JOURNAL OF RESEARCH IN SCIENCE, MATHEMATICS AND TECHNOLOGY EDUCATION
DOI: 10.31756/jrsmte.511

Measuring Opportunity Cost in Statistics Using Evaluative Space Grid Items: Results from a Pilot Study

Douglas Whitaker

Mount Saint Vincent University, Canada

Joseph Barss

Mount Saint Vincent University, Canada

Bailey Drew

Dalhousie University, Canada

Abstract: Challenges to measuring students' attitudes toward statistics remain despite decades of focused research. Measuring the expectancy-value theory (EVT) Cost construct has been especially challenging owing in part to the historical lack of research about it. To measure the EVT Cost construct better, this study asked university students to respond to items using both a Likert-type response and an Evaluative Space Grid (ESG)-type response. ESG items enable bivariate responses in a single item and permit distinguishing among two different types of neutral attitudes: indifferent and ambivalent. This pilot study evaluates the appropriateness of ESG-type items for measuring the EVT Cost construct by analyzing student response patterns to ESG-type items and comparing them with Likert-type items. Validity evidence is documented using descriptive statistics and graphs, correlations among items, and a trinomial hypothesis test. Internal consistency reliability indices are also reported. Friedman's Test is used to compare the average response times for items of different types. Results indicate that students can meaningfully respond to ESG-type items in ways that are similar to their Likert-type responses, that students respond to ESG-type items quicker with more practice, and that distinguishing among indifferent and ambivalent attitudes seems appropriate for the EVT Cost construct. These findings suggest that ESG-type items may provide new insights not possible with Likert-type items but also that more research should be conducted to better understand their advantages and disadvantages within statistics education.

Keywords: *Evaluative Space Grid; Expectancy-value Theory; Statistics Attitudes;* S*tatistics Education; Validity Evidence*

Introduction

Statistics educators have long been interested in affective constructs, and instruments measuring constructs such as *attitudes* (e.g., Roberts & Bilderback, 1980; Wise, 1985) and *anxiety* (e.g., Cruise, Cash, & Bolton, 1985) have been available for decades. Calls for an increased focus on research into these affective constructs (e.g., Gal & Ginsburg, 1994; Pearl et al., 2012) have resulted in a bevy of instruments measuring a multitude of interrelated constructs (Nolan, Beran, & Hecker, 2012). The two most widely used instruments to measure affective constructs with students in statistics education are the Survey of Attitudes Toward Statistics (SATS; Schau, 1992, 2003) and the Statistical Anxiety Rating Scale (STARS; Cruise et al., 1985). However, measuring the intended constructs can be problematic for both the SATS (Whitaker, Unfried, & Bond, in press) and the STARS (e.g., Chew, Dillon, & Swinbourne, 2018).

To address the challenges to using the SATS to measure students' attitudes about statistics, a new

instrument is being developed that is aligned to the same theoretical framework. Pilot data collection using this new instrument, the Student Survey of Motivational Attitudes toward Statistics (S-SOMAS), is ongoing (Unfried, Kerby, & Coffin, 2018; Unfried et al., 2021; Whitaker, 2021; Whitaker, Unfried, & Bond, 2019). Both the SATS and the S-SOMAS are aligned to an expectancy-value theory (Eccles & Wigfield, 2002, 2020) and both employ scales composed of Likert-type items. During the development of the S-SOMAS, the team encountered challenges with writing items for the Cost construct in the expectancy-value theory; the SATS scale aligned with Cost also has poor psychometric properties (Whitaker et al., in press). Because of this, the team held discussions about whether alternatives to Likert-type items might allow for better measurement of the Cost construct.

While instruments measuring affective constructs in statistics education have largely employed Likert-type items, there are many alternatives to Likert-type items (e.g., DeVellis, 2017). One such alternative item type is the Evaluative Space Grid (ESG) which asks participants to respond by selecting one cell from a grid (Larsen, Norris, McGraw, Hawkley, & Cacioppo, 2009). Due to characteristics of the expectancy-value theory Cost construct, ESG-type items may provide practical measurement advantages. This paper reports on a pilot study that used Likert-type and ESG-type items to measure the Cost construct with students in an introductory statistics course.

Literature Review

Two different bodies of literature support this study: Eccles and colleagues' Expectancy-Value Theory (EVT; Eccles & Wigfield, 2020), which describes the construct to be measured, and work on ESG-type items.

Situated Expectancy-Value Theory

EVT is a psychological theory of motivation explaining achievement-related choices and behaviours. Originally developed to explain mathematics achievement among adolescents (Eccles (Parsons) et al., 1983), EVT is a widely used framework for explaining motivation across many disciplines (Wigfield & Eccles, 2020) including statistics education (e.g., Ramirez, Schau, & Emmioğlu, 2012). In EVT, one's achievement-related choices and behaviours are directly affected by one's values (Subjective Task Values) and what one expects to happen (Expectancy); all other factors affecting achievement-related choices and behaviours are mediated through Subjective Task Values, Expectancies, or both (Eccles & Wigfield, 2002). The overall EVT model is broad and accounts for "distal psychological, social, situational, and cultural determinants" (Wigfield & Eccles, 2020, p. 164), and a thorough elaboration of it is beyond the scope of this paper. Rather, we will focus on the Cost construct.

Cost, sometimes referred to as the Cost of Success or Failure, was originally described as affecting how one values an activity (Eccles (Parsons) et al., 1983) and was viewed as a component of Subjective Task Values (e.g., Wigfield & Eccles, 2000). Specifically, the cost-benefit ratio of engaging in an activity was hypothesized to be related to the value assigned to the task through a reciprocal relationship (Eccles (Parsons) et al., 1983). EVT draws on social exchange theory which defines cost as "any factors that operate to inhibit or deter the performance of a sequence of behavior" (Thibaut & Kelley, 1959, p. 12). More

recent descriptions of EVT have moved away from viewing Cost as a component of Subjective Task Values and instead view Cost as affecting them.

While Cost was part of the original description of EVT and is viewed as "especially important" to the choices made by students (Wigfield, Rosenzweig, & Eccles, 2017, p. 124), it has also described as a "forgotten component of expectancy-value theory" (Flake, Barron, Hulleman, McCoach, & Welsh, 2015, p. 232) due to the historical paucity of research about it. The original conception of Cost included three components – effort, loss of valued alternatives, and the psychological cost of failure (Eccles (Parsons) et al., 1983) – but the construct is now understood to include many other costs such as emotional, social, and financial costs (Wigfield et al., 2017). There has been a recent increased focus on measuring Cost (e.g., Flake, 2012; Flake et al., 2015; Jiang, 2015; Jiang, Rosenzweig, & Gaspard, 2018) which has led to an expansion of the construct and yielded further opportunities for research (Wigfield et al., 2017). This recent work on measuring Cost has provided evidence that it can be empirically distinguished from Subjective Task Values (Jiang et al., 2018) and provided insights into how items measuring cost might be written (Flake et al., 2015).

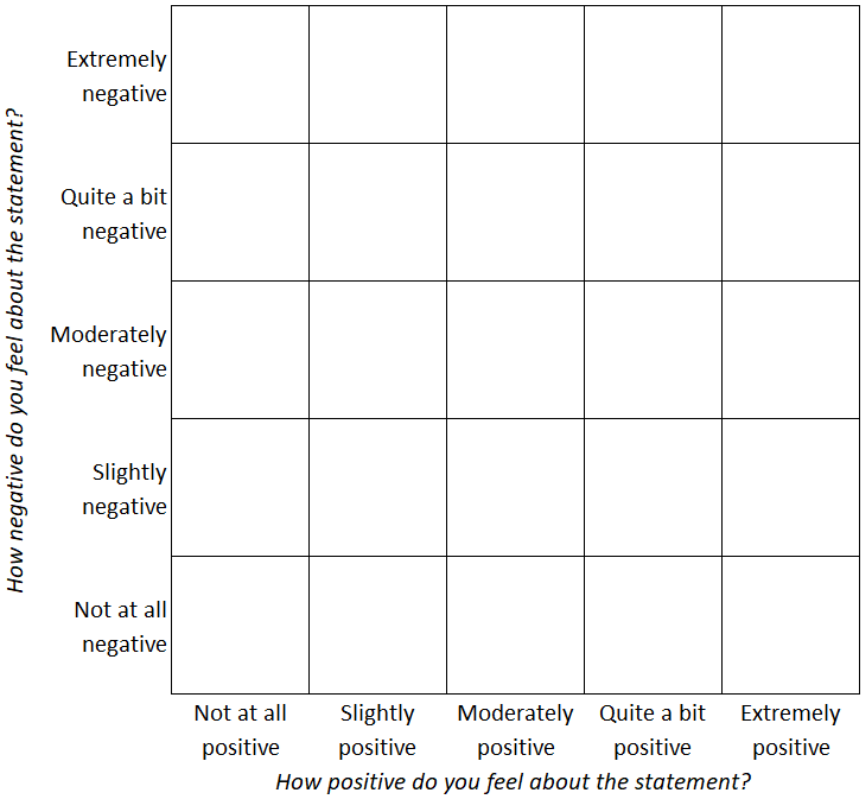
Evaluative Space Grid

Since their introduction in the 1930s (Likert, 1932/1933), items with bipolar response scales (e.g., Likert-type items) have come to dominate the field of attitude research (Bandalos, 2018; Irwing & Hughes, 2018). While the term Likert-type item might reasonably be applied to both items with both unipolar and bipolar labels (e.g., Uebersax, 2006), bipolar Likert-type items are standard (Bandalos, 2018; Likert, 1932/1933). Bipolar labels structured around Agreement/Disagreement are perhaps the most widely used form of Likert-type items, though other labels such as Positivity/Negativity or Approval/Disapproval are also common (Bandalos, 2018). Bipolar Likert-type items allow the respondent to indicate both the direction of their attitude (i.e., valence) and the magnitude of their attitude. Because a Likert-type item only allows a single response, the format of the item imposes a reciprocal structure on the construct (Cacioppo & Berntson, 1994): increases in agreement (or positivity) are necessarily matched by decreases in disagreement (or negativity).

While positive and negative attitudes are ostensibly reciprocally related for many constructs, the choice of measurement techniques in attitude research (e.g., Likert-type items) has imposed this structure on the constructs being measured (Cacioppo & Berntson, 1994). Such a simplification about the processes controlling the valences of affective constructs reflected the dominant research narratives in the early 20th century (Cacioppo, Gardner, & Berntson, 1997) and may have allowed for productive research about affective constructs at the time. However, empirical findings have shown "that positive and negative affect are not invariably reciprocally activated" (Cacioppo, Berntson, Norris, & Gollan, 2012, p. 55); this suggests that the time is now to reconsider the appropriateness of the simplified relationship imposed by the choice of Likert-type items. Ultimately, research about the extent to which affective constructs exhibit a reciprocal relationship is still nascent, and various measurement techniques have been proposed for measuring what may be a bivariate outcome including multi-item scales and ESG-type items. A generic ESG-type item is shown in Figure 1. The extent to

Figure 1

A typical formulation of the Evaluative Space Grid-type item as described by Larsen et al. (2009). Respondents are asked to select a single cell that best matches their attitudes.



which particular affective constructs do or do not exhibit a reciprocal relationship controlling positive attitudes and negative attitudes is also an open question.

The proposed advantage of ESG-type items over Likert-type items and other unidimensional assessments of attitudes is a more nuanced view of neutral responses (Larsen et al., 2009). With a Likert-type item, responses toward the middle of the continuum are associated with a neutral response – neither positive nor negative but without an indication of the intensity of the feeling. However, these neutral responses may be chosen by respondents for a variety of reasons, resulting in challenges to interpreting such responses. By using a bivariate response grid, ESG-type items allow for a distinction between two types of neutral attitudes: *indifferent attitudes*, characterized by low positivity and low negativity, and *ambivalent attitudes*, characterized by high positivity and high negativity (Larsen et al., 2009). Ambivalent attitudes might also be conceptualized as contradictory attitudes.

Research about bivariate ESG-type items has been conducted in the field of consumer satisfaction (e.g., Audrezet, 2014; Audrezet, Olsen, & Tudoran, 2016; Audrezet & Parguel, 2018; Borriello, 2017). These studies provide evidence that different latent response processes are used by respondents when choosing their

positive and negative responses using ESG-type items (Borriello, 2017). Validity evidence has been presented supporting the claim that ESG-type items can measure the intended construct (Audrezet et al., 2016; Larsen et al., 2009) and compare favourably with Likert-type items in their internal consistency (Audrezet et al., 2016). Additionally, researchers have advanced analysis methods appropriate for ESG-type items to enable direct comparison of ESG-type items with Likert-type items via a transformation (Audrezet et al., 2016) and modeling of attitudinal scores (Borriello, 2017).

