An Examination of Distributional Characteristics & Psychometric Properties of Rating Scales Under Different Scale Label Configurations

Alfonso J. Martinez

University of Iowa

Author Note

Alfonso J. Martinez, Department of Psychological and Quantitative Foundations, University of Iowa. https://orcid.org/0000-0002-5337-9654

Correspondence regarding this article should be addressed to Alfonso J. Martinez, Department of Psychological and Quantitative Foundations, University of Iowa, 224B Lindquist Center, 240 S Madison St., Iowa City, IA 52242, United States of America. Email: alfonso-martinez@uiowa.edu

Abstract

Rating scales (e.g., Likert scales) have become an indispensable tool in the social sciences for measuring latent constructs. Despite their practicality, it is well known that rating scales are sensitive to design choices made during the scale construction process. In the present study, we investigate the effects of different scale labeling configurations on the distributional characteristics and psychometric properties of rating scales by analyzing data from two 7-point rating scales that were altered with respect to their scale labels. Specifically, we investigate the differences in composite score means, variances, reliability, factor structure, and measurement invariance (i.e., configural, metric, and scalar invariance) of two scales that measure divergent constructs (life satisfaction life; academic time management & procrastination). Results indicate that distributional characteristics of rating scales are affected by differences in labeling configurations to varying degrees, however the psychometric properties of the scales are not influenced by changes in labeling. In particular, it was found that scalar invariance held indicating the scales, although different with respect to scale labels, have similar factor structures, discrimination parameters, and latent intercept parameters, irrespective of labeling configuration. Implications and recommendations for scale developers and directions for future research are discussed.

Keywords: Rating scales, surveys, scale labels, scale design, factor analysis, measurement invariance

An Examination of Distributional Characteristics & Psychometric Properties of Rating Scales Under Different Scale Label Configurations

Rating scales (RS) have become an indispensable tool in the social sciences for measuring latent, or unobserved, constructs (DeCastellarnau, 2018; Krosnick, Judd, & Wittenbrink, 2018; Likert, 1932). The RS format is preferred over alternative methods (e.g., multidimensional forced-choice format; Wetzel, Frick, & Greiff, 2020) because they perform well in a variety of settings and are straightforward to administer, score, and are cost-effective relative to alternative approaches (Simms, 2008). Combined with a rigorous construction process – which includes defining the target construct on the basis of theoretical and empirical research, item writing and selection procedures, and a plethora of validation techniques – the RS format has become the gold standard for measuring latent constructs.

Despite the popularity and practicality of RS, it is widely recognized that they are sensitive to choices made during the scale construction process (Wetzel & Greiff, 2018). In particular, the realization that RS are a function of auxiliary components – scale design characteristics that may not be directly related to the target construct – has catalyzed research evaluating these components and their effect on response data and measurement quality. Previous efforts have focused on scale design characteristics such as response-order effects (e.g., Betts & Hartley, 2012; Galesic, Tourangeau, Couper, & Conrad, 2008; Krosnick & Alwin, 1987), content polarization (e.g., Lam & Stevens, 1994), number of response categories that optimize scale reliability and validity (Churchill & Peter, 1984; Lee & Paek, 2014; Leung, 2011; Lozano, García-Cueto, & Muñiz, 2008; Preston & Colman, 2000; Spector, 1976), effects of including or excluding a middle response option (Armstrong, 1987; Chyung, Roberts, Swanson, & Hankinson, 2017; Garland, 1991; Hernández, Drasgow, & González-Romá, 2004; Nadler, Weston, & Voyles, 2015; Presser & Schuman, 1980), and scale format effects (e.g., Weijters, Cabooter, & Schillewaert, 2010; Cabooter, Weijters, Geuens, & Vermeir, 2016). Research in these areas has lead to a better understanding of the influence scale characteristics have on measurement quality. Overall, the research unequivocally supports the notion that scale format choices matter (see DeCastellarnau (2018) for a review).

Despite decades of research investigating the effects of scale characteristics, there still remain design characteristics that have received relatively little attention in the extant literature. In particular, there is little research that has examined the effects of differential scale labels on the measurement quality of RS. As explicated in the next section, the current literature in this area is inconclusive and often conflicting. This article aims to contribute to the existing literature by investigating the effects of differential scale labels on the distributional and psychometric properties of RS, particularly among fully-labeled and partially-labeled RS. Current practices generally consider scale labeling an afterthought in the scale construction process (Eutsler & Lang, 2015; Simms, 2008; Wetzel & Greiff, 2018), presumably because labels are assumed to not have any major influence on responses. Given that the scale design choices have been shown to influence measurement quality however, it is of practical importance to investigate if scale labels are a design characteristic that influence response processes and data quality and the extent to which they do so. When integrated with the larger scale development literature, this research can be used to inform future scale development practices.

We have organized the remainder of this paper as follows. The following section presents a response process framework developed by cognitive psychologists and survey methodologists that describes a mechanism by which individual's response behaviors are shaped. This provides the theoretical foundations for understanding how and why responses are affected by RS design characteristics. We then provide a review of the relevant literature on the effects of labeling schemes on the quality of response data, followed by a description of our research questions, methods, and analyses. Finally, we present our findings and discuss the implications of the research, and conclude by providing recommendations for scale developers.

Theoretical Background on Response Processes

Response process models have been developed in the survey development literature in an attempt to describe the cognitive processes that underlie responses to RS items. Here we describe the framework proposed by Krosnick et al. (2018). According to their model, respondents undergo an evaluation process consisting of three phases that ultimately influence the response behaviors that are displayed on the survey: (1) automatic activation, (2) deliberation, and (3) response.

In the automatic activation phase, the object of evaluation (i.e., the RS item) produces an initial judgement. This judgement often occurs automatically and spontaneously, without the conscious awareness of the evaluator (i.e., the respondent). Previous research has found that these processes utilize limited cognitive resources, do not require active memory recall, are the result of repetitive and frequent experiences with the evaluation object, and occur within a few milliseconds after encountering the object (Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Herring et al., 2013; Klauer, Rossnagel, & Musch, 1997; Krosnick et al., 2018; Shiffrin & Schneider, 1977). The strength of the initial judgement varies from respondent to respondent depending on the contents of the item, with some respondents producing initial judgements with enough strength to induce a final response and other respondents producing relatively impartial initial judgements.

In the deliberation phase, there is a transition from automatic evaluation to a conscious retrieval of relevant information and associations. In this phase, the respondent consciously draws on information from a variety of sources, including stored evaluations from previous experiences, relevant associations, and current situational factors. Motivation and opportunity are key prerequisites for successful information retrieval because they, in part, determine the amount and quality of information the individual retrieves. For example, an item that elicits information that is not readily available to the individual (e.g., an item whose content is unfamiliar to the respondent) may limit the respondent's ability to appropriately draw on the information needed to provide a thoughtful response.

The final phase is the response phase, in which the individual integrates the information obtained from the previous steps and maps their final evaluation onto the provided response alternatives (if the item has a predetermined response format). Although this step appears trivial, it is crucial that the response options provided sufficiently align with the individual's final evaluation. If the response format is flexible enough to accommodate unstructured responses (e.g., open-ended response items), the individual should not have much trouble providing an accurate response. However, if the response format is restricted to a few predefined response alternatives, which is often the case in the RS format, the individual may have difficulty aligning their true response with the available alternatives, especially if the provided alternatives significantly deviate from the individual's true evaluation. In such a situation the respondent would be forced to provide an inaccurate or random response.

It is known that the response processes just described are influenced by characteristics of the survey or RS (DeCastellarnau, 2018; Tourangeau, Rips, & Rasinski, 2000). In the case of RS labeling formats, such as those that are the focus of this paper, response processes and survey-response interactions may be influenced by the amount of information provided by the scale labels. Because labels are designed to clarify the meaning of the scale points, it is hypothesized (assuming a fixed number of categories) that increasing the verbal labeling will increase the likelihood of a precise response. This notion is in agreement with Krosnick (1991), who postulated that ambiguous response alternatives to survey or RS items prompt the individual to compensate for the ambiguity by attaching their own interpretations and meanings to the ambiguous alternatives. For instance, respondents may elect to an endorse a particular (unlabeled) response category based on the cognitive mechanism previously described that drives them to select that particular response (e.g., looking for clues in the item stem or interpreting numerical labels, etc.). It is possible that the attributions made for that particular response category is different across respondents, i.e., respondents interpret the same response category differently due to their unique processing strategies. This subsequently leads to measurement error and a decrease in scale reliability and validity (Alwin & Krosnick, 1991). Thus, ambiguity in the response labels is expected to produce biased responses and suboptimal data quality (Alwin, 2007).

