The Effect of Item Randomization on Scale Psychometrics

Erin M. Buchanan[1], Jeffrey Pavlacic[2], Becca Huber[3], & Alyssa Counsell[4]

[1] Harrisburg University of Science and Technology

[2] University of Mississippi

[3] Idaho State University

[4] Ryerson University

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

Correspondence concerning this article should be addressed to Erin M. Buchanan, 326 Market St., Harrisburg, PA 17101. E-mail: ebuchanan@harrisburgu.edu

13 Abstract

14 Some words here for the abstract

15 *Keywords:* scales, psychometrics, item randomization, research methods

The Effect of Item Randomization on Scale Psychometrics

MAKE A PACKRAT - figure out how to reference other markdowns

Classify by sample size, classify by scale size random versus not random (drop paper) See if/where it breaks - configural - metric - scalar - strict residuals

(Buchanan, 2002) says that stuff is stuff and stuff. (**???**) said more stuff. This file needs an update.

Measure Randomization

Self-report measures are a common assessment tool used in social sciences and can be administered in a variety of ways (e.g., paper and pencil, online surveys). Varying the administration method, generally, does not affect the equivalence of measures across administration type (Deutskens et al., 2016). To control for the methodological influences that self-report measures present, researchers will also randomize items within a measure to avoid counterbalance and method concerns (Keppel & Wickens, 2004). Randomization, too, has been shown not to alter scale psychometrics under certain conditions (Buchanan, Foreman, Johnson, Pavlacic, Swadley, & Schulenberg, 2018). Altering the delivery method and randomizing items is an appealing option that (1) allows for flexibility in data collection and (2) does not yield significant differences across items (Buchanan et al., 2018; T. Buchanan et al., 2006; Meyerson & Tyron, 2003; Ruyter & Wetzels, 2006). Of course, the use of self-report measures and techniques such as randomization are not without their limitations.

Despite some studies showing null effects in terms of equivalence across administration type, item ordering can negatively affect scale psychometrics under certain conditions (Schuman & Presser, 1981). In the information systems literature, grouping items together

based on construct similarily leads to different reliability coefficients compared to randomized items (Wilson & Lankton, 2012). Item randomization also reduces reliability coefficients in marketing research (Bradlow & Fitzimons, 2001). Given these findings, it is concerning that researchers typically choose not to report how/if items were randomized during data collection (Wilson, Srite, & Loicono, 2017). From a theoretical perspective, changes in interpreting future items based on previous items is a well-established psychological phenomenon (Tourangeau & Rasinski, 1988). Self-report measures are also exposed to other contextual influences, such as item randomization and common method variance (Edwards, 2008). Other ecological influences, such as mood states or lack of sleep, can also confound the psychometric properties of self-report measures (Trull & Ebner-Preiemer, 2009).

A less examined area related to item randomization is concerned with how randomization affects the factor structure of measures. T. Buchanan et al. (2005) show that factor structure did not replicate in online surveys compared to mailed surveys, while Weinberger, Darkes, Del Boca, Greenbaum, and Goldman (2006) show that factor structure varies depending on the order of items in a sensation seeking measure. Ultimately, the literature is not conclusive regarding how measurement method impacts factor structure (Buchanan et al., 2018). We sought to determine whether item randomization affects factor structure of meaning in life measures, given the differences that these measures show across individual measures and demographics more generally.

Meaning in Life

Broadly defined, meaning in life refers to an individual's understanding of his or her life, complemented by desired goals (Steger, Bundick, & Yeager, 2014). Meaning occurs through goal achievement, as well as positive interactions with others or the environment (Frankl, 1966, 1984). The concept of meaning is prevalent and related to positive outcomes across different contexts, including natural disaster survivors (Van Tongeren et al., 2018) and

those attending college for the first time (Bronk, Riches, & Mangan, 2018) among countless others (for a review, see Brandstetter, Baumann, Borasio, & Fegg, 2012). Individuals lacking meaning, rather, are at increased risk for developing various forms of psychopathology, such as depression (Beck, 1967) and anxiety (Bernard et al., 2017). Given the relavence of meaning to well-being and psychological suffering, understanding how to measure meaning in life is crucial to intervention and research efforts. Empirically-validated measures of meaning in life will help researchers to better understand the construct itself, as well as how it relates to treatment outcome. More generally, we also sought to demonstrate how item randomization could potential affect the factor structures of measures in a single domain.

