Design and Validation Arguments for the Student Survey Of Motivational Attitudes toward Statistics (S-SOMAS) Instrument

Douglas Whitaker

Mount Saint Vincent University

Alana Unfried

California State University, Monterey Bay

Marjorie Bond

Monmouth College

Abstract

This chapter describes the development plan for a new survey designed to measure students' attitudes toward statistics aligned with Expectancy Value Theory, the S-SOMAS. Included in this description are the context and motivation for developing a new instrument, a brief overview of the theoretical framework, and the validity evidence to be collected to support the intended uses of the S-SOMAS. The S-SOMAS is currently in development and pilot data have been collected; as such, this chapter includes both planned and enacted elements of instrument development. This chapter also provides readers with an illustration about the early stages of validation work and some choices that instrument developers may encounter.

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. Bostic, E. Krupa, & J. Shih (Eds.), *Assessment in Mathematics Education Contexts: Theoretical Frameworks and New Directions* (1st ed., pp. 120–146). New York, NY: Routledge.

Design and Validity Arguments for the Surveys Of Motivational Attitudes toward Statistics (SOMAS) Instruments

Attitudes toward statistics are important outcomes of statistics courses because they are linked with student achievement and because they affect students' lasting impressions of the discipline (Gal, Ginsburg, & Schau, 1997; Ramirez, Schau, & Emmioğlu, 2012). In short, "People forget what they do not use. But attitudes 'stick'" (Ramirez et al., 2012, p. 57). As the field of statistics education matures, instruments for measuring attitudes have been precipitated by an evolving understanding about the learning and teaching of statistics and the needs of researchers. The Surveys Of Motivational Attitudes towards Statistics (SOMAS) project is developing a family of instruments for assessing attitudes toward statistics for both of the aforementioned reasons: significant advances in the discipline have occurred since the release of the last major attitude assessment instrument in statistics education and current instruments are not designed for contemporary research needs.

This chapter outlines the development process of an instrument designed to be taken by students enrolled in undergraduate introductory-level statistics courses (S-SOMAS), including a brief overview of the motivation for developing a new instrument, the instrument development plan, and the validity evidence that will support the intended uses of the instrument. While the SOMAS project also includes instruments to be completed by instructors, only the S-SOMAS is described here. The SOMAS project is an on-going instrument development project and specific details and goals may evolve or change as the project continues.

Validity evidence to be collected supporting the intended uses of the S-SOMAS instrument is organized using the sources of evidence detailed in the 2014 *Standards for Educational and Psychological Testing* document (AERA et al., 2014). This chapter seeks to make explicit claims that may remain implicit in some validity studies. Together with a clearly-articulated purpose for developing and an using an instrument, these explicit claims and evidence statements represent a roadmap for the validation studies to be conducted for the S-SOMAS. By presenting the validity evidence claims and clear motivation for developing the S-SOMAS in one location, we hope to illuminate the process of gathering validity evidence during instrument development.

Motivation for Developing a New Instrument

Instruments for measuring attitudes toward statistics have been developed and used for decades. Today, the Survey of Attitudes Toward Statistics (SATS) instruments (Schau, 1992, 2003b) are among the most widely-used and researched attitude surveys in the statistics education literature since their initial release and subsequent updates (Nolan, Beran, & Hecker, 2012). One factor that contributes to the popularity of the SATS is that it is claimed to be congruent with Expectancy Value Theory when few extant instruments in statistics education are aligned with an established educational theory (Ramirez et al., 2012; Schau, Millar, & Petocz, 2012). There are two versions of the SATS: a 28-item version that measures four constructs related to statistics attitudes (SATS-28; Schau, 1992), and a 36-item version that is the SATS-28 with eight additional items measuring two additional constructs (SATS-36; Schau, 2003b). The SATS-36 contains the SATS-28 as a subset of items. When statements apply to both instruments, the term SATS will be used.

The SATS have been widely used (e.g. Dauphinee, Schau, & Stevens, 1997; Schau & Emmioğlu, 2012; Vanhoof, Kuppens, Sotos, Verschaffel, & Onghena, 2011), and numerous limitations and challenges have been identified as it has been studied more. Notably, the alignment of the SATS instruments to EVT was post-hoc and did not guide the development of the instruments (Schau, 2003a), limiting the extent to which the instruments measure the theoretical constructs that comprise the EVT framework. The six constructs measured by the SATS-36 are Affect, Cognitive Competence, Value, Difficulty, Interest, and Effort (Schau, 2003b). Schau (2003b) noted that these constructs are not the same as the EVT constructs described in the EVT literature (e.g. Eccles, 1983, 2014). While the alignment of the SATS instruments to EVT was innovative in statistics education when they were developed, their widespread use has highlighted areas for future research and improvement.

Additionally, the SATS instruments were developed for use with students in introductory statistics classes at the undergraduate and graduate level (Schau, Stevens, Dauphinee, & Vecchio, 1995), but contemporary research in statistics education has expanded to include research with many populations beyond undergraduate and graduate students such as instructors of statistics, students in grades K-12 (though primarily focused on students in high school), teachers of students in grades K-12, pre-service teachers, and learners of statistics in specialized contexts such as health sciences.

Adapting the SATS instruments so that they are suitable for all contexts would not be a straightforward task: some items included on the SATS instruments are written in such a way that they presuppose that the respondent is enrolled in a statistics course. Additionally, two forms for each SATS instrument exist: a version intended to be administered near the beginning of the course, and a version intended to be administered near the end of the course. Adapting the instrument for use with groups beyond students enrolled in tertiary first courses in statistics would require at least three substantial types of changes. First, at least 25% of the items on the SATS-36 that explicitly position the respondent as being enrolled in a course because the items reference courses, tests, and problems and would need to be heavily modified or replaced.

Second, other items implicitly position the respondent as being a student (e.g. by asking about future employability) which would require more than minor alterations. Third, the way some scales assess the underlying latent constructs would need to be reconceptualized (e.g. all four items in the Effort construct reference enrollment in some way). While some researchers have used the SATS instruments with other populations of interest, there is a lack of validity evidence for these uses because of the aforementioned challenges and concerns, and piecemeal modification of the instrument is problematic. Instead, new instruments are needed that have been designed for use with these populations.

Developing the S-SOMAS Instrument

The researchers working on the SOMAS project (SOMAS project team) all have prior experience researching attitudes in statistics education and using the SATS instruments. Before the decision to develop a new family of instruments was made, the team had considered working with the lead developer of the SATS instruments to update, modify, and adapt them to address concerns and for use with new populations. However, the team chose to develop new instrument because 1) using an existing educationl framework to guide development will result in stronger instruments and 2) previous experiences using and researching the SATS instruments has suggested that there are other challenges to the using these instruments beyond those already reported in the literature. The SOMAS project team instead chose to draw on their accumulated knowledge of the SATS to inform the development of new instruments.

The SOMAS project team is currently developing three online instruments: one intended to measure the attitudes toward statistics of undergraduate students enrolled in introductory statistics courses, the S-SOMAS, and two intended to be completed by instructors, allowing researchers the ability to analyze attitudes in an introductory statistics course in a broader and more cohesive way than previously possible. While the projet team anticipates that there will be a desire to use the instruments with other populations, restricting the focus to introductory college-level statistics will help ensure the development of high-quality instruments with strong validity evidence supporting their use with this population rather than attempting to meet all possible project goals simultaneously.