On the other hand, several scholars have argued that increasing verbal labeling increases the cognitive effort required by respondents by increasing the amount of information that must be processed at any given moment (Krosnick & Fabrigar, 1997; Krosnick & Presser, 2010; Kunz, 2015). To minimize their cognitive efforts, respondents are likely to develop cognitive shortcuts, such as selecting the first response they encounter, regardless of their true attitude with respect to the item's content. This behavior is known as satisficing (Krosnick, 1991; Krosnick et al., 2018) and it has been linked to less accurate survey responses, as well as inflated reliability and validity estimates (e.g., Hamby & Taylor, 2016; Keusch & Yang, 2018). Cognitive shortcuts and satisficing behaviors have been shown to be especially prevalent among individuals with lower levels of education and when respondent motivation is low (Krosnick, 1991). Thus, although explicit labeling of response categories reduces ambiguity in translating subjective responses to the provided response categories (Alwin, 2007), it comes at the potential cost of increased cognitive effort and satisficing behaviors.

Full-labeling vs. Partial-labeling of Rating Scales

Despite theoretical discourse regarding the effects of labeling format, the existing empirical literature in this area has produced conflicting results. Early studies by Huck and Jacko (1974) and Wyatt and Meyers (1987) found that endpoint-labeled scales (i.e., scales in which only the extreme response categories are explicitly defined) had lower total score means compared to their fully-labeled counterparts (i.e., scales in which all response categories are explicitly defined), however Newstead and Arnold (1989), Dixon, Bobo, and Stevick (1984), and Menold, Kaczmirek, Lenzner, and Neusar (2014) reported that group means were not affected by changes in labeling. Moreover, Eutsler and Lang (2015) found that fully-labeled scales display greater variability, have increased power, and less measurement error than partially-labeled scales, however these results are inconsistent with Chang (1997), who found that variance due to labeling configuration accounted for less than 1% of the total variation in observed scores. Wyatt and Meyers (1987) and Newstead and Arnold (1989) also reported that labeling format has a negligible influence on the variation of observed scores.

With respect to internal consistency reliability, Alwin (2007) found that fully-labeled rating scales provided the highest reliability, with an average reliability of approximately .72 while the reliability of partially-labeled scales was significantly lower with an average reliability of .51. Additional studies by Alwin and Krosnick (1991), Menold et al. (2014), Krosnick and Berent (1993), and Saris and Gallhofer (2014) have also found fully-labeled scales to be slightly more reliable, however these findings are in disagreement with Finn (1972) who found reliability to be unaffected by verbal labels and Andrews (1984) who found that fully-labeled scales exhibited lower reliability and below average data quality. Test-retest reliability has been found to be higher in fully-labeled scales, provided the scale items were highly related to the construct being measured (Weng, 2004).

More recently, the effects of labeling have been investigated with modern methodological techniques. Spratto, Leventhal, and Bandalos (2020) utilized item response tree models to model respondents' tendency to select extreme response categories in fully- and partially-labeled scales, while Moors, Kieruj, and Vermunt (2014) employed latent class analysis to uncover latent response characteristics. Spratto and Bandalos (2020) applied a mixed-methods approach where participants were recorded and interviewed as they were providing their responses in an attempt to capture their response processing in real time, and Menold et al. (2014) utilized eye-tracking technology to capture participants' eye movements as they interacted with a virtual rating scale. Their results found that fixation times – defined as time spent looking at the response scale area – were shorter when the scale was partially labeled, however the probability a given response category received attention was higher under this labeling scheme.

The Present Study

The effects of response labeling are still not well understood and the existing literature does not provide clear guidelines for scale developers. Given this background, the purpose of the present research is to investigate the effects of labeling configurations on the *distributional characteristics* as well as the *psychometric properties* of RS. Our rationale for distinguishing between the two is motivated by recognizing that previous research has predominantly focused on investigating how differential labeling affect distributional characteristics of response data (e.g., differences in composite score means). However, it is unknown if distributional differences are indicative of differences from a psychometric perspective. This is an important consideration given that RS are traditionally designed and utilized to measure latent constructs, thus it is of interest to evaluate the extent to which changes in labeling affect the measurement properties of RS (i.e., measurement of the target construct). Moreover, a psychometric perspective allows for a more nuanced approach for answering questions that cannot be answered by examining distributional differences or reliability indices alone, such as, "does labeling configuration X make items more or less difficult to endorse at the same level of the target construct than labeling configuration Y?" If differences from a psychometric perspective are found, then it is likely that labeling introduces a non-trivial amount of construct-irrelevant variance, bringing into question the quality and validity of the responses. A scale developer might find it permissible for the distributional characteristics to differ slightly as a result of labeling choices provided the psychometric information (measurement of the latent construct) remains the same.

To summarize, although different labeling configurations may alter distributional features of the response data, the extent to which such changes affect the psychometric attributes of the scale is unknown. To this end, we employ multiple group confirmatory factor analysis (MGCFA) to conduct invariance analyses and investigate these properties from a psychometric perspective.

Methods

Participants

Participants in the present study consisted of university students from several sections of an introductory psychology course who were recruited via an online subject pool recruitment system. Participants' ages ranged from 18 to 37 ($M = 19.16$, $SD = 1.62$). All participants were compensated for their participation by receiving course credit. A total of $N = 656$ subjects participated in the study, with approximately 77% ($N = 505$) of the sample identifying as female. The sample was ethnically and racially diverse, with 57% of the respondents identifying as Hispanic, 19% Asian-American, 14% White, 4% Black or African-American, and 6% indicating "other" (e.g., biracial).

Measures

We administered the following scales: Satisfaction with Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985); Academic Time Management and Procrastination Measure (ATMPM; Martinez, 2021; Won & Shirley, 2018); and the Marlowe-Crowne Social Desirability Scale (SDS; Crowne & Marlowe, 1960). These scales were selected for this study because they are published measures that have been used previously in research and practice[1], and have been shown to have favorable psychometric properties (see descriptions of measures below). Furthermore, the scales do not appear to overlap in the constructs they measure. This decision was intentional as the discordant nature of the selected scales,

[1] We note that we are not necessarily interested in making inferences or judgements about these particular scales, rather these scales were selected based on their scale design and popularity in the literature. We believe that utilizing previously established scales enhance the credibility of the results as opposed to developing a scale whose psychometric properties are unknown beforehand (see Greenleaf (1992) for a different perspective).

along with the diversity of factor structures, permit generalizability to other measures with similar scale structure.

Satisfaction with Life Scale

The SWLS is a 5-item scale designed to measure global cognitive judgments of satisfaction with one's life (Diener et al., 1985). The psychometric properties of the SWLS have been extensively investigated and the scale has been shown to exhibit strong internal consistency (e.g., Vassar, 2008), moderate temporal reliability (e.g., Pavot & Diener, 2009), as well as acceptable levels of predictive, convergent, and construct validity (e.g., Diener, Sandvik, Seidlitz, & Diener, 1993; Pavot & Diener, 2009; Pavot, Diener, Colvin, & Sandvik, 1991; Jang et al., 2017). In addition, the scale items have been shown to cover a wide range of the target domain (±2.5 standard deviations; Nima, Cloninger, Persson, Sikström, & Garcia, 2020). Scores on the SWLS range from 5 to 25, with higher scores reflecting greater overall satisfaction with life. The SWLS utilizes a 7-point response format with the following label configuration: "1 = Strongly Disagree," "2 = Disagree," "3 = Slightly Disagree," "4 = Neither Agree nor Disagree," "5 = Slightly Agree," "6 = Agree," and "7 = Strongly Agree."

Academic Time Management and Procrastination Measure

The ATMPM is a 14-item scale consisting of three subscales that quantify students' perception about their time management and level of procrastination as it pertains to educational outcomes (Won & Shirley, 2018). Five items measure a students' ability to plan their time (ATMPM-PT), four items assess a students' ability to monitor their time (ATMPM-MT), and five items measure behaviors related to academic procrastination (ATMPM-Pro). Scores on the ATMPM-PT, ATMPM-MT, and ATMPM-Pro range from 5 to 35, 4 to 28, and 5 to 35, respectively, with higher scores on ATMPM-PT and ATMPM-MT reflect greater efficacy with time management while higher scores on ATMPM-Pro reflect a greater tendency to engage in procrastination. Previous research has reported acceptable reliability estimates (coefficient ω) ranging from 0.84 to 0.92 (Martinez, 2021) and the scale has been found to function well with a variety of student subpopulations (e.g., first-generation and non-first generation college students; Martinez, 2021). The ATMPM utilizes the same 7-point response format and labeling configuration as the SWLS.