The concept of meaning in life is measured in different ways, and coming to conclusions regarding meaning in life literature can be difficult (Cosco et al., 2017). Within the 21st century alone, Hicks and Routledge (2013) catalogued 59 meaning in life measures targeting different domains (e.g., search for meaning, breadth, depth) and factor structures. For example, Ryff and Singer (1998) conceptualize meaning as dedicating time towards personal goals, while Battista and Almond (1973) consider meaning to be related to a sense of coherence or understanding. Still, others consider meaning in life as life significance (Crumbaugh & Maholick, 1964), while Steger, Frazier, Oishi, and Kaler (2006) posit meaning to have two components (i.e., Presence of Meaning and Search for Meaning). Demographically, meaning also varies across gender (Chamberlain & Zika, 1998). Recently, researchers have argued as to whether meaning should be differentiated from the concept of purpose (Martela & Steger, 2016). Bronk, Riches, and Mangan (2018) consider meaning to be a component of purpose. Meaning in life measures have ultimately been criticized for their lack of objectivity (Dyck, 1987; Frazier, Oishi, & Steger, 2003; Garfield, 1973; Klinger, 1977; Steger et al., 2006; Yalom, 1980) and heterogeneity. Understanding how randomization affects factor structure within the context of meaning in life will allow researchers to understand the influence of randomization on a specific aspect of scale psychometrics.

Multi-Group Confirmatory Factor Analysis

erin going to leave this for you because of the new stats we are doing. this section isn't scienced yet

While meaning in life measures offer an interesting opportunity to determine whether measures yield different results across demographics and administration modalities, this topic is of interest to researchers more generally. For example, how is the factor structure of a measure supported across modalities? What about across demographically-diverse samples? Multi-group Confirmatory Factor Analysis (MGCFA) applies principles from Confirmatory Factor Analysis to different groups across a measure of interest. A careful examination of measurement variance can allow the researcher to determine whether the measure yield similar attributions across groups (Beajean, 2014). Typically, MGCFAs are conducted in a stepwise approach, first by examining configural invariance. Also referred for as "equal form invariance," configural invariance afford the researcher an opportunity to statistically determine whether factor loadings are indential across specified groups (e.g., male vs. female). Then, metric invariance determines whether there is a significant difference. Assuming our model does not significantly differ in CFI after examining metric invariance (CFI difference must be less $< .01$), researchers generally conduct an analysis of scalar invariance. Scalar invariance examines indicator intercepts and determines whether or not these are equal across groups. Additionally, scale invariance determines whether or not group membership influences raw scores. Our analyses present noninvariant questions across random/non-random groups so that researchers are better able to decide whether or not it would be prudent to randomize question order. This section presents a summary of MGCFA and the rationale for its usage. It is important to conduct this type of analysis in order to improve construct validity.

Goals of Present Study

We sought to conduct MGCFA on a variety of meaning in life measures and examine measurement invariance acorss these measures. A more collectivist view of meaning in life measures can facilitate sound assessment tools for clinicians and researchers and also allows researchers to understand how item randomization can factor structure. Specifically, we utilized a stepwise approach to examine model fit across all groups (both random and non-random). Then, data were into both random and non-random data in order to examine model fit across these individual groups. Each group provides the researcher with a set of fit indices by which to examine model fit. The rationale for conducting such an analysis is to discover whether or not each group produces less than desired fit statistics. With this information, the researcher is able to tell whether or not different group reports differently on the given scale. Regardless of model fit, we continued with the suggested stepwise approach by calculating different types of invariances.