This chapter will focus on the development of the S-SOMAS instrument, intended to be used with undergraduate students. A brief overview of the development process – both planned and enacted – will be presented first. Then, the theoretical framework adopted for use with the SOMAS instruments, EVT, will be briefly described. Lastly, the plan for collecting validity evidence supporting the use and interpretations of the S-SOMAS will be presented.

Overview of Development Process

The development of the S-SOMAS instrument is a multi-year process that began May 2017 and is anticipated to conclude in 2020 (Figure 1).

[INSERT FIGURE 1 HERE]

There are numerous factors that contribute to the length of time needed for developing an instrument, including both the realities of professional life and principles of instrument development. The development process illustrated in Figure 1 reflects an iterative development process wherein data collected leads to revisions of work done previously as well as a strong focus on collecting validity evidence to support the intended uses of the S-SOMAS instrument. Note that Figure 1 is meant to illustrate the sequencing of key events in the development process rather than a timeline to be interpreted on a linear scale. The primary activities illustrated in Figure 1 are the development of and revisions to the theoretical framework (described below), the development of and revisions to items, data collection using assembled forms, and analysis and interpretation of collected to data to inform revisions. In addition to these primary activities, the collection of other data through focus groups and subject matter expert (SME) review are included. Each of these development activities supports collection of validity evidence for the use and interpretations of the S-SOMAS instruments and represents an intentional development decision in the planning phase. The SOMAS project team anticipates having a finalized version of the S-SOMAS instrument prior to Fall 2020: this would represent the initial instrument development, two rounds of pilot data collection and revisions, and the operational administration to ensure the instrument is performing as intended with no subsequent changes planned. Note that Fall 2020 was not chosen first and the development process designed to meet this goal. Rather, the development activities necessary to achieve an instrument with appropriate validity evidence for our intended purpose were determined and a timeline constructed around these activities. However, data collection with possible revisions in Fall 2020 is included in the Figure 1 timeline in recognition that more changes – even minor tweaks – may be necessary, and a large administration of the S-SOMAS is planned for Fall 2020.

While not intrinsic to the work of developing instruments, professional realities still affect the development timeframe and should be reasonably accounted for in planning. Examples of professional realities that may contribute to long development timelines include high teaching loads, time commitments for other projects, supplemental work for the development such as writing proposals for grant applications, and the sequential development process introducing bottlenecks resulting from the division of labor based on team members' expertise.

Overview of Theoretical Framework

The SOMAS project team has adopted EVT as the theoretical framework for the S-SOMAS instrument. EVT is a theory of motivation, and this theoretical focus is reflected in the term “motivational attitudes” in the SOMAS project name. Historically, researchers in statistics education have been interested in studying what has been referred to as attitudes, even when nominally aligned with established educational theories. The SOMAS project team uses the phrase *motivational attitudes* as a bridge to connect the field of statistics education’s historical focus on attitudes to the broader educational literature of motivation that informs this project.

Before presenting the S-SOMAS EVT model, it is important to note that there are several related technical terms used throughout this chapter that are closely related: construct, scale, and factor. We conceive of *constructs* as a type of latent variable, that is, a random variable that cannot be measured directly (Raykov & Marcoulides, 2011). For each constructs, a group of items – known as a *scale* – will ultimately be developed to measure it. While the term scale is sometimes conflated with the term instrument, we use term scale to mean a group of items that assess a construct (Raykov & Marcoulides, 2011); the S-SOMAS instrument is then comprised of several scales. When analyzing items and scales, the term *factor* is used to denote an unobservable variable that explains the relationships among items (Raykov & Marcoulides, 2011). Ideally the unobservable variable represented by a factor would be a construct of interest, but this is not always the case.

The EVT framework explains achievement-related outcomes by relating them to an individual’s beliefs about success on a given a given task (expectancies) and beliefs about the value of the task (values) (Eccles & Wigfield, 2002). In EVT, the choice of task, performance on the task, and persistence on the task are affected by one’s expectancies and values: all other variables and constructs that may have an effect on achievement are mediated through the expectancies and values constructs (Eccles, 1983; Eccles & Wigfield, 2002; Wigfield & Cambria, 2010). While a detailed description of the EVT models used by the SOMAS project team is beyond the scope of this manuscript, a brief description of the rationale for selecting EVT is provided in the description of validity evidence section that follows.

The SOMAS project team began by developing EVT models that explain student performance based on the constructs and relationships articulated in the EVT literature (Eccles, 1983, 2014; Eccles & Wigfield, 2002). This led to the creation of the EVT model diagram for the S-SOMAS (Figure 2), which is similar to other published EVT models (e.g. Eccles, 2014). Each of the components of the model in Figure 2 (e.g. Utility Value, Goal Orientation) is a construct, a theoretical idea that we wish to measure (Wilson, 2005). The S-SOMAS model diagram, along with other internal documents describing the constructs, are heavily referenced by the SOMAS project team throughout the development process: it informs the item-writing process, the assembly of forms, analysis of data, and interpretation of results.

[INSERT FIGURE 2 HERE]

The model in Figure 2 was developed for use throughout the development process by illustrating all constructs hypothesized to influence performance in EVT, the relationships among these constructs, and which constructs are intended to be measured by the instrument. Further details about the EVT models are available in other manuscripts (e.g. Whitaker, Unfried, & Batakci, 2018).

Overview of Validity Evidence Plan

Validity, in the context of instrument development, refers to the overall support for proposed interpretations and uses of scores from an instrument (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014; Messick, 1995). Validity is not an intrinsic property of an instrument; instead, it should always be discussed in the context of specific interpretations (AERA et al., 2014; Messick, 1995). It is only with appropriate evidence that specific interpretations and uses of instruments are supported, and collecting validity evidence is a key aspect of the instrument development process.

The SOMAS team has adopted the validity framework articulated in the 2014 *Standards for Educational and Psychological Testing* document (AERA et al., 2014) which conceptualizes as evidence supporting validity of interpretations as being derived from several sources. This framework describes five broad sources of evidence for validity claims:

- Evidence based on test content
- Evidence based on response processes
- Evidence based on internal structure
- Evidence based on relations to other variables
- Evidence for validity and consequences of testing (AERA et al., 2014, pp. 14–19)

Evidence from each of these five sources supports intended interpretations and uses of an instrument (AERA et al., 2014).

Specific validity evidence to be collected aligned with each of the five sources supporting the use of the S-SOMAS instrument are presented below and summarized in Table 1. This table also lists the primary phase of the development process in which evidence will be collected. Note that the mapping of evidence statements to development phases is approximate and does not preclude evidence collection during other phases: for example, the claim “Chosen Likert-type response scale is appropriate” is listed as being a focus in the Initial Planning and Development phase, though evidence will also be collected in the Analysis of Collected Data and Item Writing and Revisions phases. Together, Table 1 and Figure 1 attempt to capture the complexity of the iterative instrument development process that includes the collection of many types of validity evidence both concurrently and sequentially. By focusing on validity evidence throughout the development process and modifying the instruments and supporting models as warranted by the development process, the resulting instruments will be well-supported for their intended uses by researchers with introductory statistics students.