Marlowe-Crowne Social Desirability Scale

The SDS is a dichotomously-scored (True/False) 33-item scale that measures the extent to which a respondent appears to be responding in a socially desirable manner. In the present study, the SDS functioned as a distractor task and is not investigated further (see Procedure section).

Procedure

The online survey administration system Qualtrics® was used to administer the instruments. We manipulated labeling of the response categories in one of two ways: (1) labeling the extreme response categories only or (2) explicitly labeling all response categories. Two separate survey sets were created for this purpose; in the first survey set, the scales were reproduced verbatim with verbal labels, response categories, and other formatting and instructional details unaltered (i.e., they were reproduced to the authors' original specifications). The original scale labels in the SWLS and ATMPM consisted of the fully-labeled 7-point label configuration described above. In the second survey set, the intermediate response categories were removed and only response categories 1 and 7 were labeled. Figure 1 provides a visual depiction of the scale labels that were shown to respondents in each respective condition. Respondents who received the first survey set are hereafter referred to as the fully-labeled group/condition and respondents who received the modified survey set are hereafter referred to as the endpoints-only group/condition.

Survey administration was pseudorandom as respondents were required to select between one of two links that redirected them to a survey set. Both survey links were identical and adjacent to each other in the online system. Respondents were blind to the condition they were in and once they completed a survey set they were unable to access the other link. The sample sizes of the two conditions were comparable with 352 respondents (54%) in the fully-labeled condition and 304 respondents (46%) in the endpoints-only condition. Table 1 provides information regarding median time to completion between the two groups, as well as general demographic information about the sample. There were no significant differences with respect to any demographic variable between the two groups, providing evidence the quasi-randomization procedure was as good as random.

After accessing the survey set, and consenting to participate, respondents were instructed to answer each item to the best of their ability. The scales were presented in the same order for all respondents in both conditions. To ensure the scale labels were not producing carryover effects that would affect responses from one scale to the next, a distractor task (the SDS) consisting of dichotomously scored items was administered. The same version of the SDS was used across conditions. Figure 2 provides a visual diagram of the experimental design. Respondents completed the experiment in one sitting (median time to completion across both conditions: 14.98 min).

Statistical Analyses

Distributional Characteristics

To evaluate the impact of the labeling manipulation on response data we independently examined the distributional characteristics of each scale at the scale level and at the item level. At the scale level, we examined and compared total score means, variances, and internal consistency reliability (coefficient α). These procedures are consistent with previous research in this area and were used to contrast or supplement results from the psychometric analyses (described below). Item-level analyses were conducted for both scales since it is possible that the scale-level indices, which are aggregated across items and respondents, cannot capture, detect, or otherwise “cancel out” features which may be salient at the item level (Dixon et al., 1984). At the item level, the effect of the labeling manipulation was evaluated by creating an indicator variable for each item that took on a value of 1 if response category 1 or 7 was endorsed, and 0 otherwise. This variable captures the total number, and corresponding proportion, of respondents who endorsed an endpoint category for a given item (Greenleaf, 1992). If omission of the intermediate response labels has no affect on how respondents interact with the scale items, the proportion of endorsements of these two categories is expected to be the same across groups (within sampling error). Pearson chi-square tests of independence were computed for each item to determine if the proportion of endpoint endorsements differed between the two groups. All analyses were evaluated at the .05 significance level. To control for the family-wise error rate associated with the multiple tests, we employ the Dunn-Šidák procedure which gives the adjusted significance level $\alpha^\star = 1 - (1 - \alpha)^{1/m}$, where m is the number of independent tests and α is the significance level common to the tests. At the nominal $\alpha = .05$ level, the adjusted significance level for the SWLS becomes $\alpha^\star = .010206$ and $\alpha^\star = .003657$ for the ATMPM.

Psychometric Analyses: Measurement Invariance

We performed invariance analysis via multiple group confirmatory factor analysis (MGCFA; Jöreskog, 1971) to investigate if the labeling manipulation had an effect on the psychometric characteristics of the scales. In particular, we examined the impact of the labeling manipulation on the overall factor structure, factor loading estimates, and latent intercept estimates. To this end, we examined and compared three invariance models: a configural invariance model, a metric invariance model, and a scalar invariance model.

In the configural invariance model, all model parameters (factor loadings, latent intercepts, residual variances) were freely estimated and no between-group constraints were made. If the configural model was supported by the data, it was determined that the factor structure was similar, but not identical, across groups. The subsequent metric invariance model imposed equality constraints on the matrix of factor loadings across groups. Factor loadings can be conceptualized as the discriminating power of an item, therefore the metric invariance model reflects the condition where the set of items relate to the latent construct to the same extent between groups. If the metric invariance model was supported by the data (relative to the configural model), the model was constrained further by imposing equality constraints on the item intercepts across groups (scalar invariance). Item intercepts can be conceptualized as item difficulty parameters, therefore the scalar invariance model reflects the condition where items are equally difficult across groups. Satisfying the condition of scalar invariance is needed to make meaningful group comparisons with respect to the latent construct (Chen, 2007; Millsap, 2012). Moreover, evidence of scalar invariance is sufficient to demonstrate or establish measurement invariance (e.g., Gregorich, 2006; Little & Lee, 2015), thus we conclude our analysis with the scalar invariance model and we do not examine residual invariance or structural invariance.

Estimation of the invariance models was carried out in Mplus version 8.4 (Muthén & Muthén, 1998–2017) using the robust maximum likelihood estimator. We assessed model fit of the invariance models by examining absolute and comparative fit indices, including standardized root mean square residual (SRMR), root mean square error of approximation (RMSEA), comparative fit index (CFI), and the Tucker-Lewis index (TLI). The chi-square goodness-of-fit index also is reported, however we interpret it with caution as it has been shown to be sensitive in large sample settings (Brown, 2015). Following Hu and Bentler (1999), cutoff values of SRMR $\leq$ 0.08, RMSEA $\leq$ 0.06, CFI $\geq$ 0.95, and TLI $\geq$ 0.95 were used to provide evidence of acceptable model fit. To assess relative model fit of the configural, metric, and scalar models, we follow Chen (2007) who recommended that an absolute change of $\leq$ .010 in CFI, in addition to an absolute change of $\leq$ 0.15 in RMSEA or a change of $\leq$ .030 in SRMR to provide evidence of invariance. We also examine and compare information criteria across invariance models (AIC, BIC, sample size adjusted

BIC) to supplement the model fit indices and nested model comparisons (see below).

The degree to which model fit decreased as a result of parameter constraints was evaluated via nested model comparisons by computing the scaled likelihood ratio test statistic ($-2\Delta LL$; Millsap, 2012; Satorra & Bentler, 2010). The $-2\Delta LL$ test statistic is asymptotically chi-square distributed with degrees of freedom equal to the difference in degrees of freedom between the two competing models (Δdf). A statistically significant likelihood ratio test result indicates that the imposed constraints lead to a significant reduction of model fit. Modification indices were considered for any model parameter that was found to be non-invariant and applied only if the proposed modification could be substantively justified (Brown, 2015). The configural model was identified by setting the factor mean and variance to zero and one, respectively. This identification approach placed the latent variable(s) in a z-score metric and allowed all model parameters (loadings, latent intercepts, and residual variances) to be freely estimated. Metric and scalar invariance models were identified by applying the appropriate identification procedure.[2]

Results

Distributional Characteristics at the Scale Level

Table 2 provides item and scale level descriptive statistics for the SWLS and ATMPM, including group means, variances, and internal consistency coefficients. The standard deviations in the endpoints-only condition of the SWLS and ATMPM were found to be larger relative to the fully-labeled versions. To test whether the corresponding variances were statistically different from each other, Levene's test of equality of variances (Levene, 1961) was computed. For the SWLS, Levene's test indicated that the variances of the two groups did not differ significantly, $F(1, 637) = 1.62$, $p = .20$, however the variances

[2] The metric invariance model was identified by setting the factor means of both groups to 0, setting the factor variance of the fully-labeled group to 1, and estimating the factor variance of the endpoints-only group. The scalar invariance model was identified by estimating the factor mean and variance of the endpoints-only group while fixing them to 0 and 1 in the fully-labeled group, respectively.

were found to differ significantly between groups on the ATMPM, $F(1, 633) = 6.73$, $p = .009$, with greater variation in the endpoint-only group. Comparison of composite score means via t-tests from the SWLS indicated that the group means did not differ significantly, $t(637) = 1.67$, $p = .09$, $d = 0.13$, and a similar result was found for the ATMPM, $t(590.74)^3 = 0.97$, $p = .33$, $d = 0.07$. In addition, the effect size estimates via Cohen's d were relatively small by accepted standards, adding further support that the labeling manipulation had a relatively small effect on observed differences in composite score means. These results suggest that omission of the intermediate scale labels does not greatly impact the observed mean distribution of the scales, however variability of the longer scale was found to be slightly impacted.