Method

Participants

Participants in this study included 2377 students at a large Midwestern university. Participants included 924 males (39.81%) and 1397 females (60.19%) between the ages of 15 and 55 ($M = 19.57$, $SD = 3.20$). The study included multiple different ethnicities, made up of Asian participants (2.67%), Black participants (6.73%), multiple ethnicity participants (2.40%), other ethnicity participants (4.94%), and White participants (83.26%). The study also consisted of Freshmen (68.08%), Sophomores (18.86%), Juniors (9.11%), Seniors (3.57%), and Graduate/Other students (0.39%).

Materials

need someone to make a table of the scales we used

Procedure

Data analysis

Data Screening.

MGCFA - jeff wrote this but probably wrong with new model leaving for Erin. Multigroup Confirmatory Factor Analysis (MG-CFA) was conducted on individual meaning in life scales. This particular process involves applying CFA principles to multiple groups across different each individual scale. Delivery type (non-random vs. random) was used to examine model fit and whether or not randomization of scales produces a worse or better-fitting model. We utilized previously published standards for adding restrictions to each MG-CFA. This approach allowed us to first examine model fit across all groups. Subsequently, model fit across non-random and random groups was examined. Then, parameters were constrained in order to calculate different types of invariances.

Individual Groups -jeff wrote this but probably wrong with new model leaving for Erin.

Utilizing a stepwise approach allowed us to examine model fit across individual groups by means of MG-CFA. We conducted single-group solutions based on delivery method (non-random question order vs. random question order). Questions delivered on paper were excluded for final analysis in R, as they were not part of this particular analysis. Each group provided us with a set of fit indices by which to evaluate model fit and examine whether or not scale randomization impacts factor structure across each different scale. Randomized

scales not adhering to the published factor structure should warrant caution among researchers planning to deliver questions in a random format. Randomized scales adhering to published factor structure do not suggest any reason to avoid randomization of questions (Brown citation). Regardless of fit, we continued with the suggested stepwise approach by calculating different types of invariances. Each level of invariance adds restrictions to the model.

Configural Invariance -jeff wrote this but probably wrong with new model leaving for Erin.

Regardless of whether or not our individualized groups both showed adequate model fit, we progressed to calculate configural invariance. Configural invariance can also be referred to as "equal form." This test allows the researcher to understand whether or not factor structure and loadings are identical across groups, in this case non-random questionnaires vs. random questionnaires. This test utilizes the same set of fit indices explained above (assuming we will add this section in the data analysis section/insert a citatio).

Metric Invariance -jeff wrote this but probably wrong with new model leaving for Erin.

Regardless of whether or not equal forms was supported across groups, we then analyzed the data using metric invariance. Metric invariance examines factor loadings across groups. This analysis was supported if this test of invariance did not differ significantly from configural invariance. In order to meet this assumption, $\Delta\text{CFI} < .01$.

Scalar Invariance -jeff wrote this but probably wrong with new model leaving for Erin.

Assuming that metric invariance did provide a large enough decrease in CFI, we then tested scalar invariance. Scalar invariance examines indicator intercepts and determines

whether or not these are equal across groups. Additionally, scalar invariance determines whether or not group membership influences a role in raw scores across groups. If the change in CFI is not equal to or greater than .01, this assumption has been met. As with metric invariance, this analysis was supported if the test of invariance did not differ significantly from configural invariance.

Partial Invariance =jeff wrote this but probably wrong with new model leaving for Erin.

Different methods have been utilized for scales that differ when utilizing the stepwise method for conducting the different types of invariances. The scale can either be abandoned or the noninvariant items removed for further analyses. This may affect construct validity as well as the theory behind the scale (Cheung & Rensvold, 1999). We relaxes constructs of noninvariant items for the remainder of analysis, as suggested by Brown (2006) & Byrne et al. (1989).

Results

Discussion

References

Buchanan, T. (2002). Online assessment: Desirable or dangerous? *Professional Psychology: Research and Practice*, *33*(2), 148–154. doi:10.1037/0735-7028.33.2.148

Buchanan, T. (2002). Online assessment: Desirable or dangerous? *Professional Psychology: Research and Practice*, *33*(2), 148–154. doi:10.1037/0735-7028.33.2.148