However, there is not consistency in the literature about how the axes of the ESG-type items should be labeled with different researchers adopting slightly different approaches (e.g., Audrezet et al., 2016; Borriello, 2017; Larsen et al., 2009), though each has placed the positive terms (e.g., positivity, agreement) along the horizontal axis and the negative terms (e.g., negativity, disagreement) along the vertical axis. Moreover, different rules for classifying a response as Positive, Negative, Indifferent, or Ambivalent have been proposed and used (Audrezet, 2014; Audrezet et al., 2016; Borriello, 2017). While considerable work has been done to show that ESG-type items should be considered by researchers, there are still open questions about how best to analyze the responses and the situations that they are best suited for.

Methods

Research Questions

The purpose of this pilot study is to document validity evidence (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014) about the use of ESG-type items to inform their use in other studies in the field of statistics education or that adopt the EVT framework. To that end, we aim to answer the following questions:

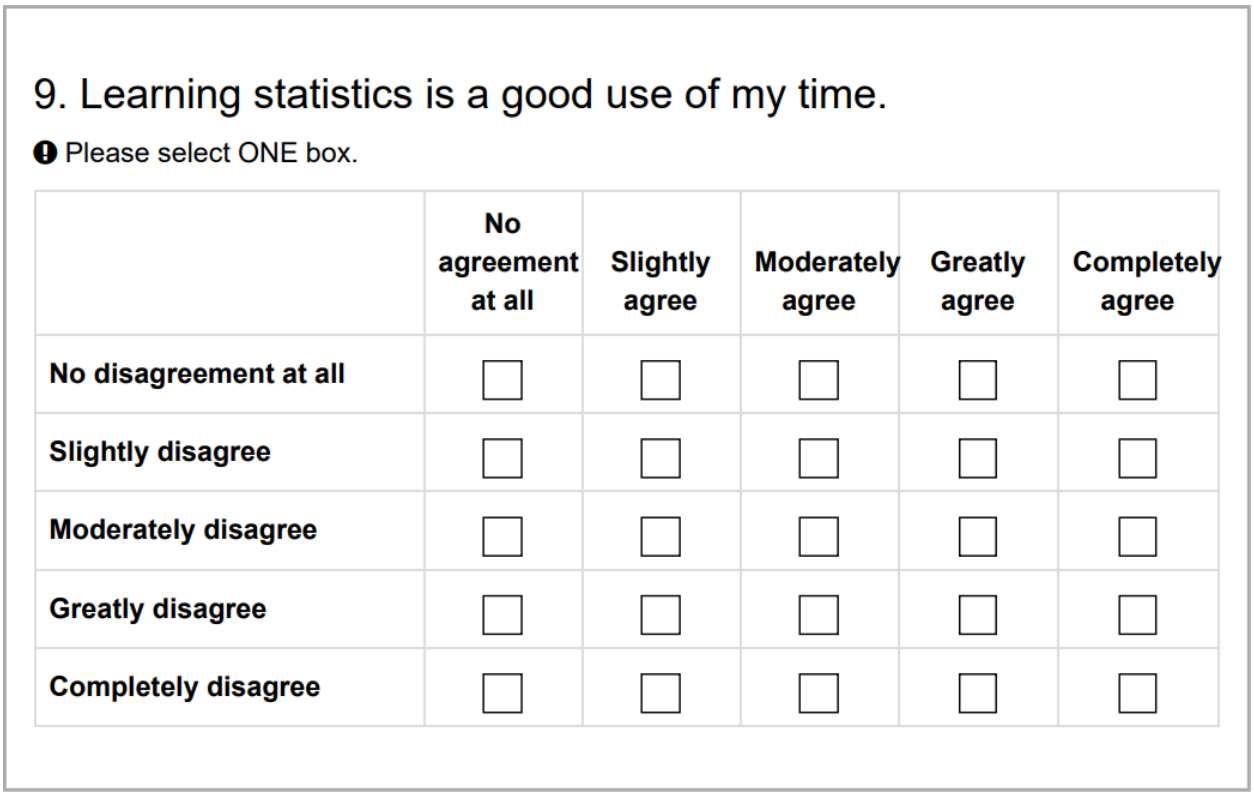
1. Are students able to understand and meaningfully respond to ESG-type items?
2. Do ESG-type items provide for a more nuanced neutral category than Likert-type items for the EVT Cost construct?

Instrument

The instrument used in this pilot study consisted of items from existing scales and items developed for this study; the items and instructions from the instrument are included in Appendix A. Using three existing scales, 19 items were presented to participants as both a Likert-type item and as an ESG-type item. Each Likert-type item was presented using a 9-point scale with the following anchors: *Completely disagree, Greatly disagree, Moderately disagree, Slightly disagree, Neither agree nor disagree, Slightly agree, Moderately agree, Greatly agree, Completely agree*. The axes for the ESG-type items were labeled: *No agreement at all, Slightly agree, Moderately agree, Greatly agree, Completely agree* (horizontal axis) and *No disagreement at all, Slightly disagree, Moderately disagree, Greatly disagree, Completely disagree* (vertical axis); an example item is shown in Figure 2. No changes were made to the item stem.

Figure 2

Example of an ESG-type item presented to participants in LimeSurvey.



9. Learning statistics is a good use of my time.

Please select ONE box.

	No agreement at all	Slightly agree	Moderately agree	Greatly agree	Completely agree
No disagreement at all	☐	☐	☐	☐	☐
Slightly disagree	☐	☐	☐	☐	☐
Moderately disagree	☐	☐	☐	☐	☐
Greatly disagree	☐	☐	☐	☐	☐
Completely disagree	☐	☐	☐	☐	☐

This study was motivated by challenges developing a Cost scale in another project: the S-SOMAS instrument (Unfried et al., 2021; Whitaker et al., 2019). To that end, items from the Cost scale from the first pilot version of the S-SOMAS instrument were chosen for the instrument in this study. The S-SOMAS Cost scale is intended to be a scale for measuring EVT Cost in general rather than focusing on a specific component of cost and will be referred to as Overall Cost throughout. Four of the eight Overall Cost scale items should be reverse-coded; to do this in an ESG-type items, the transpose of the response grid is used. Because the S-SOMAS Cost scale is in development, two established scales were also chosen: the Task Effort Cost and Emotional Cost scales developed by Flake, Barron, Hulleman, McCoach, and Welsh (2015). The Task Effort Cost scale was chosen because it measures a component of cost that has been previously studied in statistics education, and Emotional Cost was chosen because of its similarity to statistics anxiety, a distinct construct that may be accounted for in the EVT model by the Cost construct (e.g., Whitaker & White, 2020).

There were 44 fixed-choice items on the survey (19 Likert-type, 22 ESG-type, 3 multiple choice) and one free-response item total. The survey was administered using LimeSurvey, and the items and instructions are included in Appendix A. Nineteen of the items appeared in both a Likert-type and an ESG-type item, for a total of 38 items. These 19 items were randomly ordered and divided into two sets: Set A was items 1-9, and Set B was items 10-19. On the survey, these items were presented across four pages in the following

order: Set A (Likert-type), Set B (ESG-type), Set B (Likert-type), then Set A (ESG-type). Three additional ESG-type items were developed for this study and presented on page 5; three multiple-choice questions and one free-response question about the experience of taking the survey were presented to respondents on page 6. These additional items were written by the lead researcher based on discussions with colleagues and other members of the research team.

As shown in Figure 2, due to limitations in LimeSurvey the ESG-type items were not presented as they are typically discussed in the literature (cf. Figure 1). The ESG is usually shown with the horizontal axis labeled along the bottom of the item rather than the top (e.g., Audrezet et al., 2016; Larsen et al., 2009), which makes the bottom left corner a natural origin point (i.e., the bottom left cell corresponds to "No agreement at all" and "No disagreement at all"). Because LimeSurvey could only be configured to display the labels across the top of the item, the axis labels could not meet at the bottom left corner cell and instead met at the top left corner cell. We strongly felt that the origin point should where the labels for the axes meet, and so the order of the disagreement scale was reversed. That is, in the instrument used in this study, participants were asked to choose cells lower on the vertical axis to indicate greater disagreement with the statement being considered (rather than choosing cells higher on the vertical axis as in the ESG usually presented in the literature).

Participants

Anonymous data were collected from students in a multi-section statistics course at a primarily undergraduate university in Atlantic Canada; the study was cleared by the university's Research Ethics Board (Clearance File #2019-134). Data collection occurred from late February through early April 2020; the statistics course was abruptly shifted from face-to-face to remote learning in March 2020 due to the COVID-19 pandemic. We analyzed the data as a single group rather than distinguishing between before-disruption and after-disruption based on the results of a brief statistical analysis (described below). The statistics course was the second course in an introductory statistics sequence (covering topics such as multiple linear regression, analysis of variance, and chi-squared testing), so all students had completed at least one statistics course before being asked to participate in the study.

Students were recruited for participation through general emails sent by their instructors, and only students who had reached the age of majority (nineteen) were allowed to participate. A voluntary appeal with no compensation for participants was chosen because the intention of this study is not to characterize the views about statistics for a particular group of students; rather, only an examination of the characteristics of ESG-type items is of interest. The total enrollment across all sections of the second-semester statistics course from which participants were sampled was 316. A total of 42 students responded to at least one item on the survey; 24 students responded to every fixed-choice item on the survey. Six respondents completed only the first page of Likert-type items, and a further six respondents completed the first page of Likert-type items and the first page of ESG-type items. Three respondents skipped between one and three fixed-choice items. Five participants responded to the free-response item.

Data Analyses

Before performing analyses to answer the research questions, a brief analysis of the COVID-19 disruption was conducted using MANOVA. Assumptions were checked using chi-squared plots and the multivariate Shapiro-Wilk test. To answer the first research question, we examined participants' responses to the questions about their experience doing the survey using descriptive statistics. Owing to the small sample size, we also examined the average time students spent on the items for each page using Friedman's Test and Conover's post hoc paired comparisons (Conover, 1999). The hypotheses for Friedman's Test are:

H_0: The distribution of average seconds per item is the same for all pages.

H_A: The distribution of average seconds per item is different for at least two pages.

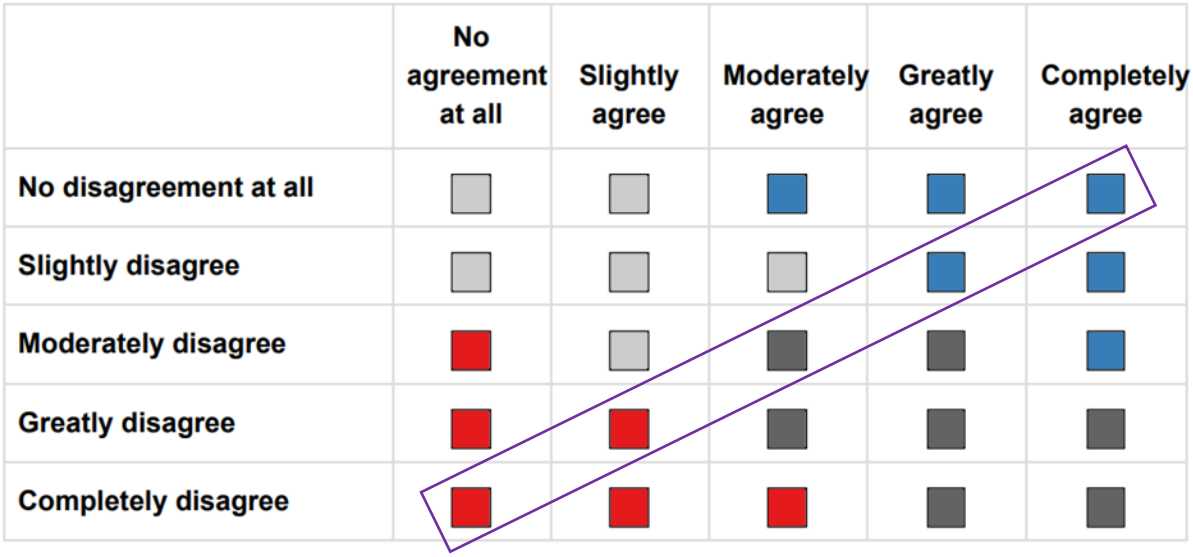
The correlation between the Likert-type responses and the unidimensional scores for the ESG-type items was computed for each pair of items using both the Pearson and Spearman methods. To compute a unidimensional score for an ESG-type item, Audrezet et al. (2016) proposed $S(i,j) = (b+2)i + bj - 1 - 6b$ where $-1 < b < 0$ and $b = -0.5$ by determining which values satisfy constraints imposed on $S(i,j)$. We also examined the correlations between the responses to the Likert-type items and the ESG-type item scores using $S(i,j)$ when $b = -1$. Lastly, we examined the results to the three ESG-type items developed for this study. These items were written specifically to elicit a particular pattern of responses, and the extent to which responses follow this pattern is evidence about the degree to which participants can meaningfully engage with these items. The expected pattern is that respondents should respond strongly negatively and not at all positively to the first item, respond not at all negatively and strongly positively to the third item, and respond between these extremes to the second item: the expected change in response from the first to second to third item is essentially movement along the diagonal associated with a reciprocal relationship.