Internal consistency reliability, as quantified by coefficient α, was higher in the endpoints-only condition for both scales. Specifically, the reliability estimates were .89 and .86 for the SWLS and .79 and .75 for the ATMPM under the endpoints-only and fully-labeled configurations, respectively. To compare these reliability estimates, we employed the k-sample test for independent α coefficients (Diedenhofen & Musch, 2016; Feldt, Woodruff, & Salih, 1987; Hakstian & Whalen, 1976). Under the null hypothesis, the test statistic is approximately chi-square distributed with $k - 1$ degrees of freedom where k is the number of groups. The two SWLS α coefficients were not significantly different from each other, $\chi^2(1) = 3.14$, $p = .07$. A similar result was found for the ATMPM, $\chi^2(1) = 2.54$, $p = .11$. These results suggest that scale reliability is not adversely affected by the omission of the intermediate scale labels. Taken together, results from comparison of means, variances, and internal consistency suggests that labeling the scale under the endpoints-only or the fully-labeled configuration has little overall affect on distributional characteristics at the scale level.

[3] The non-integer degrees of freedom comes from the Welch two sample t-test which provides an adjustment due to unequal group variances. A non-significant result was also found when assuming equal variances.

Distributional Characteristics at the Item Level

Tables 3 and 4 display the 2 × 2 item contingency tables for items in the SWLS and ATMPM, respectively. Visual inspection of the tables shows that the proportion of endpoint endorsements was higher in the endpoints-only condition for all five items of the SWLS and for 12 of the 14 items in the ATMPM. Statistical analyses from the chi-square tests revealed that proportion of endpoint endorsements were significantly different in 9 out of 19 items (47%). Of these, three came from the SWLS (60% of scale) and six from the ATMPM (43% of scale). The proportion of endpoint endorsements were significantly higher under the endpoints-only configuration for all 9 items. For instance, endpoint endorsements were 84% higher in the endpoints-only condition for item one of the SWLS, with 12.7% of respondents in the endpoints-only condition endorsing an endpoint compared to 6.9% of respondents in the fully-labeled condition. Similarly, for the ATMPM, endpoint endorsements were up to 82% higher in the endpoints-only condition. These findings suggest that, in the absence of intermediate response labels, respondents displayed a tendency to endorse an endpoint category at a significantly higher rate. This trend was also observed for the majority of the remaining items, suggesting that when intermediate response categories are left blank, respondents are more likely to shift their responses towards the labeled endpoints.

Invariance Analyses: SWLS

Model fit statistics from the invariance analyses for the SWLS and ATMPM are given in Table 5 and parameter estimates for the configural and final scalar invariance models are provided in Table 6. For the SWLS, model fit indices, with the exception of the chi-square statistic, indicated that the configural one-factor invariance model had excellent model fit, $\chi^2(10) = 24.45$, $p < .01$, SRMR = 0.019, RMSEA = .067, CFI = .987, TLI = .974, providing evidence the factor structure is not affected by changes in category labeling. The factor loadings of the configural model were larger for the endpoints-only group for all five items and all item intercepts were larger in the fully-labeled group, with the exception of item five which was larger in the endpoints-only group. Imposing equality constraints on the factor loadings between groups (metric invariance) revealed that the observed differences in item loadings were not statistically significant, suggesting the observed differences in factor loadings were most likely the result of sampling error. The fit of the metric model relative to the configural model was not significantly worse, $-2\Delta LL = 5.09$, $\Delta df = 4$, $p = 0.28$. Moreover, ΔCFI = 0.002, ΔRMSEA = 0.006, and ΔSRMR = -0.017, all satisfying the guidelines suggested by Chen (2007), supporting the metric invariance model. No modification indices were considered for the metric invariance model. In light of these results, we conclude that the respective items have the same discriminating power and measure the latent construct to the same degree, irrespective of labeling configuration.

Given that the full metric model was found to have satisfactory fit, we then imposed additional equality constraints on all item intercepts (scalar invariance). The resulting scalar model did not exhibit significantly worse fit relative to the metric model, $-2\Delta LL = 7.88$, $\Delta df = 4$, $p = 0.09$, however modification indices revealed that removing the intercept constraint from item 5 would enhance model fit. Interestingly, item 5, which reads “If I could live my life over, I would change almost nothing,” has been identified by previous research as being a relatively weak item in the SWLS compared to the remaining items (e.g., Pavot & Diener, 2009). Indeed, our analysis revealed that item 5 had the least discriminating power and smallest intercept across the two groups. Removing the intercept constraint from item 5 resulted in significant improvement in model fit over the full scalar model, $-2\Delta LL = 5.01$, $\Delta df = 1$, $p = 0.02$, but did not fit significantly worse than the metric model, $-2\Delta LL = 2.82$, $\Delta df = 3$, $p = 0.42$. Examination of information criteria, which takes into consideration model fit alongside model parsimony, further supported the scalar invariance models (see Table 5). Thus, both the full and partial scalar models are supported by the data, however the partial scalar model was found to be preferable to the full scalar model. These results indicate that, in general, items did not become more difficult for respondents as a function of labeling.

Invariance Analyses: ATMPM

For the ATMPM, the configural three-factor invariance model was found to have adequate model fit, $\chi^2(148) = 504.40$, $p < .001$, SRMR = 0.072, RMSEA = .086, CFI = .921, TLI = .903. Likelihood ratio tests and model fit indices indicated that constraining the factor loadings to be equal between groups resulted in virtually no reduction of model fit, $-2\Delta LL = 2.49$, $\Delta df = 11$, $p = 0.99$, ΔCFI = -0.001, ΔRMSEA = 0.003, and ΔSRMR = -0.001. Thus, similar to the results from the SWLS invariance analyses, the labeling manipulation appears to have little to no effect on the discriminating power of the items. Furthermore, constraining item intercepts to be equal between groups (scalar invariance) also revealed a very minimal reduction of model fit relative to the full metric model, $-2\Delta LL = 4.07$, $\Delta df = 11$, $p = 0.96$, ΔCFI = 0, ΔRMSEA = 0.003, and ΔSRMR = 0, suggesting that items were equally difficult regardless of labeling configuration. No modification indices were suggested or applied for any of the invariance models. Examination of information criteria further supported the scalar invariance model. Based on these results, we conclude that configural, full metric, and full scalar invariance are supported by the data.

Taken together, results from the invariance analyses indicate that the measurement properties of the two scales were unaffected as a function of labeling configuration. In particular, regardless of the labeling scheme, the items measure the latent construct to the same degree between groups (discriminating power is unaffected) and items are equally difficult across labeling configurations. On the basis of these results, along with results from the distributional analyses at the scale and item level, we conclude that labeling configuration (fully-labeled, endpoint-only) has minimal effect on the psychometric and distributional properties of rating scales, with the exception of proportion of endpoint endorsements at the item level which was found to be higher on several items in the endpoints-only condition.

Discussion

For decades it has been known that the quality of responses obtained from RS and surveys are influenced by the characteristics of the scale (see DeCastellarnau (2018) for a review). Research in this area has led to a deeper understanding of scale characteristics that influence responses and remains an important area of inquiry in the scale development literature. The present study investigated the impact of labeling formats, a design characteristic for which little or conflicting empirical research exists. In particular, the purpose of the present study was to investigate the effects of differential labeling schemes on the distributional characteristics and psychometric properties of RS. We manipulated category labeling and administered two fully-labeled or endpoints-only RS to a set of respondents and we examined the extent to which these labeling configurations lead to substantively different results in terms of distributional characteristics and measurement properties of the scales.

Our analyses indicated that distributional characteristics of rating scales are relatively unaffected when the unit of analysis is a composite score. In particular, labeling does not appear to have an effect on group means, however there was a difference in variation of observed scores, with the endpoints-only labeling format displaying greater variation in responses. This finding, however, only applied to the ATMPM, a scale almost three times as long as the SWLS, suggesting that longer scales are potentially more sensitive to changes in labeling. It is unclear if this observation generalizes to other scales of similar length, thus more research is needed to support or refute this hypothesis.