[INSERT TABLE 1 HERE]

Evidence Supporting Validity

Each of piece of validity evidence supporting the S-SOMAS development outlined in Table 1 is described and categorized using the five sources of validity evidence (AERA et al., 2014). These pieces of validity evidence serve a roadmap for the development of the S-SOMAS instrument and may function as a model roadmap for future instrument development.

Validity Evidence Based on Test Content

The collection of validity evidence based on test content is central to the focus of developing an instrument to measure motivational attitudes toward statistics in a manner consistent with the established framework we have adopted. The evidence described below will be collected throughout the S-SOMAS development process, including the initial planning and development, item writing and revisions, and the assembly and revision of forms (see Table 1 and Figure 1). Clearly articulating these claims is integral to a plan that will result in an authentic assessment of attitudes toward statistics.

Claim: The EVT model is appropriate for use with undergraduate students.

EVT was chosen as the theoretical framework to guide the development process, and more details about the selection and components of the proposed EVT models are planned for other manuscripts. The EVT model was initially considered as the theoretical framework for the S-SOMAS in part because it is the stated theoretical framework for the SATS-36 instrument, but there were additional reasons why the EVT model was considered and finally adopted for use in this project. The statistics education literature currently uses the language of attitudes, but it seems that researchers are interested in motivation to explain performance and behaviors. This led to considering several motivational models.

Bandura’s (1977, 1986) self-efficacy model serves as a theoretical foundation for EVT and other models such as self-regulated learning. Ultimately these models do not conflict with each other. The use of Bandura’s self-efficacy model directly was considered, but ultimately EVT was preferred because it includes additional aspects of motivation beyond self-efficacy. Self-regulated learning (Zimmerman & Labuhn, 2012) was not chosen because it is a cyclical model that would be difficult to fully investigate using a survey in one or two administrations. Ultimately, EVT (Eccles, 1983, 2014; Eccles & Wigfield, 2002) was chosen because of the many factors that it conceptualizes as influencing motivation while being consistent with other respected, widely-used theories such as Bandura’s self-efficacy model.

Another consideration that led to the adoption of EVT was that a single underlying theoretical framework explaining both students’ motivations for learning statistics (using the S-SOMAS) and instructors’ motivations for teaching statistics (using the I-SOMAS) was desired. Initial EVT models were developed for use with students who were children or adolescents (Eccles, 1983; Eccles & Wigfield, 1995). The implicit claim that the EVT model was appropriate for use with the populations of interest – undergraduate students and instructors – is made explicit.

Evidence supporting the appropriateness of EVT beyond the populations was initially proposed comes primarily from the literature. First, there is widespread consensus within the statistics education literature that EVT is appropriate for use with undergraduate statistics students as evidenced by the widespread use of the SATS-36 instrument and development of attitudinal models consistent with EVT (Ramirez et al., 2012; Schau, 2003b; Schau et al., 2012). Two empirical studies have found at least partial support for the use of EVT in statistics education, and discrepancies might be attributable to the instrument used in the studies (Hood, Creed, & Neumann, 2012; Sorge & Schau, 2002). Second, this EVT model has been widely applied with adults in various contexts, for example, with Korean female college students (Bong, 2001) and unemployed Belgians (Vansteenkiste, Lens, Witte, & Feather, 2005). Lastly, though a detailed argument based on empirical research about the appropriateness of the model with adults was not found, when Eccles and her colleagues write about EVT without a context, the term *individual* is used rather than *adolescent* or *child* (e.g. Eccles, 2014; Eccles & Wigfield, 2002). Even though Eccles and her colleagues have tended to apply the EVT model with children and adolescents, the widespread use of EVT with adults in the literature supports our use of EVT with adult populations, including undergraduate students and instructors.

Additionally, throughout the instrument development process, the proposed EVT models will be reevaluated based on data collection and revised as-needed. The goal of these revisions is to enact a model that is consistent with EVT while being responsive to the context of learning and teaching statistics. However, these models will ultimately be empirically tested once instruments have been constructed which will further confirm – or disconfirm – the appropriateness of the EVT model with undergraduate students and instructors.

Claim: Degree to which items are aligned with EVT constructs.

The items that comprise the SOMAS instruments are claimed to align with the EVT constructs in the theoretical framework. This claim is distinct from the loading of items onto statistical factors as discussed in the Internal Structure section: this claim is that the items themselves use language and ideas that are consistent with the theoretical constructs as defined in EVT rather than other frameworks.

The evidence supporting this claim is based on the item-writing and revision process used throughout the project. EVT was chosen as a theoretical framework and specific models for students and instructors were proposed prior to writing items or considering existing items for the SOMAS instruments. In this way, the SOMAS development team became familiar first with EVT and then items were written rather than doing a post-hoc alignment. Item writers consisted of the SOMAS development team and several members of ROSA. To guide the item-writing process, a document with directions, working definitions, and examples was created after the EVT models were proposed. This document was circulated to the item writers so that there were common touchstones all could reference. The working definitions were brief (a few sentences at most) and grounded in the literature, and the examples were a few statements for each construct that the development team believed would help clarify the associated definition (see Figure 3). At this point, item writers worked individually.

[INSERT FIGURE 3 HERE]

After items had been written, the SOMAS development team met to evaluate the alignment with constructs to determine an initial item pool for the S-SOMAS. Each item was evaluated for its alignment with the construct it was written for and other characteristics (e.g. readability, not double-barreled, etc.). Many items were excluded or revised substantially, and then a round of collaborative item-writing took place. The development team met again to assess item alignment and develop the initial item pool from the proposed items. During this process, the SOMAS team struggled with aligning items in the pool for several constructs and, in doing so, refined its understanding of the EVT constructs and model.

Additionally, subject-matter experts (SMEs) have been – and will continue to be – involved in the review of items. After the initial pool of items was developed, a list was developed consisting of SMEs with expertise in statistics education, STEM education, student attitudes, the SATS instrument, and educational psychology. SMEs were identified by SOMAS team members using personal contacts and by searching the literature and conference programs for authors whose work was related one of the desired categories. These SMEs were presented with all items in the pool, organized by construct, and asked to rate how essential the item was for measuring the construct (Essential; Useful, but not Essential; Not Necessary). SMEs were also asked for their feedback about the overall construct, and many SMEs used this free-response box to discuss perceived item (mis)alignment. Both the qualitative and quantitative feedback was reviewed by the SOMAS project team and used to identify items to remove, recategorize, or revise. This differs from a Delphi study wherein participants are presented with the results from the group before providing their feedback again (Helmer-Hirschberg, 1967). The SOMAS project team rather than the SMEs received the initial feedback. A similar process will be used in other phases of the project.

Claim: Created scales cover salient aspects of the constructs.

When developing instruments that are aligned with EVT, an implicit claim is that the scales that comprise the instrument measure the salient aspects of the constructs to which they are aligned. While all constructs are conceptualized as unidimensional continua, it is still not guaranteed that any set of items that have been separately identified as aligning to the construct and that load onto the statistical factor together capture the richness of the construct (because important aspects may not be assessed by any of the items). For example, the Utility Value construct in the student EVT model is the value one places on statistics because learning statistics meets some future goal (Eccles & Wigfield, 2002; Flake, Barron, Hulleman, McCoach, & Welsh, 2015). It is conceivable that a scale could be developed that measures the Utility Value construct in a limited way that ignores import some aspects, for example, valuing statistics because learning statistics specifically meets the future goal of finding a job (when really there are many future goals one might have that cause one to value statistics).