Two measures of internal consistency are reported: coefficient alpha (Cronbach, 1951) and Guttman's Lambda-6 (1945). Commonly referred to as reliability indices, these coefficients are measures of item homogeneity within scales. While these coefficients are often misinterpreted or given undue weight (Henson, 2001; Schmitt, 1996), reporting information about the reliability and precision of estimates is an important part of psychological measurement (AERA et al., 2014). These coefficients are reported so that their values with Likert-type items can be compared to previously published results for the scales (i.e., the Task Effort Cost, Emotional Cost, and Overall Cost scales), to compare the scales with Likert-type items to the scales with ESG-type items, and to provide estimates to which future work can be compared.

To answer the second research question, we classified responses as being Positive, Negative, Indifferent, or Ambivalent as described by Audrezet et al. (2016) and shown in Figure 3. In the ESG-type items, there are six possible responses for each of Positive, Negative, and Indifferent, and seven possible responses that are classified as Ambivalent. Similarly, for the Likert-type items, the positive attitudes correspond with a response of 7, 8, or 9 (*Moderately agree, Greatly agree, Completely agree*); the negative attitudes correspond with a response of 1, 2, or 3 (*Completely disagree, Greatly disagree, Moderately disagree*); the

Figure 3

The classification corresponding to each of the cells in the ESG-type items using the method proposed by Audrezet et al. (2016): blue indicates positive attitudes (top right corner), red indicates negative attitudes (bottom left corner), light gray indicates indifferent attitudes (top left corner), and dark gray indicates ambivalent attitudes (bottom right corner). A purple rectangle has been superimposed indicating the responses that would be expected if a reciprocal relationship exists between the poles on the axes of the ESG-type item (i.e., if increasing agreement necessarily means a corresponding decrease in disagreement.)



Neutral responses correspond to a 4, 5, or 6 (*Slightly disagree*, *Neither agree nor disagree*, *Slightly agree*). Note that the Likert-type items have a combined Neutral category in lieu of separate Indifferent and Ambivalent categories. Then, we present summaries of the counts and proportions of the responses in each of the four categories for each item.

Next, we aggregate the responses for the items in each scale and conduct trinomial tests (Bian, McAleer, & Wong, 2011; Ganesalingam, 1994) to determine if the Ambivalent and Indifferent categories seem appropriate for this construct. If the more nuanced Indifferent and Ambivalent neutral categories that can be identified using ESG-type items are *not* appropriate for the SEVT Cost construct, we would expect to see participants with neutral attitudes tending to respond along the reciprocal diagonal. That is, if positivity and negativity are reciprocally related for a construct, we would expect to see responses falling along the diagonal from high-negative, low-positive to low-negative, high positive (see Figure 3). Any off-diagonal responses to an ESG-type item measuring a construct for which the positive and negative poles do have a reciprocal relationship would essentially be random error. With a paucity of literature about ESG-type items, we have no theoretical basis to expect that off-diagonal responses would tend to be either above or below the diagonal, and so will assume that an off-diagonal response is equally likely to be above or below the diagonal (in the situation that a true

reciprocal relationship exists for the poles of the construct in question).

We performed three sets of trinomial tests comparing the responses:

A. above the diagonal, below the diagonal, and on the diagonal;
B. above the diagonal region, below the diagonal region, and in the diagonal region (on the diagonal, superdiagonal, or subdiagonal); and
C. classified as Indifferent, classified as Ambivalent, and classified as either Positive or Negative.

The specific hypotheses for each test will be:

$$H_0: p_1 = p_2$$
$$H_A: p_1 \neq p_2$$

where p_1 and p_2 represent the proportions for the first and second groups listed above and the third named group is viewed as the third category. The trinomial test has higher power than other tests that ignore the third category (Bian et al., 2011; Ganesalingam, 1994).

Trinomial tests A and B provide evidence about whether or not a reciprocal relationship exists for the SEVT Cost construct. In trinomial test A, responses exactly on the diagonal (the purple rectangle in Figure 3) are compared with responses above it and responses below it. In trinomial test B, responses that are either exactly on the diagonal or just off the diagonal (the diagonal region) are compared with responses above and below the diagonal region. If the trinomial test provides substantial evidence against the null hypothesis, then there would appear to be an asymmetry in how participants are responding off the diagonal. As previously stated, if there is a reciprocal relationship for a construct, off-diagonal responses are thought to be essentially random error and presumed to be equally likely to occur above or below the diagonal.

Note that it is conceivable that a reciprocal relationship could exist for a construct and off-diagonal responses occur in such a way that it is not equally likely to respond above or below the diagonal. However, considering the novelty of the ESG-type items and research into whether constructs have a reciprocal relationship, there is not yet a theoretical reason to expect that this might be the case. Characterizing potential patterns of off-diagonal responses for constructs whose poles have a reciprocal relationship is beyond the preliminary work of this manuscript. Still, conclusions from trinomial tests A and B are predicated on the appropriateness of the assumption of off-diagonal responses being equally likely to be above or below and this will be acknowledged in the discussion.

The regions above the diagonal and below the diagonal each contain two positive and two negative response options. To further explore whether the Neutral category for the SEVT Cost construct can be reasonably decomposed into Indifferent and Ambivalent categories, a trinomial test the aggregates positive and negative into a single category is performed. In trinomial test C, the analysis is no longer focused on the diagonal shown in Figure 3. Instead, responses that have a polarity (i.e., are classified as Positive or Negative) are the third category and compared with responses classified as Indifferent and responses classified as Ambivalent. The conclusions for trinomial test C are dependent upon essentially the same assumption as trinomial tests A and B.

All analyses were done in R (R Core Team, 2021). To create the graphs in this manuscript, the following packages were used: ggplot2 (Wickham, 2009), ggforce (Pedersen, 2021), pheatmap (Kolde, 2019), corrplot (Wei & Simko, 2017), and MVQuickGraphs (Whitaker & Hebert, 2021). The GridItemTools package (Whitaker, Drew, & Barss, 2021) was used to process the raw LimeSurvey data, analyze the ESG-type items, and conduct the trinomial test described by Ganesalingam (1994). The PMCMRplus package (Pohlert, 2021) was used to conduct Conover's post hoc paired comparisons. The mvnormtest package (Jarek, 2012) was used to conduct the Multivariate Shapiro-Wilk test. The psych package (Revelle, 2021) was used for calculating internal consistency coefficients.

Results

COVID-19 Disruption

Due to anonymous nature of the data collection in LimeSurvey, specific dates that students responded to the survey were not logged. However, using periodic file downloads, it was possible to assemble to two groups that approximately corresponded to students who completed the survey before the COVID-19 disruption (Before, 21 students) and those who completed it after the disruption (After, 21 students). No demographic or other identifying information was collected, so we compared the three scale scores for the students in the Before and After groups using MANOVA; boxplots of these scores are shown in Figure 4. A multivariate Shapiro-Wilk test was conducted which provided evidence against multivariate normality for the scale scores ($W = 0.9192$, $p = 0.0057$). One multivariate outlier (in the After group) was identified using a Chi-Square Plot (see Figure 5) and removed; the multivariate Shapiro-Wilk test then indicated essentially no evidence against multivariate normality using the remaining 41 observations ($W = 0.9667$, $p = 0.2671$). MANOVA was conducted to determine if there was evidence of any differences in means among the scale scores for the two groups using the 41 observations: there was essentially no evidence of any differences in means with $F(3, 37) \approx 1.62$ and $p = 0.20$. Because there was essentially no evidence of differences in mean scale scores for the Before and After groups, we will analyze the complete dataset without distinguishing between the two groups for all future analyses.

Analysis of Students' Experiences with ESG-type Items

We now focus on answering the first research question: Are students able to understand and meaningfully respond to ESG-type items? We first analyze their responses to the multiple-choice items that asked about their experiences. Then we examine the correlations between Likert-type items and ESG-type items. Lastly, we examine responses to three ESG-type items designed to elicit responses along a continuum.

Figure 4

Boxplots showing the scale scores for the students who completed the survey before and after the COVID-19 disruptions.

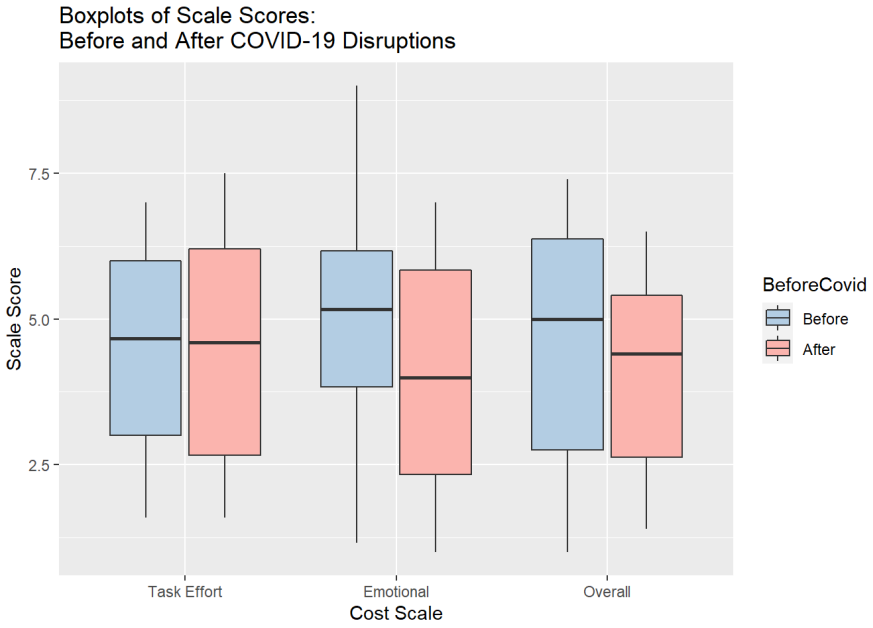
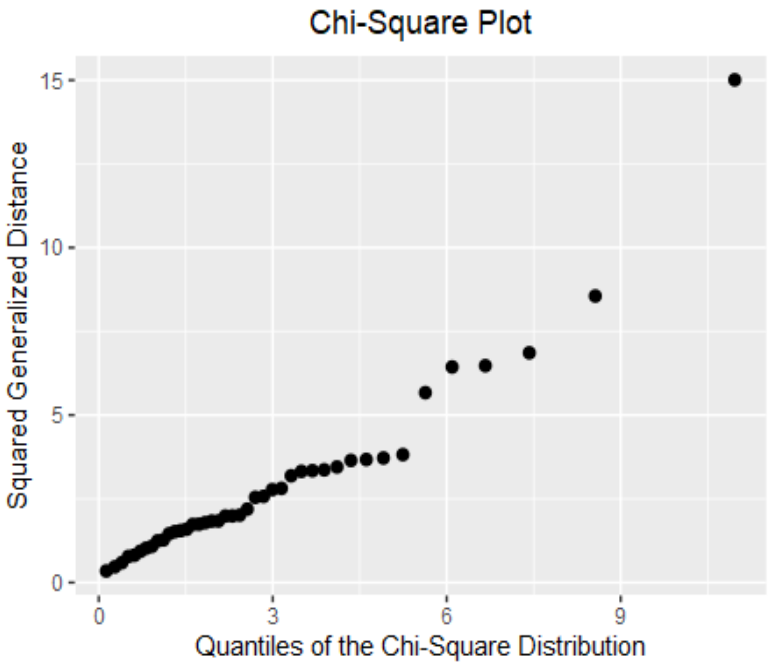


Figure 5

A Chi-Square Plot showing the squared generalized distance for each 3-tuple of scale scores plotted against the quantiles of the Chi-Square distribution with 3 degrees of freedom. One outlier is visible (furthest right point).



Multiple-choice Item Responses: On the last page of the survey, three multiple-choice items about students' experiences completing the survey were presented. Twenty-five students responded to each item, and summaries of these responses are shown in Tables 1, 2, and 3. The first question directly asked students about their understanding of using ESG-type items to respond.

Because of the novelty of the ESG-type items, examples (included in Appendix A) of how to use them to respond were presented on each page with ESG-type items. The second multiple-choice item asked about students' perceptions of these examples. The last multiple-choice item about their experiences that students were presented asked them to compare responding with Likert-type and ESG-type items.