At the item level, we found evidence that endpoint endorsements were significantly higher in the endpoints-only condition across the two scales. Our findings corroborate previous research by Eutsler and Lang (2015), Menold et al. (2014), Moors et al. (2014) and Weijters et al. (2010) who have found extreme responses to be more prevalent in partially-labeled scales. It appears that respondents are misinterpreting what the unlabeled categories represent, as hypothesized by Krosnick (1991), prompting them to disperse towards one of the defined endpoints. An alternative explanation is that respondents are engaging in extreme response style (ERS), a “personality-like” characteristic (e.g., Greenleaf, 1992) whereby respondents fail to consider all provided response options and instead choose to favor or avoid extreme categories, irrespective of item content or their true attitudes. However, given that our experiment consisted of a randomization component, this procedure would, in theory, evenly distribute individuals with varying levels of ERS across the two conditions. Our observation that a substantial number of items exhibited differences in endpoint endorsements even after taking into consideration the experimental design supports the notion that unlabeled response categories cause respondents to intentionally shift their responses towards labeled categories.

Contrary to most of the existing literature in this area, our findings indicate that labeling does not have a major impact on internal consistency reliability. Although our analyses indicated that the endpoint-only scales had higher reliability, these findings were not statistically significant according to hypothesis tests designed for reliability coefficients, suggesting that the observed difference in reliability was most likely the result of sampling error. Statistical significance notwithstanding, the finding that the endpoints-only version of the scales exhibited higher reliability may seem rather counterintuitive at first glance, as one might expect fully-labeled scales to produce higher reliability. However, as noted by Hamby and Taylor (2016), reliability estimates can be artificially inflated in partially-labeled scales due to the increased use of cognitive shortcuts, such as satisficing, or due to lack of incentive or motivation. Thus, making a determination about whether to use a fully- or partially-labeled response format on the basis of reliability is not recommended as reliability indices can lead to incorrect inferences, particularly if the underlying assumptions of the reliability coefficient are not satisfied (e.g., unidimensionality, tau-equivalence; see (McNeish, 2018)) or if respondents engage in practices that cause reliability indices to artificially increase or decrease.

Psychometric analyses via multiple group confirmatory factor analysis revealed that the structural validity of scales is not compromised due to changes in labeling. This is evidenced by similarities in model fit, factor structure, factor loadings, and latent intercept parameters across labeling conditions. Interestingly, only one item was found to be non-invariant (item 5 of the SWLS). This item was found to have a substantially smaller standardized factor loading relative to the other items, with the construct accounting for only 34% of the observed variance. It may be the case that poorly constructed items (e.g., items that are weakly related to the target construct) are more susceptible and easily influenced by changes in labeling format. This interpretation is supported by Weng (2004) who reported that items with large factor loadings were less likely to be affected by changes in scale format. Scale developers wishing to use partially-labeled scales are recommended to select items that are highly related to the target construct, while scale developers utilizing a fully-labeled format may be able to include less informative items in their assessment. We note that these recommendations are preliminary, as more research is needed to provide more definitive guidelines in this regard. Overall, our invariance analyses overwhelmingly support the notion that labeling has virtually no affect on the measurement and psychometric properties of rating scales.

A few limitations of the present study are now discussed. First, although the scales have been developed for use within a general population (the SWLS is written at a 6th grade level (Pavot & Diener, 2009); the ATMPM has been found to perform well with samples of adolescent grade-school students (Won & Shirley, 2018)), college populations may be better acquainted with various response formats as a result of their educational interactions (e.g., course evaluations, surveys administered by departments, etc.). The extent to which these results generalize to populations with varying levels of education is of interest, particularly because other scale design characteristics have been found to have a more salient effect in populations with lower levels of education (e.g., response-order effects; Knauper, 1999).

Second, we note that it is possible that fatigue and boredom effects occurred towards the tail end of the survey sets. Overall respondents were administered over 50 items, which may have been enough to induce these effects. The distractor task was selected to remove carry-over effects, however we realize that a 33-item distractor task may have been more than necessary. Third, we note that both scales used in the analyses consisted of a 7-point response format. It is unclear if these results general to other popular choices, such as the 5- or 6-point response format, however we predict that the general findings presented here would be similar under these response formats. Lastly, we restricted our labeling format to agree-disagree scales. Many other response formats are frequently employed, including asymmetric scales (e.g., "somewhat happy," "a little happy," "very happy") and continuous rating scales (e.g., Chyung, Swanson, Roberts, & Hankinson, 2018), each of which vary in the amount and type of information they provide in the scale labels. Scale developers who utilize scale formats other than the agree-disagree format should interpret the results from this study with caution as it is unknown if they generalize to such formats. Thus, we recommend that scale developers attempt to make the scale labeling decision an explicit step in the scale construction process and to ensure the selected label configuration is congruent and consistent with the purposes of the instrument (Simms, 2008).

Despite these limitations, the present study has several strengths. First, the study utilized published scales to evaluate the effects of response manipulations on distributional and psychometric characteristics. Doing so allowed us to examine these effects with empirically established measures, connecting the present research to empirical, real-life applications. Secondly, we gave considerable attention to the scales at the item-level and under a psychometric framework. This is in contrast to most of the extant literature which has focused on scale level differences. A final strength of the study is the experimental setting. Although not technically a truly randomized experiment, we believe our design sufficiently demonstrates that labeling configuration does not adversely affect the distributional characteristics and psychometric properties of RS, giving scale developers one less scale design characteristic to worry about.

References

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement* (Vol. 547). John Wiley & Sons.

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods & Research*, *20*(1), 139–181.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, *48*(2), 409–442.

Armstrong, R. L. (1987). The midpoint on a five-point likert-type scale. *Perceptual and Motor Skills*, *64*(2), 359–362.

Betts, L., & Hartley, J. (2012). The effects of changes in the order of verbal labels and numerical values on children's scores on attitude and rating scales. *British Educational Research Journal*, *38*(2), 319–331.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research.* (2nd ed.). New York, NY: The Guilford Press.

Cabooter, E., Weijters, B., Geuens, M., & Vermeir, I. (2016). Scale format effects on response option interpretation and use. *Journal of Business Research*, *69*(7), 2574–2584.

Chang, L. (1997). Dependability of anchoring labels of likert-type scales. *Educational and Psychological Measurement*, *57*(5), 800–807.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.

Churchill, G. A., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research*, *21*(4), 360–375.

Chyung, S. Y., Roberts, K., Swanson, I., & Hankinson, A. (2017). Evidence-based survey design: The use of a midpoint on the likert scale. *Performance Improvement*, *56*(10), 15–23.

Chyung, S. Y., Swanson, I., Roberts, K., & Hankinson, A. (2018). Evidence-based survey design: The use of continuous rating scales in surveys. *Performance Improvement*, *57*(5), 38–48.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349.

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: a literature review. *Quality & Quantity*, *52*(4), 1523–1559.

Diedenhofen, B., & Musch, J. (2016). cocron: A web interface and r package for the statistical comparison of Cronbach's alpha coefficients. *International Journal of Internet Science*, *11*(1), 51–60.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71–75.

Diener, E., Sandvik, E., Seidlitz, L., & Diener, M. (1993). The relationship between income and subjective well-being: Relative or absolute? *Social Indicators Research*, *28*(3), 195–223.

Dixon, P. N., Bobo, M., & Stevick, R. A. (1984). Response differences and preferences for all-category-defined and end-defined likert formats. *Educational and Psychological Measurement*, *44*(1), 61–66.

Eutsler, J., & Lang, B. (2015). Rating scales in accounting research: The impact of scale points and labels. *Behavioral Research in Accounting*, *27*(2), 35–51.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*(1), 93–103.

Finn, R. (1972). Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement*, *32*(2), 255–265.

Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, *72*(5), 892–913.

Garland, R. (1991). The mid-point on a rating scale: Is it desirable. *Marketing Bulletin*, *2*(1), 66–70.

Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly*, *56*(3), 328–351.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, *44*(11), S78–S94.

Hakstian, A. R., & Whalen, T. E. (1976). A *k*-sample significance test for independent alpha coefficients. *Psychometrika*, *41*(2), 219–231.

Hamby, T., & Taylor, W. (2016). Survey satisficing inflates reliability and validity measures: An experimental comparison of college and amazon mechanical turk samples. *Educational and Psychological Measurement*, *76*(6), 912–932.

Hernández, A., Drasgow, F., & González-Romá, V. (2004). Investigating the functioning of a middle category by means of a mixed-measurement model. *Journal of Applied Psychology*, *89*(4), 687–699.

Herring, D. R., White, K. R., Jabeen, L. N., Hinojos, M., Terrazas, G., Reyes, S. M., . . . Crites Jr., S. L. (2013). On the automatic activation of attitudes: A quarter century of evaluative priming research. *Psychological Bulletin*, *139*(5), 1062–1089.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Huck, S. W., & Jacko, E. J. (1974). Effect of varying the response format of the alpert-haber achievement anxiety test. *Journal of Counseling Psychology*, *21*(2), 159–163.