Evidence supporting this claim will be collected from a review by SMEs with expertise in EVT. These experts will be asked whether they believe the scales measure the salient aspects of the constructs, if any important aspects have been missed, or if there are irrelevant aspects that have been included. SMEs will be identified using a process similar to the one described above. Additionally, interitem correlations will be calculated within scales to identify potential redundancies or scales that might be too homogenous and thus candidates for further review. However, this analysis is not sufficient for identifying when salient aspects have been missed.

Focus groups with various stakeholders are also planned to further support a robust view of each construct and have already been conducted with undergraduate students. Focus groups will also be held with instructors about the S-SOMAS instrument because they are an intended user of the instrument. These focus groups will be conducted early in the development process to inform item-writing and construct alignment rather than centered on near-final instruments.

Claim: Operationalized model is consistent with EVT model.

EVT is an adaptable framework: the full model has many constructs and relationships (Eccles, 2014; Eccles & Wigfield, 2002) while specific uses of EVT may make adaptations (e.g. Eccles, 2014), and a given instrument may not measure every construct. While EVT is adaptable, modifications and simplifications can threaten the extent to which an instrument is consistent with EVT. By using EVT as the theoretical framework for this family of instruments and developing models aligned with it, an explicit claim is made that the final, operationalized models are consistent with EVT.

As with assessing the scales for alignment with and coverage of EVT constructs described above, SMEs with expertise in EVT will be asked to provide feedback about the proposed EVT models and any simplifications, adaptations, or decisions about measurement that have been made to ensure that important aspects have not been sacrificed for some other end. This feedback will be solicited prior to finalizing instruments to allow time to respond to feedback and modify the models as-needed.

Claims Requiring Evidence Based on Response Processes.

One definition of response processes is "the mechanisms that underlie what people do, think, or feel when interacting with, and responding to, the item or task" (Hubley & Zumbo, 2017, p. 2). There are many ways of collecting validity evidence for response processes (e.g. Padilla & Benítez, 2014), which include but are not limited to focus groups, interviews, and analyzing other specialized data from participants such as response times or eye movements. Hubley and Zumbo (2017) note that response process validity evidence may either be descriptive of how participants respond or explore why people respond in the ways that they do more deeply, and that while reporting of all types of response process validity evidence is limited, most work has been descriptive. The claims presented below for the SOMAS project are descriptive in the sense offered by Hubley and Zumbo. Validity evidence supporting these claims about the ways in which respondents interact with the instrument will be collected during planning and development phases as well as when the assessments are in pilot and operational phases (see Table 1 and Figure 1). Because validation is never *complete*, these descriptive claims are for initial response process validity evidence, but future work may expand to further explore why participants answer in the ways that they do. Like many instruments used in education, the S-SOMAS is an online survey. However, this does not abrogate the need to collect validity evidence based on response processes and plan accordingly.

Claim: Constructs have different levels on a continuum.

Consistent with Coombs's formulation that endorsing a response on a Likert-type item is based on the relative location of an individual and an item on the underlying continuum for that construct (1964; Roberts, Laughlin, & Wedell, 1999), each construct in the EVT model is conceptualized as being a continuum, and each respondent will have some level of the construct on that continuum. The probability of endorsing a particular response in a Likert-type item is therefore based on an individual's location on an underlying continuum, and it is this hypothesized continuum that is the focus of this claim. For the S-SOMAS instrument, this continuum will be manifest in the ordering of respondents based on their responses to the items in the scale rather than an ordering of items within each scale along a continuum (Wilson, 2005). For example, for the utility value construct items will not be written that have an explicit order and mapping from low utility value to high utility value along a continuum. Instead, we expect to see students respond in such a way that the scale scores for students fall along a continuum from having low to high utility value for statistics. The primary way that validity evidence for this claim will be collected is the construction of an Wright Map (Wilson, 2005, 2011) during analysis of data collected from pilot and operational administrations of the S-SOMAS. A Wright Map, sometimes called an item-person map, is a figure illustrating the item difficulty continuum and person ability continuum on the same axis and scale (Wilson, 2011). It is hypothesized that the scales measuring EVT constructs will produce Wright maps that show respondents at various levels of the construct, thus illustrating some continuum for the construct that influences an individual's response process.

Claim: No differential item functioning will be present.

The SOMAS instruments will also be investigated for the possibility of differential item functioning (DIF), which occurs when items perform differently for respondents of the same level on the continuum of interest who belong to different groups. DIF is undesirable and reflects items performing differently across groups rather than underlying group differences (Wilson, 2005). Initially, DIF will only be investigated between groups of respondents who self-identify as female or male on the student instrument due to sample size requirements. However, as more data are collected DIF will be assessed for others identifiable groups (e.g. based on self-identified race). While the initial administrations of the instruments will be the United States, it is hoped that the instruments will be used broadly: as uses and data support such analyses, DIF will be investigated across global cultures.

Claim: Chosen Likert-type response scale is appropriate.

All items on the SOMAS instruments will use a 7-point Likert-type response scale (Strongly Disagree to Strongly Agree with a Neutral category). There are two claims related to the selection of the 7-point scale that must be justified: (1) the 7-point scale provides enough granularity to account for respondents across the hypothesized continuum for each construct and (2) the inclusion of a midpoint is appropriate. Evidence for both claims is drawn primarily from the literature rather than empirically-derived using data collected during this project.

The use of a midpoint in the Likert-type items is supported because a neutral response might be theoretically more appropriate for a respondent on any given item. Additionally, the choice to use or not use a neutral response category has not been shown to have a substantial effect on data or conclusions (Dillman, Smyth, & Christian, 2014). For Likert-type items, four to seven points is a widely-used guideline (Dillman et al., 2014). In the statistics education literature, 7-point scales are widely-used and selecting the same scale may aid in simplifying interpretations and comparisons of instruments.

Claim: Online surveys are appropriate for data collection.

An implicit claim made of the SOMAS instruments is that an online administration format is appropriate. The use of online survey tools is ubiquitous in contemporary social science research (Dillman et al., 2014; Hewson & Stewart, 2016), and the populations of interest – students and instructors of statistics – seem well-suited to web-based surveys. Concerns about the use of internet surveys with certain populations generally stems from potential added costs, lack of access, or low computer literacy (Dillman et al., 2014). Students and teachers of statistics have increasingly used computers for decades (suggesting at least basic computer skills for most students and their teachers), and the use of technology in statistics is a widely-accepted best-practice (GAISE College Report ASA Revision Committee, 2016).

Additional concerns about low response rates to internet surveys are not generally applicable to the use of the SOMAS instruments because the anticipated mode of administration is for students to be solicited for participation by their instructor, an individual known to them rather than a stranger. There is a potential for non-response rates for instructors to be high because the researchers may be strangers to them, but the statistics education community is friendly and targeted efforts will be made to recruit statistics instructors.