Response Time Analysis: The Likert-type and ESG-type items were presented across five pages of the survey in this order: nine Likert-type items (page 1), ten ESG-type items (page 2), ten Likert-type items (page 3), nine ESG-type items (page 4), then three ESG-type items (page 5). We wished to compare the average number of seconds spent per item for the pages; only students who responded to items on each page of the survey are included in these analyses. Due to the novelty of the ESG-type items, we expected that the average time per item would be higher for these pages than for the items with Likert-type items. Figure 6 shows boxplots of the average times per item by page after removing one outlier (a student who took more than an hour to respond to the items on page 3), and Table 4 presents the summary statistics for each page.

Table 1

Summaries of the responses to the first multiple-choice item (Which of the following best describes your understanding of the grid item?).

Response	Count (Percentage)
I understood how to use the grid to respond right away.	10 (40%)
It took some time to understand how to use the grid, but I understood it by the end.	12 (48%)
I didn't really understand how to use the grid to respond.	3 (12%)

Table 2

Summaries of the responses to the second multiple-choice item (Were the examples for the grid response items helpful?).

Response	Count (Percentage)
Yes, the examples were good.	15 (60%)
Yes, the examples helped but could be improved.	6 (24%)
No, the examples did not help.	3 (12%)
No, the examples confused me more.	1 (4%)

Table 3

Summaries of the responses to the third multiple-choice item (Did you find the grid item allowed you to give a response that better reflected your reaction to each item compared to the traditional 1-9 Disagree to Agree scale?).

Response	Count Percentage)
Yes, I was able to give better responses using the grid.	13 (52%)
Neither method was better than the other for responding.	2 (8%)
No, I think I was able to give better responses using the traditional scale.	10 (40%)

Figure 6

Boxplots showing the average number of seconds per item taken by respondents on each page of the survey. Boxplots are coloured to indicate which type of items was presented on the page. One outlier has been removed.

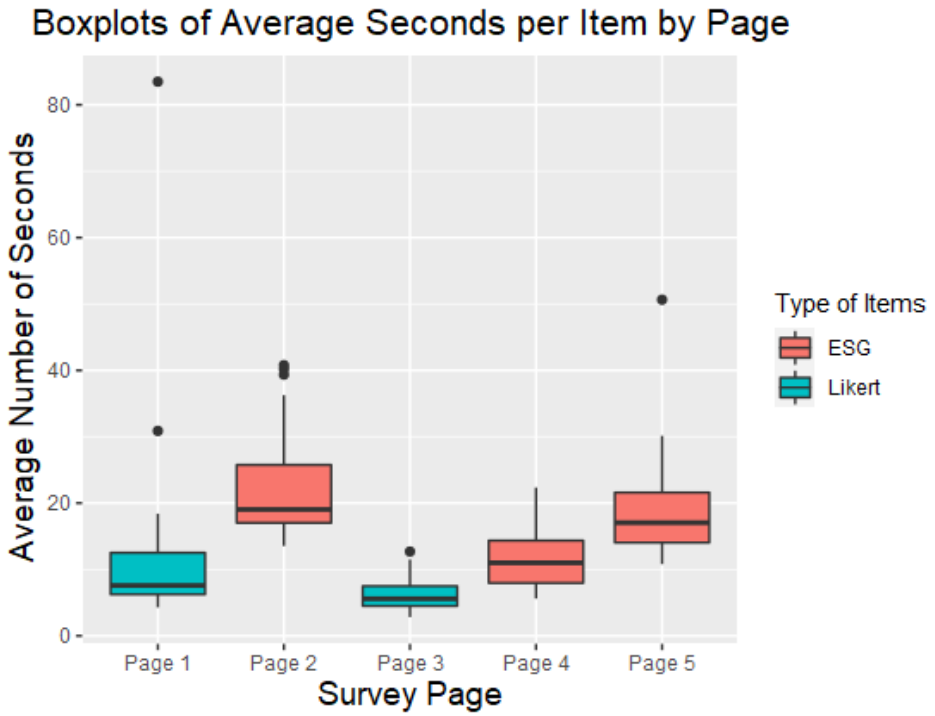


Table 4

Summary statistics for the average number of seconds per item taken by respondents on each page of the survey. One outlier has been removed.

Page	Mean	SD	Min	Q_1	Median	Q_3	Max
Page 1	13.02	16.13	4.28	6.24	7.59	12.52	83.49
Page 2	22.63	8.64	13.54	17.02	19.05	25.76	40.78
Page 3	6.45	2.72	2.83	4.50	5.59	7.50	12.70
Page 4	11.67	4.88	5.61	7.96	11.01	14.36	22.35
Page 5	19.09	8.27	10.80	14.05	17.05	21.59	50.64

Using Friedman's Test, we found overwhelming evidence against the null hypothesis that all pages had the same distribution of the average number of seconds per item for the survey pages with X^2 (4, N = 24) = 70.37 and $p < 0.0001$. The p-values from Conover's post hoc paired comparison tests are shown in Table 5. Based on these p-values, there is strong evidence of differences in distributions among some of the pages. For example, there is overwhelming evidence of a difference in distributions ($p < 0.0001$) between the first page of Likert-type items (Page 1) and the first page of ESG-type items (Page 2). However, there is also essentially no evidence of differences in distributions among some of the pages. For example, there is essentially no evidence of a difference in distributions (p = 0.8575)

between the first page of Likert-type items (Page 1) and the second page of ESG-type items (Page 4).

Table 5

The p-value for Conover's post hoc paired comparison test comparing each pair of pages.

Pages Compared		*p*-value
Page 1 (Likert)	Page 2 (ESG)	< 0.0001
Page 1 (Likert)	Page 3 (Likert)	0.0660
Page 1 (Likert)	Page 4 (ESG)	0.8575
Page 1 (Likert)	Page 5 (ESG)	0.0019
Page 2 (ESG)	Page 3 (Likert)	< 0.0001
Page 2 (ESG)	Page 4 (ESG)	0.0019
Page 2 (ESG)	Page 5 (ESG)	0.8575
Page 3 (Likert)	Page 4 (ESG)	0.0027
Page 3 (Likert)	Page 5 (ESG)	< 0.0001
Page 4 (ESG)	Page 5 (ESG)	0.0518

Internal Consistency of Scales: Guttman's Lambda-6 and a 95% confidence interval for coefficient alpha are presented in Table 6; each coefficient was computed for all three scales using the Likert-type and ESG-type items. For the ESG-type items, responses were transformed to a unidimensional value using $S(i,j)$ with the values $b = -0.5$ and $b = -1$. The values of coefficient alpha for the Emotional and Task Effort scales in this study, 0.92 and 0.95, respectively, are similar to the values reported by Flake et al. (2015), 0.94 and 0.95, respectively. Coefficient alpha has not been previously reported for the Overall Cost scale. For the Emotional and Task Effort scales, the values of the reliability indices are largely similar for Likert-type and ESG-type scales. For the Overall Cost scale, the values of Guttman's Lambda-6 and coefficient alpha are lower for ESG-type scales than the Likert-type scale, but the 95% confidence intervals for coefficient alpha overlap.

Table 6

Internal consistency coefficients computed for the scales using the Likert-type items and ESG-type items. Prior to calculating the coefficients, the ESG-type items were scored using $S(i,j)$ with the values $b = -0.5$ and $b = -1$. Coefficient alpha (with a 95% confidence interval) and Guttman's Lambda-6 are presented.

		Coefficient Alpha			
Item Type	Cost Scale	95% CI Lower Limit	Estimate	95% CI Upper Limit	Guttman's Lambda-6
Likert	Emotional	0.89	0.92	0.95	0.93
	Task Effort	0.93	0.95	0.97	0.94
	Overall	0.83	0.88	0.92	0.92
ESG scored with b = -0.5	Emotional	0.86	0.90	0.94	0.92
	Task Effort	0.90	0.93	0.96	0.94
	Overall	0.72	0.80	0.88	0.89
ESG scored with b = -1	Emotional	0.89	0.92	0.95	0.93
	Task Effort	0.90	0.93	0.96	0.93
	Overall	0.75	0.82	0.89	0.89

Correlations Among Likert-type and ESG-type Items: Each of the nineteen items presented in both a Likert-type form and an ESG-type form used the same item stem: only the response options differed. Therefore, we expect moderate-to-strong correlations between responses to Likert-type items and responses to ESG-type items converted to a unidimensional score using the function given by Audrezet et al. (2016). In the function presented by Audrezet et al., the value of b is constrained to $-1 < b < 0$ by a requirement that ambivalence increases for scores higher on the diagonal (e.g., a response of *Extremely negative* [5] and *Extremely positive* [5] indicates greater ambivalence than a response of *Quite a bit negative* [4] and *Quite a bit positive* [4]). Though the value $b = -0.5$ is recommended by Audrezet et al., we will also examine correlations using $b = -1$ which corresponds to an equal amount of neutrality along the diagonal (e.g., a response of *Extremely negative* [5] and *Extremely positive* [5] indicates the same degree of neutrality as a response of *Quite a bit negative* [4] and *Quite a bit positive* [4]) because this seems more consistent with the aim of studying the similarity between Likert-type and ESG-type responses.

Pearson correlations and Spearman correlations for each pair of Likert-type and ESG-type items for each value of b are shown in Table 7 along with the number of responses to that item and the proportion of responses that would be classified as ambivalent or indifferent (see Figure 3). Eighteen of the 19 items have a correlation of at least 0.40 using both the Pearson and Spearman methods for both values of b. For $b = -0.5$, four of the 19 items had a correlation of at least 0.70, and for $b = -1$ seven of the 19 items have such a correlation using both methods. A graph illustrating the Pearson correlations among all 19 Likert-type and 19 ESG-type items is included in Appendix B.

Examination of Three ESG-type Items: Three ESG-type items were designed to elicit responses along a continuum: one was designed to elicit mostly negative responses (*I worked on my homework for 20 hours, and I didn't really understand it.* [Patterned ESG Item 1]), another was designed to elicit mostly positive responses (*I worked on my homework for 20 hours, and I understood all of it and learned some new things*. [Patterned ESG Item 3]), and yet another was designed to elicit mixed responses (*I worked on my homework for 20 hours, and I think I understood most of it.* [Patterned ESG Item 2]). It was expected that, if students understood how to use the ESG-type items, that there would be a tendency for people to respond strongly negative (toward the bottom left corner) to the first item (Patterned ESG Item 1) and to respond strongly positively (toward the top right corner) to the third item (Patterned ESG Item 3); this would be observable in both the aggregated responses and in individual response patterns. These response patterns were expected because the items were written so that Patterned ESG Item 1 would describe a universally negative situation in an educational context (working for many hours without understanding), while Patterned ESG Items 2 and 3 were written to be relatively more positive experiences (working for the same number of hours but with increasing amounts of understanding).