Jang, S., Kim, E. S., Cao, C., Allen, T. D., Cooper, C. L., Lapierre, L. M., … Woo, J. (2017). Measurement invariance of the satisfaction with life scale across 26 countries. *Journal of Cross-Cultural Psychology*, *48*(4), 560–576.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426.

Keusch, F., & Yang, T. (2018). Is satisficing responsible for response order effects in rating scale questions? *Survey Research Methods*, *12*(3), 259–270.

Klauer, K. C., Rossnagel, C., & Musch, J. (1997). List-context effects in evaluative priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 246–255.

Knauper, B. (1999). The impact of age and education on response order effects in attitude measurement. *Public Opinion Quarterly*, *63*(3), 347–370.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*(2), 201–219.

Krosnick, J. A., & Berent, M. K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science*, *37*(3), 941–964.

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In *Survey measurement and process quality* (p. 141-164). John Wiley & Sons, Ltd.

Krosnick, J. A., Judd, C. M., & Wittenbrink, B. (2018). The measurement of attitudes. In *The handbook of attitudes* (pp. 45–105). Routledge.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In *Handbook of survey research* (pp. 263–313). Emerald Group Publishing.

Kunz, T. (2015). *Rating scales in web surveys. a test of new drag-and-drop rating procedures* (Unpublished doctoral dissertation). Technische Universität.

Lam, T. C., & Stevens, J. J. (1994). Effects of content polarization, item wording, and rating scale width on rating response. *Applied Measurement in Education*, *7*(2), 141–158.

Lee, J., & Paek, I. (2014). In search of the optimal number of response categories in a rating scale. *Journal of Psychoeducational Assessment*, *32*(7), 663–673.

Leung, S. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point likert scales. *Journal of Social Service Research*, *37*(4), 412–421.

Levene, H. (1961). Robust tests for equality of variances. *Contributions to probability and statistics*, 279–292.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Little, T. D., & Lee, J. (2015). Factor analysis: Multiple groups. In *Wiley statsref: Statistics reference online* (p. 1-10). John Wiley & Sons, Ltd.

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, *4*(2), 73–79.

Martinez, A. J. (2021). Factor structure and measurement invariance of the academic time management and procrastination measure. *Journal of Psychoeducational Assessment*, *0*(0), 1–11.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433.

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales? *Field Methods*, *26*(1), 21–39.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance.* Routledge.

Moors, G., Kieruj, N. D., & Vermunt, J. K. (2014). The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociological Methodology*, *44*(1), 369–399.

Muthén, B., & Muthén, L. (1998–2017). *Mplus user's guide.* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Nadler, J. T., Weston, R., & Voyles, E. C. (2015). Stuck in the middle: The use and interpretation of mid-points in items on questionnaires. *The Journal of General Psychology*, *142* (2), 71–89.

Newstead, S. E., & Arnold, J. (1989). The effect of response format on ratings of teaching. *Educational and Psychological Measurement*, *49* (1), 33–43.

Nima, A. A., Cloninger, K. M., Persson, B. N., Sikström, S., & Garcia, D. (2020). Validation of subjective well-being measures using item response theory. *Frontiers in psychology*, *10*, 3036.

Pavot, W., & Diener, E. (2009). Review of the satisfaction with life scale. In *Assessing well-being* (pp. 101–117). Springer.

Pavot, W., Diener, E., Colvin, C. R., & Sandvik, E. (1991). Further validation of the satisfaction with life scale: Evidence for the cross-method convergence of well-being measures. *Journal of Personality Assessment*, *57* (1), 149–161.

Presser, S., & Schuman, H. (1980). The measurement of a middle position in attitude surveys. *Public Opinion Quarterly*, *44* (1), 70–85.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104* (1), 1–15.

Saris, W. E., & Gallhofer, I. N. (2014). *Design, evaluation, and analysis of questionnaires for survey research.* John Wiley & Sons.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75* (2), 243–248.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84* (2), 127–190.

Simms, L. J. (2008). Classical and modern methods of psychological scale construction. *Social and Personality Psychology Compass*, *2*(1), 414–433.

Spector, P. E. (1976). Choosing response categories for summated rating scales. *Journal of Applied Psychology*, *61*(3), 374–375.

Spratto, E. M., & Bandalos, D. L. (2020). Attitudinal survey characteristics impacting participant responses. *The Journal of Experimental Education*, *88*(4), 620–642.

Spratto, E. M., Leventhal, B. C., & Bandalos, D. L. (2020). Seeing the forest and the trees: Comparison of two irtree models to investigate the impact of full versus endpoint-only response option labeling. *Educational and Psychological Measurement*, *81*(1), 39–60.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge University Press.

Vassar, M. (2008). A note on the score reliability for the satisfaction with life scale: An RG study. *Social Indicators Research*, *86*(1), 47–57.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236–247.

Weng, L. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, *64*(6), 956–972.

Wetzel, E., Frick, S., & Greiff, S. (2020). The multidimensional forced-choice format as an alternative for rating scales. *European Journal of Psychological Assessment*, *36*(4), 511-515.

Wetzel, E., & Greiff, S. (2018). The world beyond rating scales. *European Journal of Psychological Assessment*, *34*(1), 1–5.

Won, S., & Shirley, L. Y. (2018). Relations of perceived parental autonomy support and control with adolescents' academic time management and procrastination. *Learning and Individual Differences*, *61*, 205–215.

Wyatt, R. C., & Meyers, L. S. (1987). Psychometric properties of four 5-point likert type response scales. *Educational and Psychological Measurement*, *47*(1), 27–35.

Table 1

Respondent Demographic and Background Information

Demographics & Background Information		Fully-labeled	Endpoints-only	Test Statistic
Gender	Male	77 (12%)	69 (11%)	$\chi^2(1) = 0.017$, $p = .895$
	Female	262 (41%)	229 (36%)	
Age	Range	18-31	18-37	$t(637) = -1.306$, $p = .192$
	Mean	19.08	19.25	
	Standard Deviation	1.44	1.81	
Time to Survey Completion	1st Quantile	11.24 min	11.36 min	$t(654) = 0.762$, $p = .446$
	Mean	86.91 min	50.93 min	
	Median	15.21 min	14.75 min	
	3rd Quantile	21.29 min	19.40 min	
Ethnicity/Race	African-American	18 (3%)	9 (1%)	$\chi^2(4) = 5.952$, $p = .203$
	Asian-American	71 (11%)	50 (8%)	
	Caucasian	50 (8%)	39 (6%)	
	Hispanic	183 (29%)	180 (28%)	
	Other	18 (3%)	22 (3%)	
Did Any Parent Attend College?	Yes	176 (28%)	141 (22%)	$\chi^2(1) = 1.351$, $p = .245$
	No	164 (26%)	158 (24%)	

Table 2

Item Descriptive Statistics for the SWLS and ATMPM

	Fully-labeled					Endpoints-only				
Scale/Item	*M*	*SD*	Skewness	Kurtosis	Coefficient α	*M*	*SD*	Skewness	Kurtosis	Coefficient α
SWLS: Satisfaction with Life Scale										
1. In most ways my life is close to ideal.	4.40	1.48	-0.42	-0.60		4.24	1.52	-0.22	-0.37	
2. The conditions of my life are excellent.	4.46	1.45	-0.33	-0.68		4.42	1.48	-0.12	-0.52	
3. I am satisfied with my life.	4.81	1.45	-0.49	-0.46		4.61	1.56	-0.44	-0.37	
4. So far I have gotten the important things I want in life.	4.81	1.47	-0.5	-0.43		4.48	1.67	-0.18	-0.81	
5. If I could live my life over, I would change almost nothing.	3.76	1.73	0.26	-0.91		3.83	1.87	0.13	-1.02	
Total score	22.45	6.10			.86	21.59	6.81			.89
ATMPM: Academic Time Management & Procrastination Measure										
1. I set deadlines for myself when I set out to accomplish an assignment.	4.84	1.57	-0.62	-0.29		4.98	1.68	-0.55	-0.52	
2. I set short-term goals for the studying I want to accomplish in a few days or weeks.	4.69	1.52	-0.55	-0.31		4.78	1.67	-0.32	-0.75	
3. I have a system for managing the time I spend on my academic work.	4.22	1.75	-0.18	-0.98		4.44	1.82	-0.16	-1.01	
4. I have specific times set aside during the week to get my schoolwork done.	4.28	1.80	-0.26	-0.97		4.44	1.80	-0.28	-0.92	
5. I often set goals or make lists regarding what I need to get done each day.	4.79	1.70	-0.53	-0.57		4.85	1.80	-0.49	-0.74	
6. I look at a planner, schedule or calendar every day to see what I need to get done.	4.58	1.85	-0.33	-0.95		4.69	2.01	-0.47	-0.98	
7. I frequently use a planner, schedule or calendar to organize all my time commitments.	4.48	1.92	-0.27	-1.02		4.65	2.05	-0.35	-1.17	
8. I make a list of things to do each day and check off each task as it is accomplished.	4.24	1.93	-0.13	-1.11		4.41	2.06	-0.22	-1.27	
9. To make sure I don't forget to do my schoolwork, I often write myself notes or reminders.	5.10	1.70	-0.74	-0.28		5.09	1.82	-0.71	-0.6	
10. I often find excuses for not starting the work for my classes.	4.41	1.68	-0.35	-0.62		4.03	1.75	0.05	-0.89	
11. I promise myself I will do my schoolwork, then put it off anyway.	4.52	1.67	-0.27	-0.77		4.16	1.83	-0.09	-1.03	
12. I frequently put off getting started on the readings and assignments for my classes.	4.70	1.64	-0.38	-0.58		4.31	1.73	-0.22	-0.79	
13. I delay studying for my classes, even when it is important.	4.42	1.73	-0.28	-0.83		4.04	1.83	-0.02	-0.99	
14. I postpone doing the work for my classes until the last minute.	4.51	1.80	-0.32	-0.85		4.03	1.88	-0.05	-1.09	
Total score	63.84	11.80			.75	62.86	13.48			.79