Potential trust issues related to the collection and storage of data will also be mitigated using a professional survey administration platform. Initially Qualtrics (a professional platform with many options for securely storing data and ensuring respondents can use a multitude of devices to respond) will be used for the S-SOMAS with a goal of later moving to a customized, professional website modeled on existing successful websites used for data collection in statistics education. We believe that this approach for recruiting participants and collecting data ameliorates many of the challenges associated with using internet surveys and that the benefits greatly outweigh the challenges.

Claims Requiring Evidence Based on Internal Structure

The validity evidence based on internal structure is closely related to properties of the instrument, scales, and items, and primarily a focus during the analysis of data collected from pilot and operational administrations of the instrument, though some validity evidence stems from the item writing and revision process (see Table 1 and Figure 1). While validity evidence based on internal structure is among the most common reported (Hubley & Zumbo, 2017), planning for the analyses required and understanding how they will affect revisions to instruments and models is key.

Claim: Items load onto hypothesized factors.

A particularly important claim that is made about the SOMAS assessments is that the assignment of items to scales is appropriate. That is, items presented as part of a scale designed to measure an underlying construct should measure that construct and not another. While this claim may seem to be easily met on its surface, early item-writing efforts based on the initial S-SOMAS EVT model revealed unforeseen similarities among items written for different constructs. For example, when reviewing items written to measure the Goal Orientation construct and the Attainment Value construct, the research team experienced difficulties determining for an individual item which construct it actually measured. An example of the similarity among items from different constructs can be seen in these provisional items, each from a different construct:

- If I am unable to interpret statistical results, I feel insecure. (Attainment Value)
- It is challenging to solve a problem that requires using statistics. (Perceived Difficulty)
- I can complete tasks that require basic statistical skills. (Expectancies)
- I lack the skills to do well in statistics. (Self-Concept of Statistics)

Each of the four items above was written for and aligns with the construct indicated in parentheses but viewed together there are considerable similarities. In response to this difficulty, the complete pool of items as of Fall 2017 was administered to students enrolled in introductory statistics courses in two forms composed of scales designed to measure constructs that were perceived as similar by the research team. All items were included on at least one form, neither form contained items from all constructs, and the form that an individual received was randomized. The purpose of this early pilot administration (Pilot 1 in Figure 1) is to provide feedback to the SOMAS team about the nature of the items and constructs. The data collected will be analyzed using exploratory factor analysis with a focus on item loadings: items with high loadings in only one construct are desired for this project. The analysis of data collected in this pilot administration will inform the assignment of items to scales, the item-writing process, and potential revisions to the EVT models. Throughout development factor loadings for items will be examined to determine if the item is appropriately categorized.

Claim: Each subscale is internally consistent.

Related to the above claim is the claim that the final scales on the SOMAS instruments will have reasonable internal consistency, that is, “the degree to which the items on a test jointly measure the same construct” (Henson, 2001, p. 177). However, while internal consistency is an important aspect of reliability, particularly high values for measures of internal consistency for each scale are not of paramount importance for this project. This is because measures of internal consistency tend to be higher for longer scales, and the number of constructs to be measured would result in prohibitively long S-SOMAS and I-SOMAS instruments that would be unlikely to see widespread use. Depending on the specific development of scales, longer scales measuring a construct may be developed and then pared down to produce shorter scales that still have acceptable internal consistency.

The omega coefficient (Raykov & Marcoulides, 2011) will be used as the primary measure of internal consistency of scales on SOMAS instrumentsto be consistent with the confirmatory factor analysis approach to be used when finalizing instruments. Other measures of internal consistency – such as coefficient alpha (Cronbach, 1951) – may be calculated and used throughout the project for communicating with other audiences but only with an understanding of their uses and limitations (e.g. Henson, 2001; Sijtsma, 2009). Because of coefficient alpha's ubiquity, a few limitations will be briefly discussed.

Sijtsma (2009) provides a thorough description of coefficient alpha and details several fundamental problems with its typical use. First, Sijtsma notes that alpha is grounded in the paradigm of classical test theory, and so its ad-hoc use with other paradigms such as item response theory is not advisable. Second, while alpha is correlated with other statistical measures, it is neither a measure of internal consistency nor a measure of unidimensionality and conveys little information on its own (Sijtsma, 2009). Third, alpha's use as a lower-bound estimate for an instrument's reliability (its correlation between the scores on an instrument an on a parallel version) based on a single administration is often inappropriate because better lower bounds have been proposed (Sijtsma, 2009). Lastly, a focus on reporting alpha may serve to oversimplify and conflate the distinct concepts of reliability and internal consistency.

Claim: Measurement invariance for each construct.

Another claim that evidence will be collected is that the measurement is invariant for each construct, that is, the distances between respondents and the distances between responses on a Wright map are interpretable regardless of location on the continuum (Wilson, 2003, 2005). Evidence related to this claim will be gathered during data analysis: measurement invariance equates to each item having the same slope in the measurement model that is adopted. Because of the 7-point Likert-type scale that is being used with the S-SOMAS instrument, a polytomous model such as the Generalized Partial Credit Model (Muraki, 1992) or Graded Response Model (Samejima, 1969) will be employed during the analysis of data. Models for analyzing polytomous responses differ in the parameters that are estimated. While a comparison of each model is beyond the scope of this paper, one way in which these models differ is whether the model for all items has the same slope (discrimination) or if each item has a slope parameter calculated. Together with the researchers' judgement of the appropriateness of the models, statistical techniques will be used to compare models to determine if equality of slope parameters for each item is reasonable.

If a model that allows each item to have a different slope parameter fits best, then two options may be considered. First, the items that measure the construct under consideration will be revisited, and items may be added or deleted to obtain a measure of the construct that is invariant (Wilson, 2005). In this case, the items with slope parameters most dissimilar from the others would be examined first to determine if they are candidates for removal; the difference in item slope parameters may be a determining factor in the deletion of an item in addition to other reasons. If this is not successful, then the interpretations of the construct under consideration may be revised to account for this invariances through a more complex interpretation (Wilson, 2005). The first option is preferable for this project because the final family of instruments are intended to be made widely-available to researchers, instructors, and other interested parties, and parsimonious construct interpretations may be more desirable than complex ones. However, proposed interpretations of constructs may be made more complex if needed (and supplemented with appropriate training materials).

Claim: Constructs are unidimensional.

We plan to use item response theory (IRT) to analyze responses, and IRT assumes that underlying constructs measured by each scale are unidimensional. This is consistent with the EVT framework guiding the development. There are two related claims about unidimensionality for this project: first, that the EVT constructs to be measured by the SOMAS instruments are unidimensional; and second, that a unidimensional structure is appropriate for each scale created to measure an EVT construct.

Each EVT construct proposed by Eccles and colleagues (e.g. Eccles & Wigfield, 2002) is not explicitly unidimensional, though many of the constructs are expected to be unidimensional when operationalized. For example, the EVT construct Utility Value, in the context of learning statistics, is expected to be a unidimensional construct: the SATS-36 Value construct is most closely aligned with Utility Value (Whitaker & Gorney, 2017) and has been shown to be a distinct unidimensional factor based on confirmatory factor analysis (Vanhoof et al., 2011).