Table 7

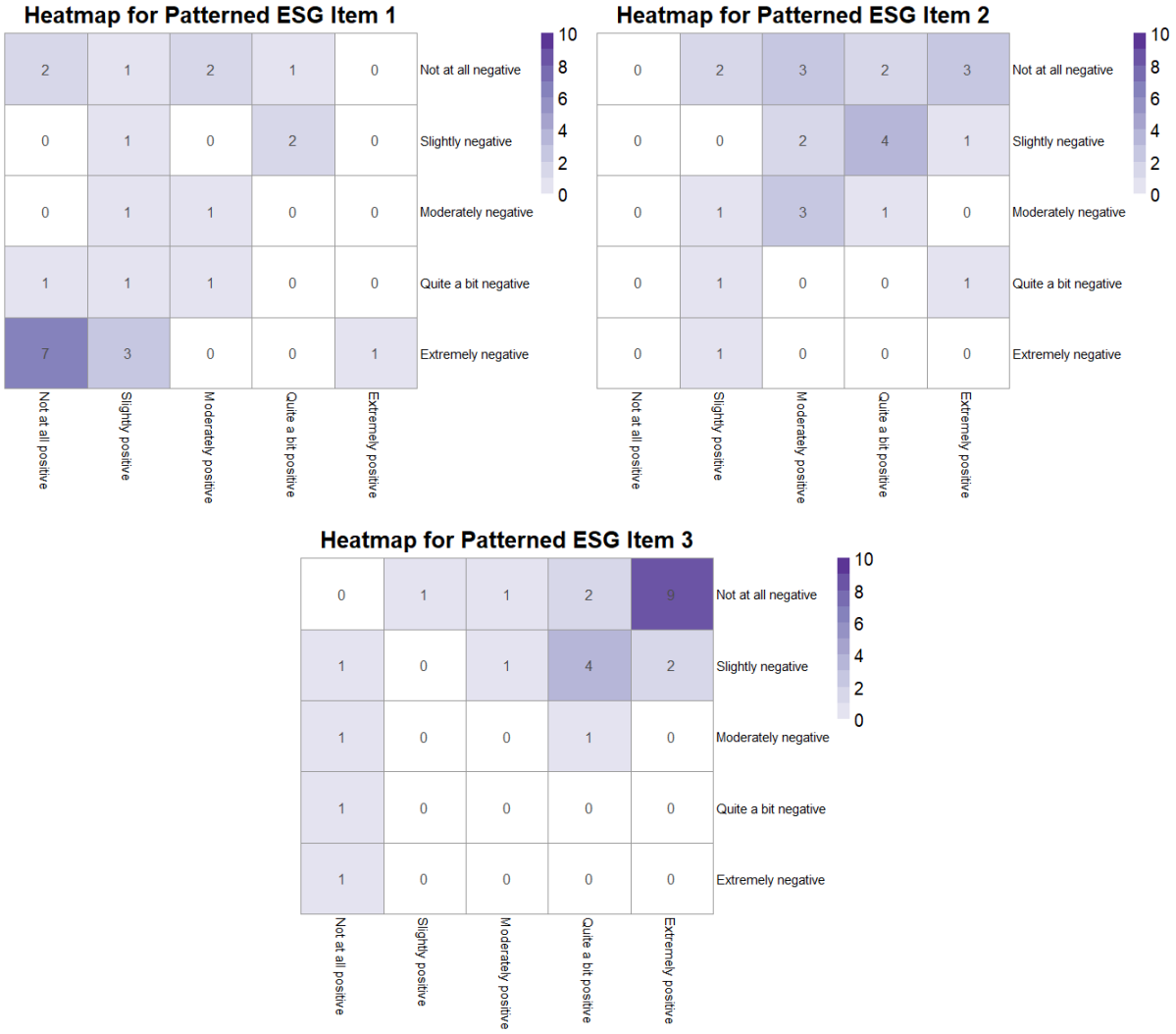
A Table showing the pairwise correlations between the Likert-type and ESG-type items. A "-" in the item name indicates that the Likert-type item is reverse-coded.

	Pearson		Spearman		Number of
Item	b = -0.5	b = -1	b = -0.5	b = -1	Responses
Emotional Cost 1	0.53	0.59	0.51	0.55	33
Emotional Cost 2	0.73	0.84	0.71	0.81	33
Emotional Cost 3	0.76	0.82	0.77	0.82	33
Emotional Cost 4	0.59	0.67	0.62	0.70	33
Emotional Cost 5	0.63	0.78	0.62	0.79	33
Emotional Cost 6	0.63	0.65	0.61	0.60	33
Task Effort Cost 1	0.65	0.78	0.62	0.76	27
Task Effort Cost 2	0.42	0.55	0.48	0.58	27
Task Effort Cost 3	0.72	0.71	0.70	0.72	27
Task Effort Cost 4	0.62	0.66	0.59	0.64	27
Task Effort Cost 5	0.53	0.56	0.47	0.57	26
Overall Cost 1-	0.64	0.63	0.65	0.55	32
Overall Cost 2	0.60	0.69	0.61	0.66	33
Overall Cost 3	0.55	0.75	0.58	0.73	33
Overall Cost 4	0.71	0.81	0.72	0.82	32
Overall Cost 5	0.33	0.26	0.33	0.31	27
Overall Cost 6-	0.52	0.57	0.51	0.54	27
Overall Cost 7-	0.60	0.65	0.49	0.54	27
Overall Cost 8-	0.69	0.65	0.68	0.67	27

Figure 7 shows the counts of the number of respondents in each cell for each of these three ESG-type items. In the aggregate, the expected pattern can be seen; Appendix B includes graphs that illustrate that individuals responded in the expected ways to produce the aggregate pattern. Responses to Patterned ESG Item 1 tended to be strongly negative and not at all positive, responses to Patterned ESG Item 3 tended to be strongly positive and not at all negative, and responses to Patterned ESG Item 2 were in between these two extremes. Of the 25 students who responded to these three items, 17 showed an increase in positivity from the first item to the third item. Five other students showed no change in positivity, but for three of these five their negativity decreased. Three students showed a decrease in positivity from the first to the third item. These patterns are consistent with the expected responses to the items.

Figure 7

Heatmaps showing the number of responses in each cell to the three ESG-type items designed for this study.

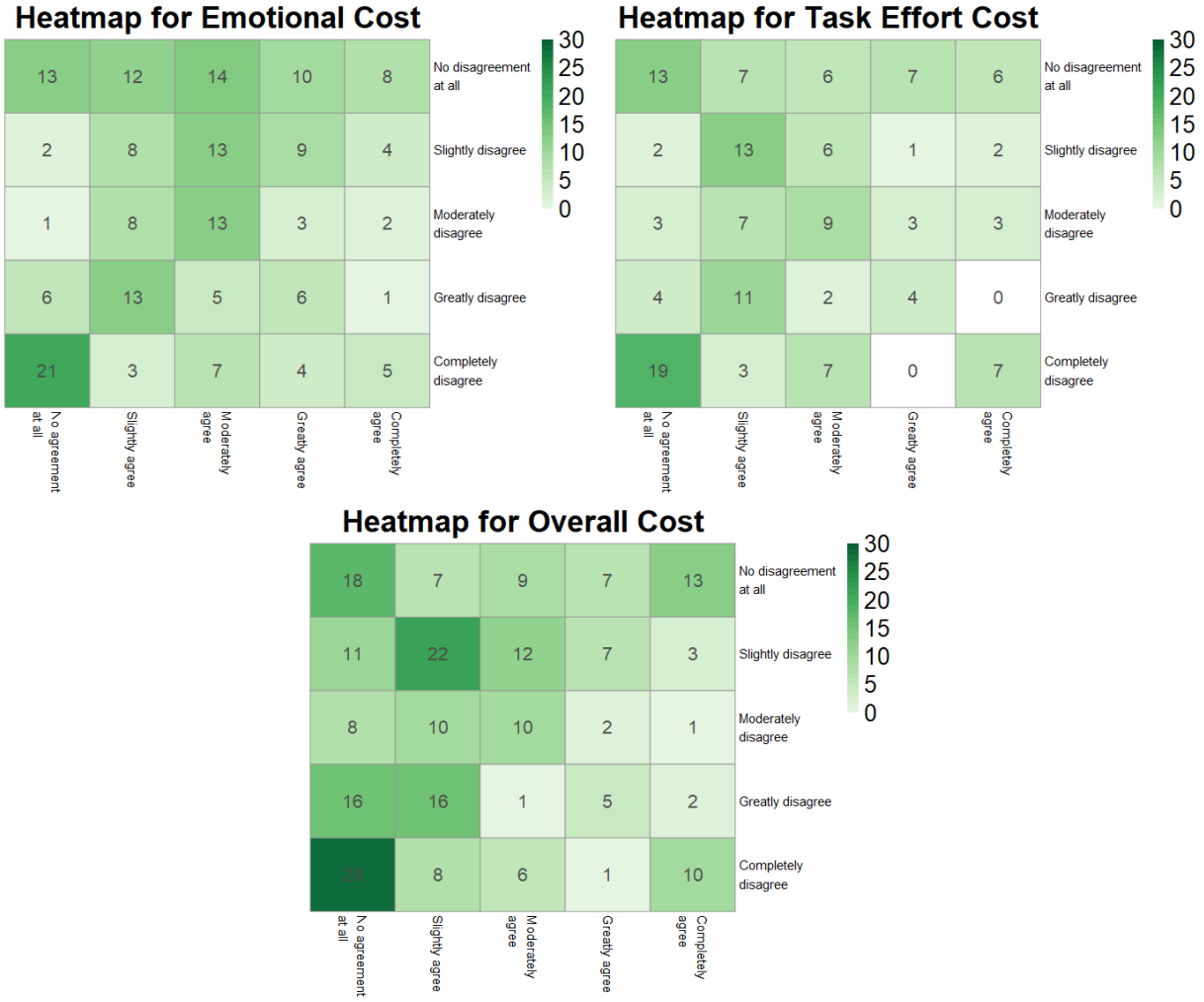


Analysis of Neutral, Indifferent, and Ambivalent Responses

The trinomial test will be used in three different analyses to assess whether there appears to be a reciprocal relationship between positivity and negativity for the EVT Cost construct. Heatmaps for the total number of responses in each cell for the ESG-type items across all items for each scale are shown in Figure 8. There were 191 responses across the six items in the Emotional Cost scale, 145 responses across the five items in the Task Effort Cost scale, and 234 responses across the eight items in the Overall Cost scale.

Figure 8

Heatmaps for the total number of responses in each cell for the ESG-type items across all items for each scale.



The trinomial test is used to compare the proportions for the first two of each of these groups:

A. Table 8 presents the results for comparing above the diagonal with below the diagonal; on the diagonal is the third group.
B. Table 9 presents the results for comparing above the diagonal region with below the diagonal region; in the diagonal region (on the diagonal, superdiagonal, or subdiagonal) is the third group.
C. Table 10 presents the results for comparing classified as Indifferent with classified as Ambivalent; classified as either Positive or Negative is the third group.

For trinomial test A, there is weak evidence against $H_0: p_{above} = p_{below}$ for the Task Effort Cost scale (p = 0.0972), strong evidence against the null hypothesis for the Overall Cost scale (p = 0.0010), and essentially no evidence against the null hypothesis for the Emotional Cost scale (p = 0.6492). For trinomial test B, there is strong evidence against $H_0: p_{aboveR} = p_{belowR}$ for the Task Effort Cost scale (p = 0.0112) and Overall Cost (p = 0.0001), but essentially no evidence against the null hypothesis for Emotional Cost scale (p = 0.7237). For trinomial test C, there is borderline

evidence against $H_0: p_{ind} = p_{amb}$ for the Emotional Cost scale (p = 0.0549), strong evidence against the null hypothesis for the Task Effort Cost scale (p = 0.0083), and overwhelming evidence against the null hypothesis for the Overall Cost scale ($p < 0.0001$)

Table 8

Counts and proportions for the number of responses above the diagonal (above), below the diagonal (below), and on the diagonal (ondiag) for each scale along with the p-value from the trinomial test.

Cost Scale	n_{above}	n_{ondiag}	n_{below}	$\hat{p}_{above}$	$\hat{p}_{ondiag}$	$\hat{p}_{below}$	p-value
Emotional	76	45	70	0.40	0.24	0.37	0.6492
Task Effort	41	46	58	0.28	0.32	0.40	0.0972
Overall	63	65	106	0.27	0.28	0.45	0.0010

Table 9

Counts and proportions for the number of responses above the diagonal region (aboveR), below the diagonal region (belowR), and in the diagonal region (indiagR) for each scale along with the p-value from the trinomial test.

Cost Scale	n_{aboveR}	$n_{indiagR}$	n_{belowR}	$\hat{p}_{aboveR}$	$\hat{p}_{indiagR}$	$\hat{p}_{belowR}$	p-value
Emotional	47	93	51	0.25	0.49	0.27	0.7237
Task Effort	25	73	47	0.17	0.50	0.32	0.0112
Overall	40	111	83	0.17	0.47	0.35	0.0001

Table 10

Counts and proportions for the number of responses classified as Indifferent (ind), classified as Ambivalent (amb), and classified as either Positive or Negative (posneg) for each scale along with the p-value from the trinomial test.

Cost Scale	n_{ind}	n_{posneg}	n_{amb}	$\hat{p}_{ind}$	$\hat{p}_{posneg}$	$\hat{p}_{amb}$	p-value
Emotional	56	98	37	0.29	0.51	0.19	0.0549
Task Effort	48	72	25	0.33	0.50	0.17	0.0083
Overall	80	123	31	0.34	0.53	0.13	< 0.0001

Discussion

This study sought to provide preliminary evidence about the appropriateness of using ESG-type items to measure EVT constructs within the context of statistics education. The results from this study will contribute to the response process validity evidence (AERA et al., 2014) for future studies that employ ESG-type items. The specific research questions addressed were:

1. Are students able to understand and meaningfully respond to ESG-type items?
2. Do ESG-type items provide for a more nuanced neutral category than Likert-type items for the EVT Cost construct?

We will now synthesize the results presented above to directly answer these questions.

