11*(2), 86–94. https://doi.org/10.52041/serj.v11i2.331

Schau, C., & Millar, A. M. (2011). Letter to the Editor. *Statistics Education Research Journal*, *10*(2), 77–79. https://doi.org/10.52041/serj.v10i2.349

Schau, C., Stevens, J., Dauphinee, T. L., & Del Vecchio, A. (1995). The development and validation of the Survey of Attitudes Toward Statistics. *Educational and Psychological Measurement*, *55*(5), 868–875. https://doi.org/10.1177/0013164495055005022

Sloane, F. C., & Wilkins, J. L. M. (2017). Aligning statistical modeling with theories of learning in mathematics education research. In J. Cai (Ed.), *Compendium for research in mathematics education* (pp. 183–207). National Council of Teachers of Mathematics.

Sorge, C., & Schau, C. (2002). *Impact of engineering students' attitudes on achievement in statistics*. Paper presented at the American Educational Research Association Annual Meeting, New Orleans. https://web.archive.org/web/20170810073100/http://evaluationandstatistics.com/AERA2002.pdf

Strobl, C., Leisch, F., Dittrich, C., Seiler, C., & Hackensperger, S. (2010). Measurement and predictors of a negative attitude towards statistics among LMU students. In T. Kneib & G. Tutz (Eds.), *Statistical modelling and regression structures: Festschrift in honour of Ludwig Fahrmeir* (pp. 217–230). Physica.

Tempelaar, D. T., Gijselaers, W. H., Schim van der Loeff, S., & Nijhuis, J. F. H. (2007). A structural equation model analyzing the relationship of student achievement motivations and personality factors in a range of academic subject-matter areas. *Contemporary Educational Psychology*, *32*(1), 105–131. https://doi.org/10.1016/j.cedpsych.2006.10.004

Tempelaar, D. T., Van der Loeff, S. S., & Gijselaers, W. H. (2007). A structural equation model analyzing the relationship of students' attitudes toward statistics, prior reasoning abilities and course performance. *Statistics Education Research Journal*, *6*(2), 78–102. https://iase-web.org/documents/SERJ/SERJ6(2)_Tempelaar.pdf

Torchiano, M. (2019). effsize: Efficient Effect Size Computation (Version R package version 0.7.6). https://doi.org/10.5281/zenodo.1480624

Unfried, A., Kerby, A., & Coffin, S. (2018). Developing a student survey of motivational attitudes toward statistics. *2018 JSM Proceedings*. Joint Statistical Meetings 2018, Vancouver, Canada.

Vanhoof, S., Kuppens, S., Sotos, A. E. C., Verschaffel, L., & Onghena, P. (2011). Measuring statistics attitudes: Structure of the Survey of Attitudes Toward Statistics (SATS-36). *Statistics Education Research Journal*, *10*(1), 35–51. https://doi.org/10.52041/serj.v10i1.354

Vendramini, C. M. M., Silva, C. B., Kataoka, V. Y., & Cazorla, I. M. (2011). Validity evidences of the attitudes towards statistics scale SATSPORTUGUÊS: A study with Brazilian students. *Proceedings of the 58th World Statistical Congress*, 4 (pp. 5997–6000). International Statistical Institute. http://2011.isiproceedings.org/papers/950155.pdf

Whitaker, D. (2020). Epistemic and mathematical beliefs of exemplary statistics teachers. *International Journal for Mathematics Teaching and Learning*, *21*(2), 119–142. https://www.cimt.org.uk/ijmtl/index.php/IJMTL/article/view/253

Whitaker, D. (2021, June). *Developing and revising the student survey of motivational attitudes toward statistics: Results from a pilot study*. Presentation at the Statistics Society of Canada 2021 Annual Meeting, Virtual.

Whitaker, D., & Gorney, K. (2017). *Surveys of attitudes about statistics: An analysis of items* [Poster]. 39th Annual Conference of the North American Chapter of the International Group for the Psychology of Mathematics Education (PME-NA), Indianapolis, IN.

Whitaker, D., Unfried, A., & Batakci, L. (2018). A framework and survey for measuring students' motivational Attitudes toward statistics. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics* (ICOTS10), Kyoto, Japan, July 8-14. http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_C200.pdf

Whitaker, D., Unfried, A., & Bond, M. (2019). Design and validation arguments for the Student Survey of Motivational Attitudes toward Statistics (S-SOMAS) instrument. In J. D. Bostic, E. E. Krupa, & J. C. Shih (Eds.), *Assessment in mathematics education contexts: Theoretical frameworks and new directions* (1st ed., pp. 120–146). Routledge. http://dc.msvu.ca:8080/xmlui/handle/10587/2125

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.

Wise, S. L. (1985). The development and validation of a scale measuring attitudes toward statistics. *Educational and Psychological Measurement*, *45*(2), 401–405. https://journals.sagepub.com/doi/abs/10.1177/001316448504500226

Xu, C., & Schau, C. (2019). Exploring method effects in the six-factor structure of the Survey of Attitudes Toward Statistics (SATS-36). *Statistics Education Research Journal*, *18*(2), 39–53. https://doi.org/10.52041/serj.v18i2.139

Zapata Cardona, L., & Rocha Salamanca, P. (2011). *Actitudes de profesores hacia la estadística y su enseñanza*. Presented at the XIII CIAEM-IACME, Recife, Brazil. http://www.cimm.ucr.ac.cr/ocs/files/conferences/1/schedConfs/1/papers/1712/public/1712-8255-1-PB.pdf

Zhang, Y., Shang, L., Wang, R., Zhao, Q., Li, C., Xu, Y., & Su, H. (2012). Attitudes toward statistics in medical postgraduates: Measuring, evaluating and monitoring. *BMC Medical Education*, *12*(117). https://doi.org/10.1186/1472-6920-12-117

Zieffler, A., Park, J., Garfield, J., delMas, R., & Bjornsdottir, A. (2012). The statistics teaching inventory: A survey on statistics teachers' classroom practices and beliefs. *Journal of Statistics Education*, *20*(1), 1–29. https://amstat.tandfonline.com/doi/abs/10.1080/10691898.2012.11889632

Zumbrun, C. M. (2016). Crossing the boundary: Results from the Teacher Attitudes and Beliefs Toward Statistics Survey (TABSS). *Proceedings of the 38th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education,* Tucson, AZ (pp. 1004–1007). https://www.pmena.org/pmenaproceedings/PMENA%2038%202016%20Proceedings.pdf#page=1018

DOUGLAS WHITAKER
Mount Saint Vincent University
douglas.whitaker@msvu.ca