It is likely, though, that some proposed EVT constructs will not be facially unidimensional. During an early item-writing activity, the Goal Orientation construct was identified as being potentially multidimensional. Initially, a unidimensional continuum of intrinsic orientation to extrinsic orientation was assumed. However, difficulties in writing and reviewing items led to the decision to view Goal Orientation as two distinct but related unidimensional constructs: Intrinsic Goal Orientation and Extrinsic Goal Orientation, each on a low to high continuum. When data suggest that a construct may not be unidimensional and a theoretical explanation can be supported, broader constructs may be reconceptualized as narrower unidimensional constructs. The Intrinsic and Extrinsic Goal Orientation constructs reflect different orientations that have been proposed in the literature such as mastery and performance orientation, respectively (Wigfield & Cambria, 2010). The proposed EVT models will then be refined.

Moreover, it is possible that some proposed EVT constructs might not be practically distinguishable from each other and therefore best viewed not as distinct unidimensional constructs but rather as a single unidimensional construct. Eccles and Wigfield (2002) define expectancies for success by synthesizing the extant work of Eccles and her colleagues: “individuals’ beliefs about how well they will do on upcoming tasks, either in the immediate or longer term future” (p. 119). While Eccles and Wigfield include two different theoretical types of beliefs in their definition – beliefs about the immediate future and beliefs about the longer-term future – they note that these “are highly related and empirically indistinguishable” (p. 119). When data suggest that theoretically-distinct constructs may not be empirically-distinguishable and a theoretical explanation can be supported, narrower constructs may be collapsed to form broader unidimensional constructs. The proposed EVT models will then be refined.

Evidence about unidimensionality will be gathered throughout the project. First, as items are written for specific constructs and preliminary data collected, exploratory factor analysis will be used to guide and refine the assignment of items to constructs. Additionally, items will be rewritten, omitted, or replaced as-needed to support the creation of scales that perform in ways consistent with measuring unidimensional constructs. The creation of scales that measure unidimensional constructs in the EVT model will be an iterative process where the theoretical framework guides the creation of scales and data collected from the use of the scales informs revisions to the theoretical framework. Ultimately, prior to supporting the widespread use of the SOMAS instruments by other researchers, model fit in a confirmatory factor analysis setting will be assessed to support the unidimensionality of the constructs measured by the revised instruments aligned with revised EVT frameworks.

Claim: Items are locally independent.

Another assumption needed for IRT is local independence, that is, responses to items are independent given a respondent's individual trait characteristics. To assess local independence, pairwise inter-item correlations will be computed and examined. For at least one pilot administration, items will be administered either grouped by constructs and in a randomized order before a final choice is made.

Claims Requiring Evidence Based on Relations to Other Variables

As part of the initial validity study, other instruments will be administered simultaneously with the S-SOMAS instrument. There is a strong presence in the planning phases to determine which instruments to administer and to whom. These instruments will not all be administered to all participants. Instead, subsets of respondents will be selected to take the S-SOMAS instrument and another instrument to determine if hypothesized relationships are observed. Relationships between the S-SOMAS instrument and two other instruments will be the primary focus: the SATS-36 (Schau, 2003b) and the Levels of Conceptual Understanding in Statistics Intermediate/Advanced online form (LOCUS; Jacobbe, Case, Whitaker, & Foti, 2014; Whitaker, Foti, & Jacobbe, 2015). These hypotheses, which are grounded in the statistics education literature, motivate the data collection plan and gathering of validity evidence in a deliberate way rather than being incidental components of later research.

Because the SATS-36 and S-SOMAS instrument both aim to measure students' attitudes toward statistics in a manner consistent with EVT, positive correlations between comparable constructs are expected. In particular, we would expect to see moderate positive correlations for the following factors:

- SATS-36 Affect and S-SOMAS Subjective Task Value subconstructs
- SATS-36 Cognitive Competence and S-SOMAS Self-Concept of Statistics Ability
- SATS-36 Value and S-SOMAS Utility Value
- SATS-36 Difficulty and S-SOMAS Perception of Difficulty
- SATS-36 Interest and S-SOMAS Interest/Enjoyment

These hypothesized relationships are based on the existing statistics education literature using the SATS instruments (e.g. Schau, 2003a; Schau & Emmioğlu, 2012; Vanhoof et al., 2011). Additionally, we do not expect to see a moderate correlation between SATS-36 Effort and S-SOMAS Cost because SATS-36 Effort construct has previously been noted to produce responses that might be too high to reflect reality (Schau & Emmioğlu, 2012). Other weak to moderate correlations among the constructs are expected to be observed, but these predictions are articulated because the scales on each instrument are attempts to measure similar underlying constructs.

Evidence for Validity and Consequences of Testing

Validity evidence will also be collected to support specific uses of the S-SOMAS. The developers intend that these instruments will be appropriate for longitudinal and pre/post research designs as well as snapshots of motivational attitudes with students and instructors at higher education institutions within the United States. To support this broad claim, several claims have been articulated above and these specific research purposes precipitated the data collection plan for the S-SOMAS. Initial validation work will use the S-SOMAS instruments in a pre/post design as that has been popular in statistics education and was the intended use of the SATS-36, but S-SOMAS is being designed so that researchers are not restricted to only using it in a pre/post setting but instead are appropriate for use longitudinally to support more sophisticated research designs such as latent growth models and time series (e.g. Sloane & Wilkins, 2017). Additionally, data collection will include a diverse set of United States higher education institutions such as large public research universities, small liberal arts schools, primarily undergraduate institutions, and community colleges.

While the SOMAS development team does not anticipate uses of the instruments beyond research in statistics education, to clarify the claims about intended uses an explicit anti-endorsement of use is offered: the SOMAS instruments should not be used in high-stakes settings such as hiring or firing instructors or placing students into courses. There are likely other uses for SOMAS instruments that might be appropriate that have not been articulated by the development team, but these other uses would require explicit validity evidence to be collected to support them and cannot rely only on the validity evidence collected by the development team.

Use of the Validity Plan in Guiding Development

This chapter aimed to illustrate the development process for an instrument to be used for research purposes with a focus on validation work with explicit attention to several aspects that may be hidden from or unknown to those new to instrument development: the development process is not linear – revisions occur in many areas, including in the documents guiding the development as more is learned – and intended uses are supported by a variety of claims for which evidence must be collected. Each preceding claim, along with its potential evidence, was written intentionally to support the the SOMAS project's goals and the intended use of the S-SOMAS with undergraduate students enrolled in statistics courses. While some claims may be made in many validation studies (e.g. "Items are locally independent" is a requirement of Item Response Theory), each claim was made to support specific uses of the S-SOMAS for specific purposes. The ultimate goal of the SOMAS project is not to develop an instrument: we want to deepen the field's knowledge what affects student outcomes in statistics and to what degree. To that end, a new family of instruments is needed as there is a paucity of validity evidence for existing instruments supporting our desired uses. Our intended uses of the S-SOMAS instrument motivate the development of the instrument and the validity evidence to be collected.