Evidence of Students' Understanding of ESG-Type Items

Across the analyses, there seems to be evidence that students understood how to respond using the grid for the ESG-type items. Based on the 25 responses to the first evaluation item (*Which of the following best describes your understanding of the grid item?*), 88% of respondents reported understanding the ESG-type items either right away or by the end of the study (see Table 1). Continued exposure to ESG-type items seems to result in increased familiarity with the item type: 48% of the respondents to the first evaluation item chose the option *I understood how to use the grid to respond right away* (see Table 2). Moreover, students were able to respond more quickly to the second page of ESG-type items, providing further evidence that increased familiarity may overcome some of the challenges associated with the novelty of the items. Conover's post hoc paired comparisons show strong evidence of a difference (p = 0.0019) between the first page (M = 22.63, SD = 8.64) and second page (M = 11.67, SD = 4.88) of ESG-type items (Pages 2 and 4, respectively), and essentially no evidence of a difference (p = 0.8575) between the first page of Likert-type items (M = 13.02, SD = 16.13) and the second page of ESG-type items (Pages 1 and 4, respectively). The results are shown in both Table 5 and Figure 6.

Some of the decrease in average response time per item on Page 4 can likely be attributed to familiarity with the instructions because the average response time per item for the third page of ESG-type items (Page 5) was higher (M = 19.09, SD = 8.27); Pages 2 and 4 had identical sets of instructions, while Page 5 had new instructions. Note that the average time per item may be artificially high for Page 5 because the time respondents spent on the page included time spent answering items and time spent reading the instructions; because there were only three items on Page 5, the time spent reading the instructions is not accounted for among as many items as other pages. While the average time per item was slower for the ESG-type items than for the Likert-type items, continued exposure to the items may result in students becoming more familiar with them and thus able to respond more quickly. If the ESG-type items provide more nuanced information about neutral-type responses, then the additional burden of asking participants to respond to these items may be justifiable.

A key part of this study was the presentation of 19 stems with both a Likert-type and an ESG-type response format. Because the stems were the same, for each pair of Likert-type and ESG-type items, correlations with a moderate-to-strong magnitude were anticipated as the items should be measuring quite similar constructs. As shown in Table 7 and Figure B1, almost all of the item pairs (18 of 19) had correlations of at least 0.40 and four and seven of the items had correlations of at least 0.70 for values of $b = -0.5$ and $b = -1$, respectively. This provides evidence that students were responding to the ESG-type items in ways that were consistent with their responses to the Likert-type items. Moreover, the internal consistency of the scales composed of Likert-type items and scales composed of ESG-type items was similar based on the values of reliability indices, consistent with prior studies of ESG-type items (e.g., Audrezet et al., 2016). Taken together, these results

provide evidence that students were able to understand the ESG-type items and respond to them in ways that indicate attitudes that are similar to the attitudes indicated by their responses to Likert-type items.

Further evidence that students were able to use the ESG-type items to respond as anticipated comes from the analysis of the three items written for this study. These items were designed so that the first item should elicit a more negative response and the third item should elicit a more positive response; the second item was designed to elicit a response that was mixed. These items were not intended to provide information about the distinction between the Indifferent and Ambivalent responses but focused instead on extreme situations. Evidence was presented that these items tended to elicit the anticipated responses based on both aggregate results (see Figure 7) and individual responses (see Appendix B). This provides evidence that students were able to respond to the items as intended.

Indifferent and Ambivalent Responses

Likert-type items are ubiquitous, and the use of ESG-type items is motivated by the two categories of Neutral responses that are possible: Indifferent (low positive and low negative attitudes) and Ambivalent (high positive and high negative attitudes). The appropriateness of these two neutral categories is predicated on the lack of a reciprocal relationship between positive and negative attitudes: this is a key assumption that may be met for some constructs and not met for others. The choice to study ESG-type items in the context of the EVT Cost construct was made because of noted historical measurement difficulties (Whitaker et al., in press; Flake, 2012; Flake et al., 2015) which could perhaps be attributed to a lack of reciprocal relationship. How participants would respond to ESG-type items if a reciprocal relationship between positivity and negativity did exist for a construct is not known. However, distinguishing between indifferent and ambivalent attitudes is of particular interest in education because students may not have formed attitudes prior to taking a course.

In this study, we assumed that, if a reciprocal relationship did exist for the construct, participants would tend to respond along the diagonal and the proportion responding in ways seemingly indicative of indifferent attitudes would be essentially the same as the proportion responding in ways seemingly indicative of ambivalent attitudes. Under this assumption, substantially different proportions responding either off the diagonal or in ways that would be classified as either Indifferent or Ambivalent would be evidence that there is not a reciprocal relationship and that ESG-type items could be appropriate for measuring the construct. However, the initial assumption may be incorrect, and further research into its appropriateness is needed.

In this study, three trinomial tests were conducted to assess whether there appears to be a reciprocal relationship between positivity and negativity for the EVT Cost construct. For the two trinomial tests that compared the proportions of responses above and below the part of the grid associated with neutral responses, there was evidence against a reciprocal relationship for the Task Effort and Overall Cost scales, but not for Emotional Cost (see Tables 8 and 9). However, when applying the trinomial test to assess the specific Indifferent and Ambivalent categories that are possible with ESG-type items, there was evidence supporting a difference in proportions

for all three scales: Emotional Cost (weak evidence), Task Effort Cost (strong evidence), Overall Cost (overwhelming evidence); see Table 10. While the effects of multiple comparisons should be considered, these trinomial test results – viewed through the lens of the preliminary nature of this work and relatively small sample size – suggest that further research using ESG-type items to investigate the EVT Cost construct would be appropriate.

Conclusion

This study provides initial evidence supporting the use of ESG-type items to measure the EVT Cost construct in the context of statistics education. Because of the preliminary nature of this pilot study, definitive conclusions about the appropriateness of ESG-type items cannot be reached, but the results provide evidence that students can understand ESG-type items and that the Indifferent and Ambivalent classifications afforded by ESG-type items are appropriate for the EVT Cost construct. However, practical questions about best practices for writing or adapting items and selecting constructs appropriate for ESG-type items remain.

Beyond the small nature of the pilot study, there were several limitations to this study due to LimeSurvey software used. Constructing an ESG-type item in LimeSurvey required adapting an existing item type by using question validation logic which resulted in the labels for the positive axis being placed along the top of the grid rather than the bottom; to account for this, the order of the negative labels was reversed. While this decision seemed reasonable because of the similarity of this conceptualization to the Cartesian coordinate system, additional data to support the appropriateness of this should be collected. Moreover, one participant noted in their response to the free response item that the ESG-type items were presented improperly when completing the survey using a smartphone which suggests that technical limitations may be a barrier to widespread use of these items. Regardless, further research and practical efforts may result in ESG-type items being used in some contexts to reveal nuanced information about neutral attitudes.

Acknowledgements

This research was supported by a grant from Mount Saint Vincent University.

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. Retrieved from https://www.testingstandards.net/open-access-files.html

Audrezet, A. (2014). *L'ambivalence des consommateurs: Proposition d'un nouvel outil de mesure* (Thèse de doctorat, Université Paris-Dauphine). Université Paris-Dauphine, Paris. Retrieved from https://tel.archives-ouvertes.fr/tel-01084068/document

Audrezet, A., Olsen, S. O., & Tudoran, A. A. (2016). The GRID scale: A new tool for measuring service mixed satisfaction. *Journal of Services Marketing*, *30*(1), 29–47. https://doi.org/10.1108/JSM-01-2015-0060

Audrezet, A., & Parguel, B. (2018). Using the Evaluative Space Grid to better capture manifest ambivalence in customer satisfaction surveys. *Journal of Retailing and Consumer Services*, *43*, 285–295. https://doi.org/10.1016/j.jretconser.2018.04.008

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

Bian, G., McAleer, M., & Wong, W.-K. (2011). A trinomial test for paired data when there are many ties. *Mathematics and Computers in Simulation*, *81*(6), 1153–1160. https://doi.org/10.1016/j.matcom.2010.11.002

Borriello, A. (2017). *The role of attitudes in determining individual behavior in transportation—From psychology to discrete choice modeling* (Dissertation, Università della Svizzera Italiana). Università della Svizzera Italiana, Lugano, Switzerland. Retrieved from http://doc.rero.ch/record/306643/files/2018ECO001.pdf

Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin*, *115*(3), 401–423. https://doi.org/10.1037/0033-2909.115.3.401

Cacioppo, J. T., Berntson, G. G., Norris, C. J., & Gollan, J. K. (2012). The Evaluative Space Model. In P. A. M. V. Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology: Volume One* (pp. 50–72). Los Angeles: SAGE.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond Bipolar Conceptualizations and Measures: The Case of Attitudes and Evaluative Space. *Personality and Social Psychology Review*, *1*(1), 3–25. https://doi.org/10.1207/s15327957pspr0101_2

Chew, P. K. H., Dillon, D. B., & Swinbourne, A. L. (2018). An examination of the internal consistency and structure of the Statistical Anxiety Rating Scale (STARS). *PLOS ONE*, *13*(3), e0194195. https://doi.org/10.1371/journal.pone.0194195

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed). New York: Wiley.

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

Cruise, R. J., Cash, R. W., & Bolton, D. L. (1985). Development and validation of an instrument to measure statistical anxiety. *Proceedings of the Section on Statistical Education, American Statistical Association*, 92–98.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (Fourth edition). Los Angeles: SAGE.

Eccles, J. S., & Wigfield, A. (2002). Motivational Beliefs, Values, and Goals. *Annual Review of Psychology*, *53*, 109–132. Retrieved from http://outreach.mines.edu/cont_ed/Eng-Edu/eccles.pdf

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, *61*, 101859. https://doi.org/10.1016/j.cedpsych.2020.101859

Eccles (Parsons), J., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives: Psychological and sociological approaches.* San Francisco: W.H. Freeman. Retrieved from http://web.archive.org/web/20170701031033/http://www.rcgd.isr.umich.edu/garp/articles/ecclesparsons83b.pdf

Flake, J. K. (2012). *Measuring cost: The forgotten component of expectancy-value theory* (Doctoral Dissertation, James Madison University). James Madison University, Harrisonburg, Virginia. Retrieved from http://commons.lib.jmu.edu/cgi/viewcontent.cgi?article=1221&context=master201019

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. https://doi.org/10.1016/j.cedpsych.2015.03.002

Gal, I., & Ginsburg, L. (1994). The role of beliefs and attitudes in learning statistics: Towards an assessment framework. *Journal of Statistics Education*, *2*(2), 1–15. Retrieved from https://ww2.amstat.org/publications/jse/v2n2/gal.html

Ganesalingam, S. (1994). A Nonparametric Test Procedure for Paired Data Sampling, in the Presence of Considerable Number of Ties. *Proceedings of the 30th Annual Meeting of the Operational Research Society of New Zealand and 45th Annual Conference of the New Zealand Statistical Association*, 384–389. Palmerston North, New Zealand. Retrieved from

http://www.thebookshelf.auckland.ac.nz/docs/NZOperationalResearch/conferenceproceedings/1994-proceedings/ORSNZ-proceedings-1994-69.pdf

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. https://doi.org/10.1007/BF02288892

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.

Irwing, P., & Hughes, D. J. (2018). Test Development. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale, and test development* (First Edition, pp. 3–48). Hoboken: Wiley.

Jarek, S. (2012). mvnormtest: Normality test for multivariate variables (Version R package version 0.1-9). Retrieved from https://CRAN.R-project.org/package=mvnormtest

Jiang, Y. (2015). *The Role of Perceived Cost in Students' Academic Motivation and Achievement* (Korea University). Korea University. Retrieved from http://dcollection.korea.ac.kr/public_resource/pdf/000000060090_20191030101628.pdf

Jiang, Y., Rosenzweig, E. Q., & Gaspard, H. (2018). An expectancy-value-cost approach in predicting adolescent students' academic motivation and achievement. *Contemporary Educational Psychology*, *54*, 139–152. https://doi.org/10.1016/j.cedpsych.2018.06.005

Kolde, R. (2019). pheatmap: Pretty Heatmaps (Version R package version 1.0.12). Retrieved from https://CRAN.R-project.org/package=pheatmap

Larsen, J. T., Norris, C. J., McGraw, A. P., Hawkley, L. C., & Cacioppo, J. T. (2009). The evaluative space grid: A single-item measure of positivity and negativity. *Cognition & Emotion*, *23*(3), 453–480. https://doi.org/10.1080/02699930801994054

Likert, R. (1933/1967). The method of constructing an attitude scale. In M. Fishbein (Ed.), *Readings in attitude theory and measurement* (pp. 90–95). New York: Wiley. (Original work published 1932)

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.