Table 3

Satisfaction with Life Scale Item Contingency Table

	Scale Label Configuration	Did Not Endorse Endpoint, *n* (%)	Endorsed Endpoint, *n* (%)	Percentage increase[a]
Item 1[*]	Fully Labeled	322 (93.1%)	24 (6.9%)	84%
	Endpoints Only	262 (87.3%)	38 (12.7%)	
Item 2	Fully Labeled	314 (91.0%)	31 (9.0%)	30%
	Endpoints Only	265 (88.3%)	35 (11.7%)	
Item 3	Fully Labeled	305 (88.7%)	39 (11.3%)	39%
	Endpoints Only	252 (84.3%)	47 (15.7%)	
Item 4[*]	Fully Labeled	302 (87.5%)	43 (12.5%)	47%
	Endpoints Only	244 (81.6%)	55 (18.4%)	
Item 5[*]	Fully Labeled	288 (83.5%)	57 (16.5%)	50%
	Endpoints Only	226 (75.3%)	74 (24.7%)	

$^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

[a] Percentage increase was calculated as follows: $(\%E_F - \%E_E)/\%E_F$ where $\%E_F$ is the proportion of endpoint endorsements of the fully-labeled condition and $\%E_E$ is the proportion of endpoint endorsements of the endpoints-only condition.

Table 4

Academic Time Management & Procrastination Measure Item Contingency Table

	Scale Label Configuration	Did Not Endorse Endpoint, *n* (%)	Endorsed Endpoint, *n* (%)	Percentage increase[a]
Item 1***	Fully Labeled	286 (82.7%)	60 (17.3%)	64%
	Endpoints Only	215 (71.7%)	85 (28.3%)	
Item 2***	Fully Labeled	300 (87.0%)	45 (13.0%)	82%
	Endpoints Only	229 (76.3%)	71 (23.7%)	
Item 3*	Fully Labeled	287 (82.9%)	59 (17.1%)	42%
	Endpoints Only	226 (75.8%)	72 (24.2%)	
Item 4	Fully Labeled	277 (80.3%)	68 (19.7%)	9%
	Endpoints Only	235 (78.6%)	64 (21.4%)	
Item 5*	Fully Labeled	267 (77.4%)	78 (22.6%)	33%
	Endpoints Only	210 (70.0%)	90 (30.0%)	
Item 6**	Fully Labeled	250 (72.7%)	94 (27.3%)	36%
	Endpoints Only	189 (63.0%)	111 (37.0%)	
Item 7*	Fully Labeled	243 (70.2%)	103 (29.8%)	30%
	Endpoints Only	184 (61.3%)	116 (38.7%)	
Item 8	Fully Labeled	250 (72.3%)	96 (27.7%)	23%
	Endpoints Only	198 (66.0%)	102 (34.0%)	
Item 9	Fully Labeled	239 (69.5%)	105 (30.5%)	12%
	Endpoints Only	197 (65.9%)	102 (34.1%)	
Item 10	Fully Labeled	283 (81.8%)	63 (18.2%)	3%
	Endpoints Only	244 (81.3%)	56 (18.7%)	
Item 11	Fully Labeled	284 (82.1%)	62 (17.9%)	21%
	Endpoints Only	235 (78.3%)	65 (21.7%)	
Item 12	Fully Labeled	273 (79.1%)	72 (20.9%)	-9%[b]
	Endpoints Only	243 (81.0%)	57 (19.0%)	
Item 13	Fully Labeled	280 (80.9%)	66 (19.1%)	19%
	Endpoints Only	232 (77.3%)	68 (22.7%)	
Item 14	Fully Labeled	263 (76.0%)	83 (24.0%)	-3%[b]
	Endpoints Only	230 (76.7%)	70 (23.3%)	

$^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

[a] Percentage increase was calculated as follows: $(\%E_F - \%E_E)/\%E_F$ where $\%E_F$ is the proportion of endpoint endorsements of the fully-labeled condition and $\%E_E$ is the proportion of endpoint endorsements of the endpoints-only condition.

[b] Negative percentages indicate that the proportion of endpoint endorsements were higher in the fully-labeled condition.

Table 5

Configural, Metric, and Scalar Invariance Model Fit Indices

	Satisfaction with Life Scale				Academic Time Management & Procrastination Measure		
	Configural	Metric	Scalar	Partial Scalar	Configural	Metric	Scalar
Chi-square	24.45	31.03	39.07	34.55	504.4	513.29	523.56
Degrees of freedom	10	14	18	17	148	159	170
Log-likelihood	-5134.08	-5136.30	-5140.28	-5137.72	-14916.17	-14917.47	-14919.54
$-2\Delta LL$		5.088	7.877	5.007		2.485	4.066
CFI	0.987	0.985	0.981	0.984	0.921	0.922	0.922
TLI	0.974	0.978	0.979	0.982	0.903	0.910	0.916
RMSEA (90% CI)	0.067 (0.033, 0.101)	0.061 (0.032, 0.091)	0.06 (0.034, 0.086)	0.057 (0.029, 0.084)	0.086 (0.078, 0.095)	0.083 (0.075, 0.091)	0.08 (0.072, 0.088)
SRMR	0.019	0.036	0.043	0.041	0.072	0.073	0.073
AIC	10328.15	10324.6	10324.56	10321.44	30012.33	29992.94	29975.08
BIC	10462.28	10440.84	10422.92	10424.27	30414.7	30346.13	30279.09
aBIC	10367.03	10358.29	10353.07	10351.25	30128.96	30095.31	30063.2

$^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$

Notes: CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; AIC = Akaike information criterion; BIC = Bayesian information criterion; aBIC = sample size adjusted BIC, CI = confidence interval.

Table 6

Standardized (Unstandardized) Configural and Final Scalar Model Parameter Estimates

Satisfaction with Life Scale													
Configural Model							Partial Scalar Model						
	Fully-labeled			Endpoints-only				Fully-labeled			Endpoints-only		
Item	Factor Loading	Intercept	Residual Variance	Factor Loading	Intercept	Residual Variance	Item	Factor Loading	Intercept	Residual Variance	Factor Loading	Intercept	Residual Variance
1	1.13 (0.77)	4.40 (2.99)	0.90 (0.42)	1.29 (0.85)	4.24 (2.80)	0.42 (0.27)	1	1.14 (0.77)	4.43 (2.99)	0.90 (0.41)	1.14 (0.85)	4.43 (2.95)	0.63 (0.28)
2	1.21 (0.84)	4.65 (3.22)	0.63 (0.3)	1.26 (0.86)	4.42 (3.00)	0.30 (0.27)	2	1.16 (0.82)	4.65 (3.28)	0.66 (0.33)	1.16 (0.87)	4.65 (3.09)	0.56 (0.25)
3	1.25 (0.86)	4.82 (3.33)	0.54 (0.26)	1.37 (0.88)	4.61 (2.95)	0.26 (0.23)	3	1.23 (0.86)	4.83 (3.36)	0.55 (0.27)	1.23 (0.88)	4.83 (3.07)	0.55 (0.22)
4	1.04 (0.71)	4.8 (3.27)	1.07 (0.5)	1.32 (0.79)	4.48 (2.68)	0.5 (0.38)	4	1.12 (0.74)	4.75 (3.13)	1.05 (0.46)	1.12 (0.77)	4.75 (2.92)	1.08 (0.41)
5	1.00 (0.58)	3.76 (2.18)	1.98 (0.67)	1.16 (0.62)	3.83 (2.05)	0.67 (0.61)	5	1.02 (0.59)	3.76 (2.17)	1.97 (0.65)	1.02 (0.62)	4.02 (2.17)	2.13 (0.62)