As illustrated in Figure 1, the development of the S-SOMAS instrument is underway but with much work remaining. As of November 2018, the EVT framework has been developed and revised multiple times, a pool of items has been written and revised, SMEs have been consulted, focus groups have been conducted, and data have been collected in a pilot administration to inform revisions. While this is still relatively early in the overall conception of the larger SOMAS project, we have already needed many of the validity claims listed above for guiding and focusing our discussions and research activities. For example, prior to the starting any item writing, some evidence for the following validity claims was needed:

- The EVT model is appropriate for use with students,
- Chosen Likert-type response scale is appropriate, and
- Online surveys are appropriate for data collection.

Additionally, the validity statements concerning appropriate use of the final instruments were also critical to have articulated before item writing began. The validity statements serve to focus the work done by the SOMAS project team and to help coordinate the many research tasks that will be undertaken in the coming years.

Conclusion and Discussion

Developing high-quality instruments for which there is documentable evidence supporting their intended uses requires substantial planning and commitment. The SOMAS project is an on-going instrument development project, and specific details and goals described in this chapter evolve or change as the project continues – but will be guided by the same focus on explicit evidence supporting intended uses that was demonstrated here. A timeline of the development process is illustrated in Figure 1; this timeline shows the sequencing of key development activities and highlights the iterative nature of the instrument development process: the revisions to the outcomes of previous phases of development are explicitly accounted for based on the collection and analysis of additional data. The process of developing a high-quality instrument that is suitable for its intended purpose takes substantial time: developing an instrument is not just writing a survey or test and it cannot be completed overnight or even in a few weeks. While many factors affect the amount of time needed to develop an instrument, planning on a timeframe that is measured in *months* rather than *days* is advisable.

The development of the S-SOMAS instrument exemplifies what this planning can look like with a focus on the collection of validity evidence consistent with the five sources of validity evidence in the *Standards for Educational and Psychological Testing* (AERA et al., 2014). The validity evidence described in this chapter was summarized in Table 1. Together, Table 1 and Figure 1 attempt to capture the complexity of a multi-year instrument development project. By making explicit validity evidence that will support the intended uses of the S-SOMAS instrument early in the development process, these evidence statements serve to guide development toward the intended uses rather than resulting in a piecemeal approach to validity undertaken after development has concluded.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, N.J.: Prentice Hall.

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

Bong, M. (2001). Role of Self-Efficacy and Task-Value in Predicting College Students' Course Performance and Future Enrollment Intentions. *Contemporary Educational Psychology*, *26*(4), 553–570.

Coombs, C.H. (1964). *A theory of data*. Oxford, England: Wiley.

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

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.

Dillman, D.A., Smyth, J.D., & Christian, L.M. (2014). *Internet, phone, mail, and mixed-mode surveys: the tailored design method* (4th edition). Hoboken: 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). San Francisco: W.H. Freeman.

Eccles, J.S. (2014). Expectancy-Value Theory. In R. Eklund & G. Tenenbaum (Eds.), *Encyclopedia of Sport and Exercise Psychology*. Thousand Oaks, CA: SAGE Publications, Inc.

Eccles, J.S., & Wigfield, A. (1995). In the mind of the achiever: The structure of adolescents' academic achievement related-beliefs and self-perceptions. *Personality and Social Psychology Bulletin*, *21*(3), 215–225.

Eccles, J.S., & Wigfield, A. (2002). Motivational Beliefs, Values, and Goals. *Annual Review of Psychology*, *53*, 109–132.

Flake, J.K., Barron, K.E., Hulleman, C., McCoach, B.D., & Welsh, M.E. (2015). Measuring cost: The forgotten component of expectancy-value theory. *Contemporary Educational Psychology*, *41*, 232–244.

GAISE College Report ASA Revision Committee. (2016). Guidelines for Assessment and Instruction in Statistics Education College Report 2016. Retrieved from https://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf

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). Amsterdam: IOS Press.

Helmer-Hirschberg, O. (1967). *Analysis of the Future: The Delphi Method*. Santa Monica, CA: RAND Corporation.

Henson, R.K. (2001). Understanding Internal Consistency Reliability Estimates: A Conceptual Primer on Coefficient Alpha. *Measurement and Evaluation in Counseling and Development*, *34*(3), 177–189.

Hewson, C., & Stewart, D.W. (2016). Internet Research Methods. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, & J.L. Teugels (Eds.), *Wiley StatsRef: Statistics Reference Online* (pp. 1–6). Chichester, UK: John Wiley & Sons, Ltd.

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.

Hubley, A.M., & Zumbo, B.D. (2017). Response Processes in the Context of Validity: Setting the Stage. In B.D. Zumbo & A.M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 1–12). Cham: Springer.

Jacobbe, T., Case, C., Whitaker, D., & Foti, S. (2014). Establishing the validity of the LOCUS assessments through an evidenced-centered design approach. In K. Makar & R. Gould (Eds.), *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute.

Jacobbe, T., Whitaker, D., Case, C., & Foti, S. (2014). The LOCUS assessment at the college level: conceptual understanding in introductory statistics. In K. Makar & R. Gould (Eds.), *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute.

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.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, *16*(2), 159–176.

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.

Padilla, J.L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, (26.1), 136–144.

Ramirez, C., Schau, C., & Emmioğlu, E. (2012). The Importance of Attitudes in Statistics Education. *Statistics Education Research Journal*, *11*(2), 57–71.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York: Routledge.

Roberts, J.S., Laughlin, J.E., & Wedell, D.H. (1999). Validity Issues in the Likert and Thurstone Approaches to Attitude Measurement. *Educational and Psychological Measurement*, *59*(2), 211–233.

Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores*. Richmond, VA: Psychometric Society.

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

Schau, C. (2003a). Students' attitudes: The "other" important outcome in statistics education (pp. 3673–3683). Presented at the Joint Statistics Meetings, San Francisco, CA.

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

Schau, C., & Emmioğlu, E. (2012). Do introductory statistics courses in the United States improve students' attitudes? *Statistics Education Research Journal*, *11*(2), 86–94.

Schau, C., Millar, M., & Petocz, P. (2012). Research on attitudes towards statistics. *Statistics Education Research Journal*, *11*(2), 2–5.

Schau, C., Stevens, J., Dauphinee, T. L., & Vecchio, A. D. (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

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107.

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). Reston, VA: 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, LA.

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.

Vansteenkiste, V., Lens, W., Witte, H., & Feather, N.T. (2005). Understanding unemployed people's job search behaviour, unemployment experience and well-being: A comparison of expectancy-value theory and self-determination theory. *British Journal of Social Psychology*, *44*(2), 269–287.

Whitaker, D., Foti, S., & Jacobbe, T. (2015). The Levels of Conceptual Understanding in Statistics Project: Results of the Pilot Study. *Numeracy*, *8*(2), Article 4.

Whitaker, D., & Gorney, K. (2017). *Surveys of Attitudes About Statistics: An Analysis of Items*. Poster presented at the 39th Annual Conference of the North American Chapter of the International Group for the Psychology of Mathematics Education, 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.), *Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018)*. Voorburg, The Netherlands: International Statistical Institute.

Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review*, *30*(1), 1–35.

Wilson, M. (2003). On choosing a model for measuring. *Methods of Psychological Research Online*, *8*(3), 1–22.

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. Mahwah, N.J: Lawrence Erlbaum Associates.

Wilson, M. (2011). Some Notes on the Term: "Wright Map." *Rasch Measurement: Transactions of the Rasch Measurement SIG American Educational Research Association*, *25*(3), 1331.