Pearl, D. K., Garfield, J. B., delMas, R., Groth, R. E., Kaplan, J. J., McGowan, H., & Lee, H. S. (2012). *Connecting Research to Practice in a Culture of Assessment for Introductory College-level Statistics*. Retrieved from https://www.causeweb.org/cause/archive/research/guidelines/ResearchReport_2012.pdf

Pedersen, T. L. (2021). ggforce: Accelerating "ggplot2" (Version R package version 0.3.3). Retrieved from https://CRAN.R-project.org/package=ggforce

Pohlert, T. (2021). PMCMRplus: Calculate Pairwise Multiple Comparisons of Mean Rank Sums Extended (Version R package version 1.9.0). Retrieved from https://CRAN.R-project.org/package=PMCMRplus

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Ramirez, C., Schau, C., & Emmioğlu, E. (2012). The Importance of Attitudes in Statistics Education. *Statistics Education Research Journal*, *11*(2), 57–71. Retrieved from https://iase-web.org/documents/SERJ/SERJ11(2)_Ramirez.pdf

Revelle, W. (2021). psych: Procedures for Personality and Psychological Research (Version 2.1.3). Evanston, IL: Northwestern University. Retrieved from https://cran.r-project.org/web/packages/psych/index.html

Roberts, D. M., & Bilderback, E. W. (1980). Reliability and Validity of a Statistics Attitude Survey. *Educational and Psychological Measurement*, *40*(1), 235–238. https://doi.org/10.1177/001316448004000138

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

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

Schmitt, N. (1996). Uses and Abuses of Coefficient Alpha. *Psychological Assessment*, *8*(4), 350–353. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.615.4053&rep=rep1&type=pdf

Thibaut, J. W., & Kelley, H. H. (1959). *The social psychology of groups*. New York: Wiley. Retrieved from https://archive.org/details/socialpsychology00thib

Uebersax, J. S. (2006). Likert scales: Dispelling the confusion. Retrieved from Statistical Methods for Rater Agreement website: https://john-uebersax.com/stat/likert.htm

Unfried, A., Kerby, A., & Coffin, S. (2018). Developing a Student Survey of Motivational Attitudes Toward Statistics. *2018 JSM Proceedings*. Presented at the 2018 Joint Statistical Meetings, Vancouver, Canada.

Retrieved from http://sdsattitudes.com/wp/wp-content/uploads/2021/08/Unfried-et-al_2018_Developing-a-Student-Survey-of-Motivational-Attitudes-Toward-Statistics.pdf

Unfried, A., Posner, M., Bond, M., Kerby-Helm, A., Bolon, W., Whitaker, D., & Batakci, L. (2021, August). *Why Do We Need Yet ANOTHER Instrument Measuring Student Attitudes?* Presented at the 2021 Joint Statistics Meetings, Virtual. Retrieved from http://sdsattitudes.com/wp/jsm2021/

Wei, T., & Simko, V. (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Retrieved from https://github.com/taiyun/corrplot

Whitaker, D. (2021, June). *Developing and revising the Student Survey of Motivational Attitudes Toward Statistics: Results from a pilot study*. Presented at the Statistics Society of Canada 2021 Annual Meeting, Virtual.

Whitaker, D., Drew, B., & Barss, J. (2021) *GridItemTools: Grid item tools.* R package version 0.0.12. https://github.com/douglaswhitaker/GridItemTools

Whitaker, D. & Hebert, N. (2021). *MVQuickGraphs: Quick multivariate graphs*. R package version 0.1.6. https://github.com/douglaswhitaker/MVQuickGraphs

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). New York, NY: Routledge.

Whitaker, D., Unfried, A., & Bond, M.E. (in press). Challenges associated with measuring attitudes using the SATS family of instruments. *Statistics Education Research Journal*.

Whitaker, D., & White, A. (2020, May). *Measuring Statistics Attitudes and Anxieties*. Poster presented at The Fifth Biennial Electronic Conference on Teaching Statistics (eCOTS). Retrieved from https://www.causeweb.org/cause/ecots/ecots20/posters/2-05

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. New York: 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

Wigfield, A., & Eccles, J. S. (2020). 35 years of research on students' subjective task values and motivation: A look back and a look forward. In *Advances in Motivation Science* (Vol. 7, pp. 161–198). Elsevier. https://doi.org/10.1016/bs.adms.2019.05.002

Wigfield, A., Rosenzweig, E. Q., & Eccles, J. S. (2017). Achievement Values: Interactions, Interventions, and Future Directions. In A. J. Elliot, C. S. Dweck, & D. S. Yeager (Eds.), *Handbook of competence and motivation: Theory and application* (Second edition, pp. 116–134). New York: Guilford Press.

Wise, S. L. (1985). The development and validation of a scale measuring attitudes toward statistics. *Educational and Psychological Measurement*, *45*(2), 401–405. Retrieved from http://epm.sagepub.com/content/45/2/401.short

Appendix A

The items (see Table A1) and instructions that comprised the survey instrument are below.

Table A1. The text of the items, the scale they are from, the page of the survey, the order on the survey, and response options. The Emotional Cost and Task Effort Cost scales were developed by Flake et al. (2015), and the Overall Cost scale was developed for another study (Unfried et al., 2018, 2021; Whitaker, 2021; Whitaker et al., 2019).

Pilot Order	Item	Scale	Likert-type Page	ESG-type Page	Response Options
1	This class is emotionally draining.	Emotional Cost	1	4	Likert; ESG Agreement/Disagreement
2*	Acquiring statistical skills is worth the effort.	Overall Cost	1	4	Likert; ESG Agreement/Disagreement
3	I prioritize other tasks over statistics	Overall Cost	1	4	Likert; ESG Agreement/Disagreement
4	This class is too much work.	Task Effort Cost	1	4	Likert; ESG Agreement/Disagreement
5	I have more important things to do than spending time learning statistics.	Overall Cost	1	4	Likert; ESG Agreement/Disagreement
6	Taking statistics will limit my future prospects (for example, lower my GPA).	Overall Cost	1	4	Likert; ESG Agreement/Disagreement
7*	Learning statistics is worth spending money on.	Overall Cost	1	4	Likert; ESG Agreement/Disagreement
8	This class takes up too much time.	Task Effort Cost	1	4	Likert; ESG Agreement/Disagreement
9	This class requires too much effort	Task Effort Cost	1	4	Likert; ESG Agreement/Disagreement
10	This class is too stressful.	Emotional Cost	3	2	Likert; ESG Agreement/Disagreement
11	I worry too much about this class.	Emotional Cost	3	2	Likert; ESG Agreement/Disagreement

12	This class is too exhausting.	Emotional Cost	3	2	Likert; ESG Agreement/Disagreement
13	I avoid working on statistics because it makes me feel bad	Overall Cost	3	2	Likert; ESG Agreement/Disagreement
14	This class demands too much of my time.	Task Effort Cost	3	2	Likert; ESG Agreement/Disagreement
15*	If I had to take another course, I would choose a statistics course.	Overall Cost	3	2	Likert; ESG Agreement/Disagreement
16	I have to put too much energy into this class.	Task Effort Cost	3	2	Likert; ESG Agreement/Disagreement
17	This class is too frustrating.	Emotional Cost	3	2	Likert; ESG Agreement/Disagreement
18*	Learning statistics is a good use of my time.	Overall Cost	3	2	Likert; ESG Agreement/Disagreement
19	This class makes me feel too anxious	Emotional Cost	3	2	Likert; ESG Agreement/Disagreement
20	I worked on my homework for 20 hours, and I didn't really understand it.			5	ESG Positive/Negative
21	I worked on my homework for 20 hours, and I think I understood most of it.			5	ESG Positive/Negative
22	I worked on my homework for 20 hours, and I understood all of it and learned some new things.			5	ESG Positive/Negative
23	Which of the following best describes your understanding of the grid item?				Multiple Choice
24	Were the examples for the grid response items helpful?				Multiple Choice
25	Did you find the grid item allowed you to give a response that better reflected your reaction to each item compared to the traditional 1-9 Disagree to Agree scale?				Multiple Choice

Note: * indicates the item should be reverse-coded.

The instructions for each page follow.

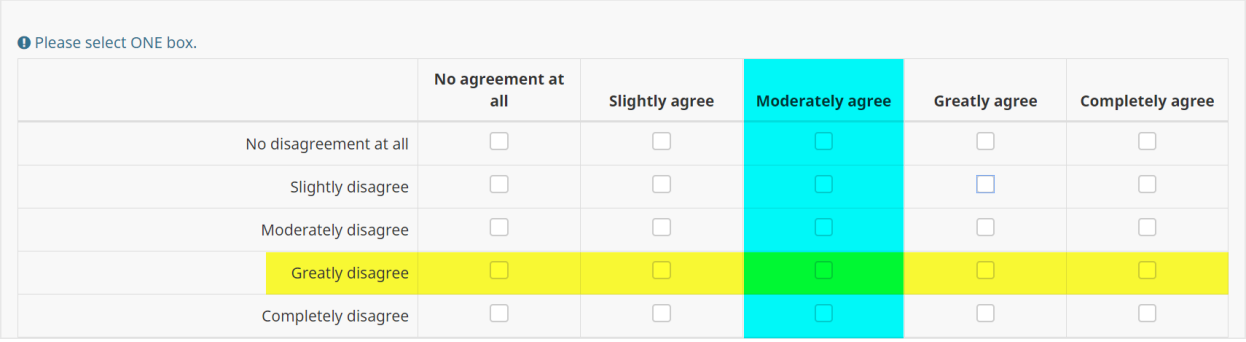
[Pages 1 and 3]

Instructions: Think about your experience in MATH 2209. For each of the following statements, please select the one response that best represents the degree to which you agree or disagree with the statement. Try not to think too much about each response. Please choose the appropriate response for each item:

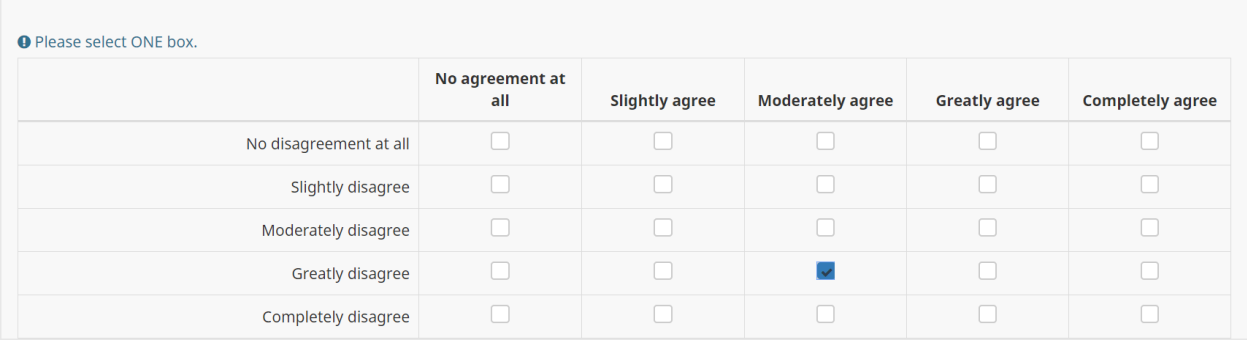
[Pages 2 and 4]

Instructions: These items all use a grid to record your responses. Using the grid, you will be able to record the extent to which you agree and the extent to which you disagree with each statement. Please select the ONE box that best describes your OVERALL feeling about each statement.

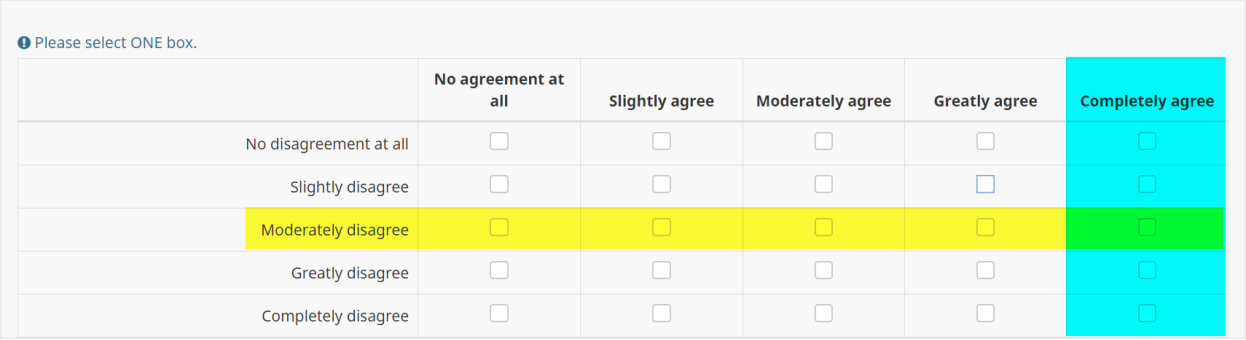
Example 1: Taylor is responding to the item "I like eating kale." Taylor really dislikes the taste of kale, but also knows that kale has a lot of nutrients. Taylor chooses the box that corresponds to "Greatly disagree" (because Taylor does not like the taste) and "Moderately agree" (because Taylor appreciates the nutritional value of kale).