Academic Time Management & Procrastination Measure													
Configural Model							Scalar Model						
	Fully-labeled			Endpoints-only				Fully-labeled			Endpoints-only		
Item	Factor Loading	Intercept	Residual Variance	Factor Loading	Intercept	Residual Variance	Item	Factor Loading	Intercept	Residual Variance	Factor Loading	Intercept	Residual Variance
1	1.23 (0.79)	4.84 (3.09)	0.94 (0.38)	1.34 (0.8)	4.98 (2.96)	1.02 (0.36)	1	1.25 (0.79)	4.84 (3.07)	0.93 (0.38)	1.25 (0.79)	4.84 (2.89)	1.03 (0.37)
2	1.17 (0.77)	4.69 (3.1)	0.92 (0.4)	1.31 (0.79)	4.78 (2.88)	1.04 (0.38)	2	1.2 (0.78)	4.67 (3.05)	0.91 (0.39)	1.2 (0.78)	4.67 (2.85)	1.06 (0.39)
3	1.47 (0.84)	4.22 (2.41)	0.89 (0.29)	1.55 (0.86)	4.44 (2.45)	0.88 (0.27)	3	1.46 (0.84)	4.24 (2.43)	0.9 (0.3)	1.46 (0.86)	4.24 (2.33)	0.87 (0.26)
4	1.38 (0.77)	4.28 (2.38)	1.33 (0.41)	1.43 (0.79)	4.44 (2.47)	1.19 (0.37)	4	1.36 (0.76)	4.28 (2.4)	1.35 (0.42)	1.36 (0.8)	4.28 (2.37)	1.18 (0.36)
5	1.24 (0.73)	4.79 (2.83)	1.32 (0.46)	1.31 (0.73)	4.85 (2.69)	1.53 (0.47)	5	1.23 (0.73)	4.76 (2.82)	1.33 (0.47)	1.23 (0.73)	4.76 (2.63)	1.53 (0.47)
6	1.73 (0.94)	4.57 (2.47)	0.42 (0.12)	1.91 (0.95)	4.69 (2.34)	0.39 (0.1)	6	1.74 (0.94)	4.56 (2.47)	0.42 (0.12)	1.74 (0.95)	4.56 (2.28)	0.39 (0.1)
7	1.73 (0.91)	4.48 (2.34)	0.66 (0.18)	1.93 (0.94)	4.65 (2.27)	0.49 (0.12)	7	1.75 (0.91)	4.49 (2.34)	0.65 (0.18)	1.75 (0.94)	4.49 (2.2)	0.5 (0.12)
8	1.49 (0.78)	4.24 (2.2)	1.49 (0.4)	1.52 (0.74)	4.41 (2.14)	1.93 (0.46)	8	1.45 (0.76)	4.26 (2.25)	1.5 (0.42)	1.45 (0.75)	4.26 (2.02)	1.92 (0.43)
9	1.05 (0.62)	5.1 (3)	1.79 (0.62)	1.18 (0.65)	5.09 (2.8)	1.92 (0.58)	9	1.06 (0.62)	5.06 (2.96)	1.79 (0.62)	1.06 (0.64)	5.06 (2.8)	1.92 (0.59)
10	1.28 (0.77)	4.41 (2.63)	1.16 (0.41)	1.36 (0.78)	4.03 (2.31)	1.2 (0.39)	10	1.29 (0.77)	4.4 (2.62)	1.16 (0.41)	1.29 (0.78)	4.4 (2.53)	1.2 (0.4)
11	1.42 (0.85)	4.52 (2.71)	0.76 (0.27)	1.47 (0.8)	4.17 (2.28)	1.19 (0.36)	11	1.41 (0.85)	4.53 (2.74)	0.76 (0.28)	1.41 (0.8)	4.53 (2.46)	1.2 (0.35)
12	1.42 (0.87)	4.7 (2.87)	0.67 (0.25)	1.52 (0.88)	4.31 (2.5)	0.67 (0.23)	12	1.43 (0.87)	4.7 (2.85)	0.67 (0.25)	1.43 (0.88)	4.7 (2.74)	0.68 (0.23)
13	1.51 (0.87)	4.43 (2.56)	0.7 (0.24)	1.59 (0.87)	4.04 (2.21)	0.82 (0.25)	13	1.51 (0.87)	4.44 (2.57)	0.7 (0.24)	1.51 (0.87)	4.44 (2.43)	0.82 (0.25)
14	1.54 (0.86)	4.51 (2.5)	0.87 (0.27)	1.59 (0.85)	4.03 (2.15)	1.01 (0.29)	14	1.53 (0.85)	4.48 (2.5)	0.87 (0.27)	1.53 (0.85)	4.48 (2.37)	1 (0.28)

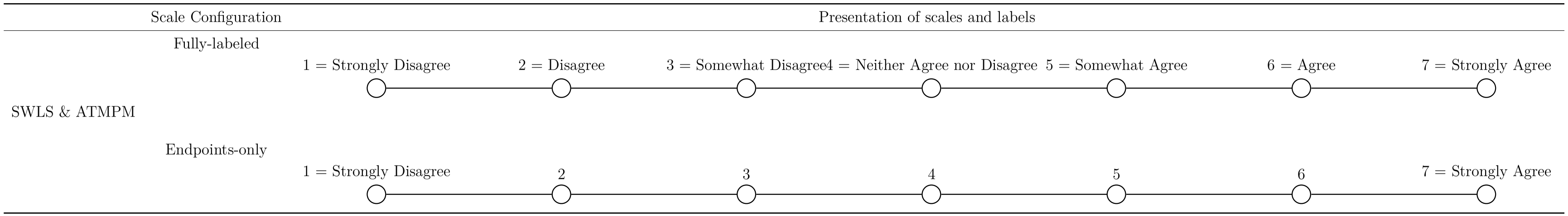

Figure 1

Scale Labels for Fully-labeled and Endpoints-only Labeling Conditions

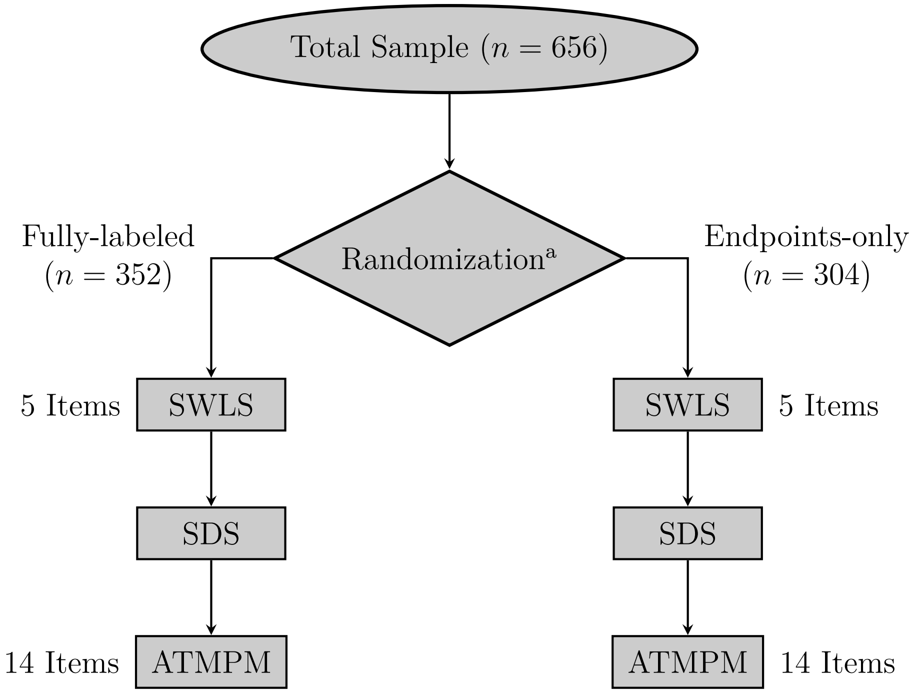

Figure 2

Experimental Design

[a] Randomization was actually pseudo-random as respondents selected between one of two identical links. Respondents were not aware of the condition they were in and once they selected one link, the remaining link became unavailable.