Zimmerman, B.J., & Labuhn, A.S. (2012). Self-regulation of learning: Process approaches to personal development. In K.R. Harris, S. Graham, T. Urdan, & J. Brophy (Eds.), *APA*

Educational Psychology Handbook (pp. 399–426). Washington, D.C.: American Psychological Association.

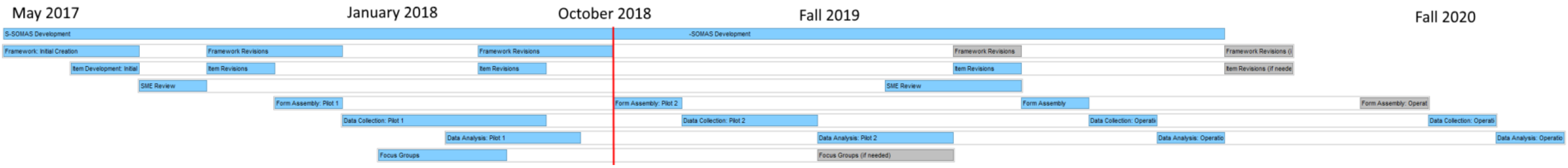


Figure 1. Timeline of S-SOMAS development.

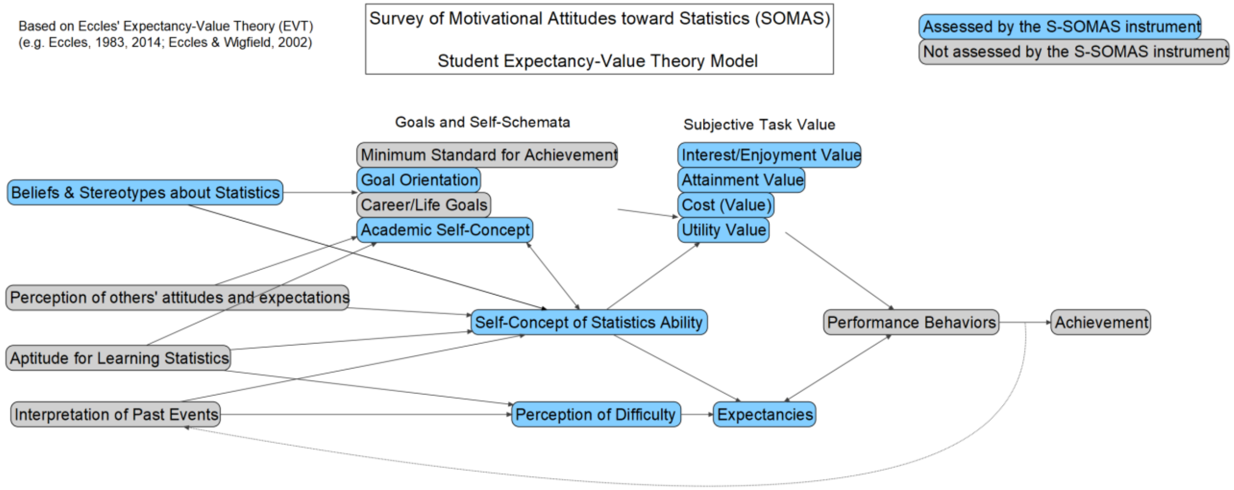
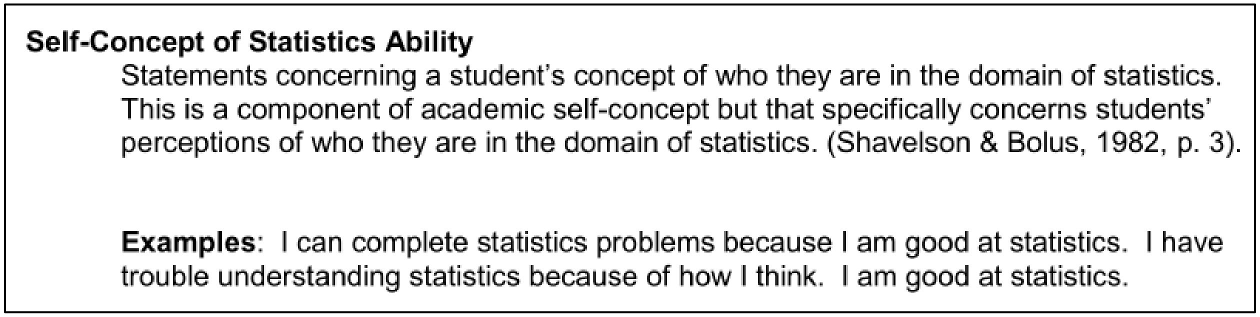

Based on Eccles' Expectancy-Value Theory (EVT)
(e.g. Eccles, 1983, 2014; Eccles & Wigfield, 2002)
Survey of Motivational Attitudes toward Statistics (SOMAS)
Student Expectancy-Value Theory Model
Assessed by the S-SOMAS instrument
Not assessed by the S-SOMAS instrument
Goals and Self-Schemata
Subjective Task Value
Minimum Standard for Achievement
Goal Orientation
Career/Life Goals
Academic Self-Concept
Interest/Enjoyment Value
Attainment Value
Cost (Value)
Utility Value
Beliefs & Stereotypes about Statistics
Perception of others' attitudes and expectations
Aptitude for Learning Statistics
Interpretation of Past Events
Self-Concept of Statistics Ability
Perception of Difficulty
Expectancies
Performance Behaviors
Achievement

Figure 2. The Student Model based on Expectancy-Value Theory.

Source of Validity	Validity Evidence Claims	Primary Development Phase
Test Content	The EVT model is appropriate for use with undergraduate students	Initial Planning and Development
	Items are aligned with EVT constructs	Item Writing and Revisions
	Created scales cover salient aspects of the constructs	Assembly of forms and revisions
	Operationalized model is consistent with EVT model	Assembly of forms and revisions
Response Processes	Constructs have different levels on a continuum	Analysis of Collected Data
	No differential item functioning will be present	Analysis of Collected Data
	Chosen Likert-type response scale is appropriate	Initial Planning and Development
	Online surveys are appropriate for data collection	Initial Planning and Development
Internal Structure	Items load onto hypothesized factors	Item Writing and Revisions
	Each scale is internally consistent	Analysis of Collected Data
	Measurement invariance for each construct	Analysis of Collected Data
	Constructs are unidimensional	Analysis of Collected Data
	Items are locally independent	Analysis of Collected Data
Relations to Other Variables	Moderate positive correlations expected for the following factors	Analysis of Collected Data
Uses and Consequences	Appropriate for specific uses, inappropriate for others	Intentional Design and Data Collection Throughout

Table 1. A list claims supporting validity of interpretations for the S-SOMAS instrument for which evidence will be gathered as well as the primary development phase in which each statement will be the focus.



Self-Concept of Statistics Ability

Statements concerning a student's concept of who they are in the domain of statistics. This is a component of academic self-concept but that specifically concerns students' perceptions of who they are in the domain of statistics. (Shavelson & Bolus, 1982, p. 3).

Examples: I can complete statistics problems because I am good at statistics. I have trouble understanding statistics because of how I think. I am good at statistics.

Figure 3. An example of the guidelines given to item writers.