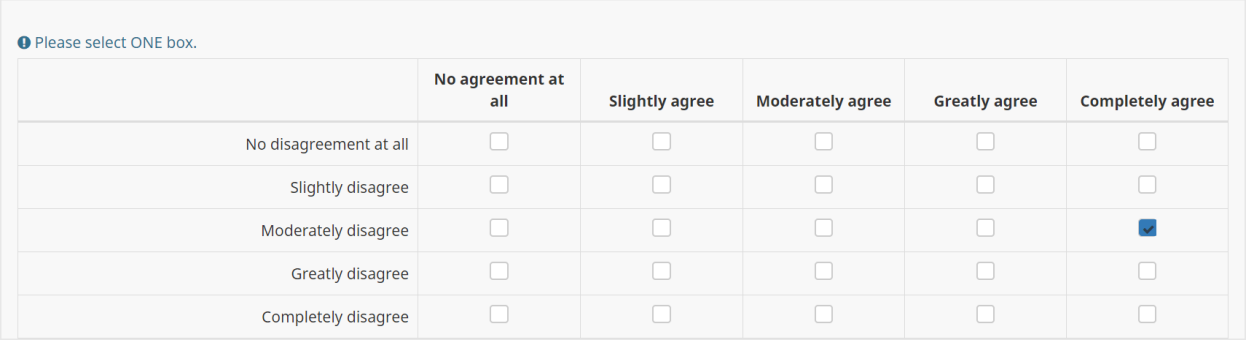
The highlighting represents Taylor's thinking – there will be no highlighting on the survey. Taylor selects the box in the corresponding row and column.



Example 2: Drew is responding to the item "I dislike driving in Halifax." Drew really hates the traffic during rush hour, but also finds driving to be more convenient than other transportation options. Drew chooses the box that corresponds to "Completely agree" (because Drew hates driving in heavy traffic) and "Moderately disagree" (because Drew appreciates the convenience driving).



The highlighting represents Drew's thinking – there will be no highlighting on the survey. Drew selects the box in the corresponding row and column.



[Page 5]

Now imagine that you have been working on a long statistics assignment for homework. For each of the statements below, use the grid to indicate how POSITIVE and how NEGATIVE you would feel.

[Page 6]

Now please think about your experience taking this survey.

[At the end of pages 1-5]

The survey continues on the next page. Note that you may stop participating at any time or skip items, particularly if you feel uncomfortable or experience discomfort while taking this survey.

Appendix B

This appendix contains additional graphs that provide more granular information about the results from ESG-type items.

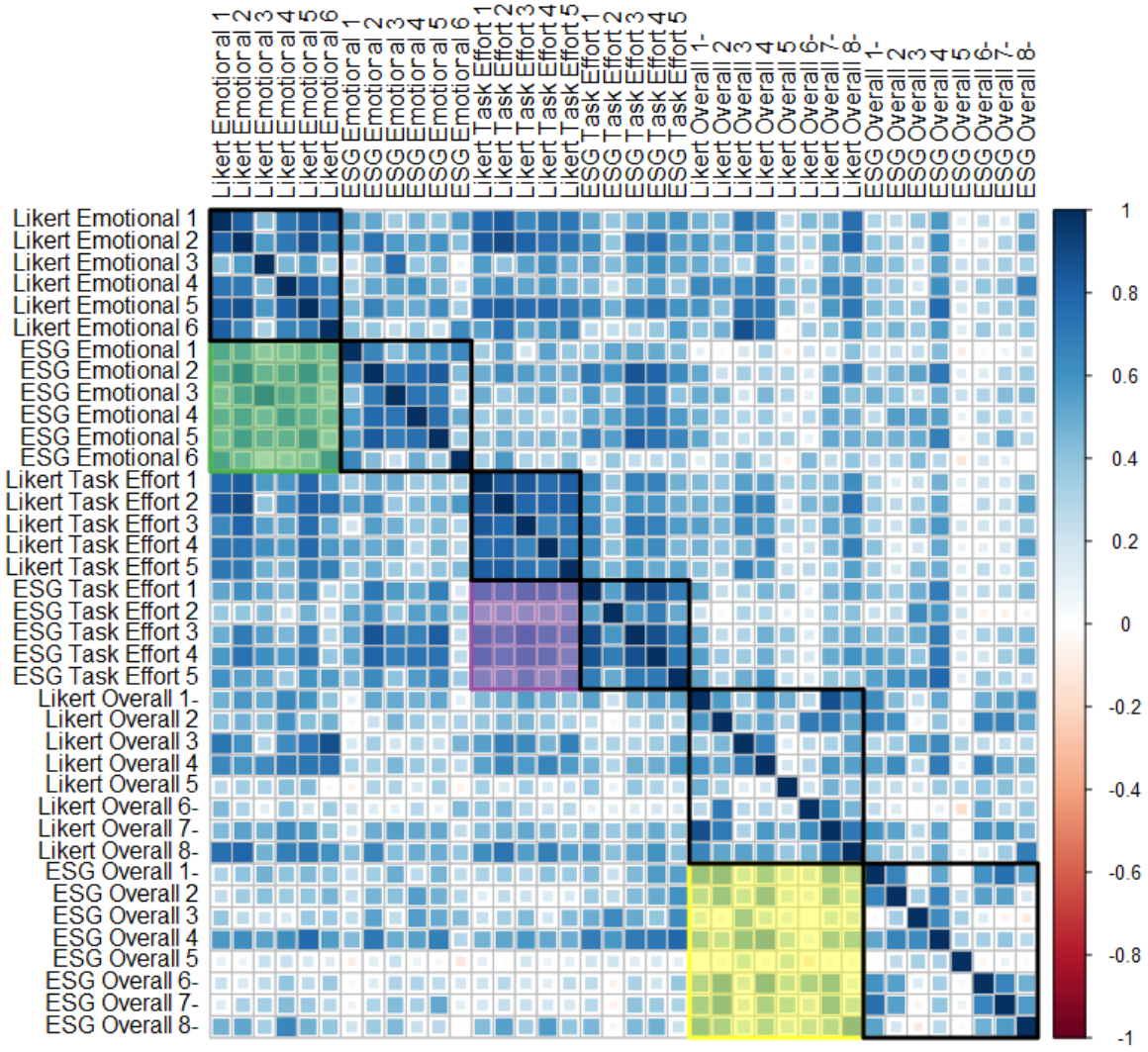


Figure B1. Pearson correlations for the 19 Likert-type items and 19 ESG-type items; the size of the square within each cell indicates the magnitude of the correlation. A "-" in the item name indicates that the Likert-type item is reverse-coded. Three transparent coloured squares in the lower region are used to draw attention to correlations among items from the same scale presented using the two types of items. Black squares are used to demarcate the items on a single scale.

Individuals' responses are shown in Figures B2 and B3. Individuals' changes in responses for the items are shown in Figures B2 and B3 (cf. Figure 7, which shows aggregated responses). Figure B2 shows the cell chosen by each individual in the first and second items (Patterned ESG Items 1 and 2) with a line, and it also shows the cell chosen

by each individual in the second and third items (Patterned ESG Items 2 and 3) with a line. In both graphs shown in Figure B2 there appears to be a tendency for individuals' responses to become more positive. To emphasize the overall change, Figure B3 shows the change from the cell chosen by the individual in Patterned ESG Item 1 to the cell chosen in Patterned ESG Item 3 (i.e., the items with the expected most negative and most positive responses, omitting the item for which a mixed response was expected).

Figure B2. Graphs showing the change in individuals' responses from the first to second (left) and second to third (right) ESG-type items created for this study. The lines are narrowest for the first question (left) and second question (right) and widest for the second question (left) and third question (right). The colour of the lines indicates the individual.

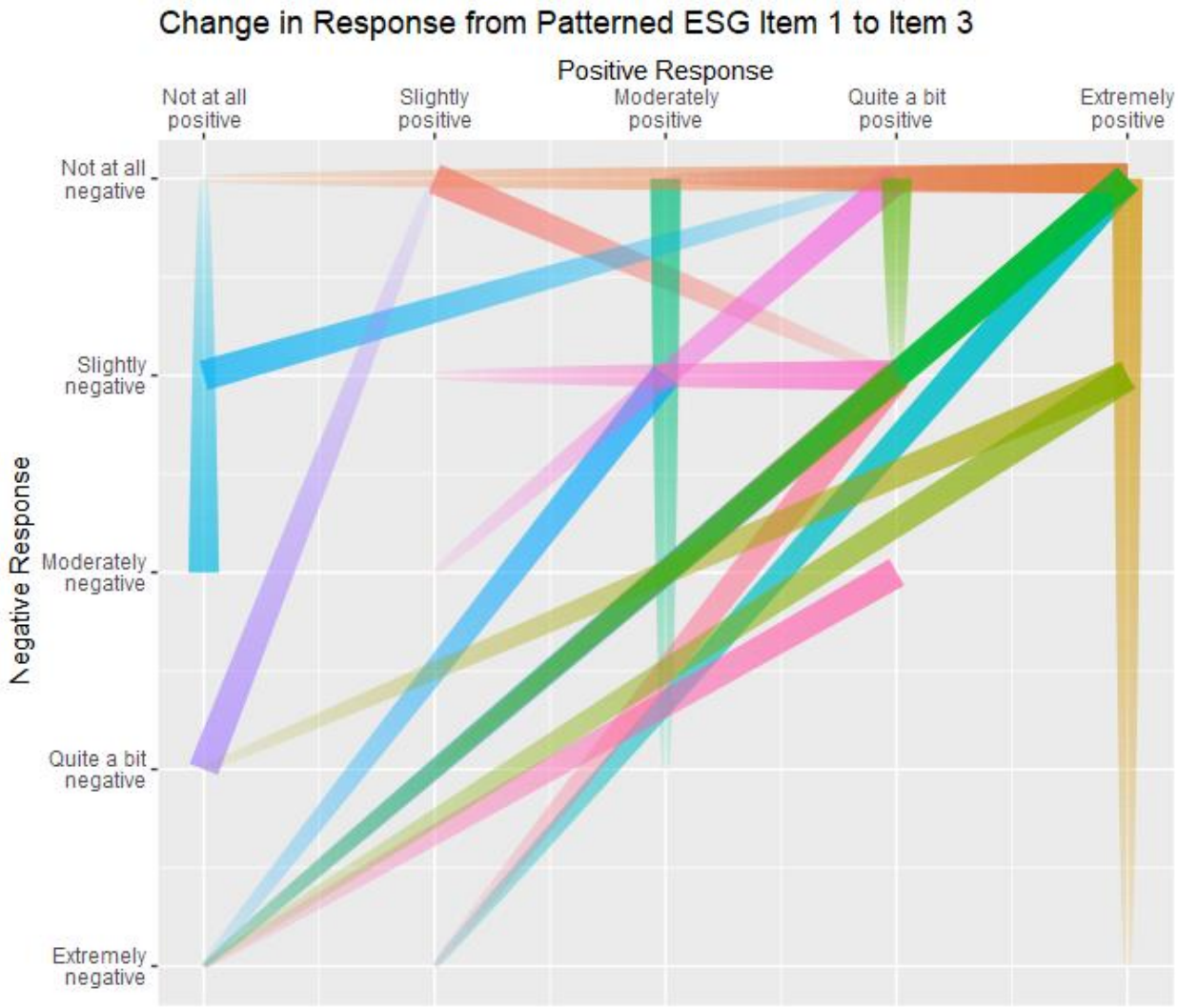


Figure B3. A graph showing the change in individuals' responses from the first to third ESG-type items created for this study. The lines are narrowest for the first question and widest for the third question. The colour of the lines indicates the individual.

While interpreting Figure B2 is more complicated due to the second item eliciting mixed responses, Figure B3 clearly shows a tendency for students to provide a more positive and less negative response to the third item than the first item because of the overall tendency of the lines to go from bottom left to top right.

Corresponding Author Contact Information:

Author name: Douglas Whitaker

Department: Department of Mathematics and Statistics

Faculty: Faculty of Science

University, Country: Mount Saint Vincent University, Canada

Email: douglas.whitaker@msvu.ca

Please Cite: Whitaker, D., & Barss, J., Drew, B. (2022). Measuring Opportunity Cost in Statistics Using Evaluative Space Grid Items: Results from a Pilot Study. *Journal of Research in Science, Mathematics and Technology Education*, 5(*1*), 1-36. DOI: https://doi.org/10.31756/jrsmte.511

Received: 14 July 2021 ▪ Accepted: 16 September 2021