**Scale Development**

Scale development typically begins with an underlying latent construct of interest to the researcher (DeVellis citation). (Olson 2010 citation) suggests compiling a list of items thought to represent the construct of interest. These items are generally evaluated by experts in the specified area. This concept is termed as construct or face validity (citation on construct or face validity). After eliminating problematic questions, data are collected and examined through classical test methods, specifically exploratory and confirmatory factor analysis (EFA; CFA; Worthington citation) or Item Response Theory methods (cite any IRT article or probably a seminal IRT article). Generally, EFA and CFA examine correlations between items in regards to how they relate to a latent construct, also termed factor. Specific items will sometimes correlate to form multiple factors, or psychological constructs, depending on the measure of interest. For example, the DASS (cite DASS or literally any other scale with multiple factors) will lead to clusters of items correlated with different underlying constructs (depression, anxiety, stress). Unfortunately, researchers will sometimes make questionable decisions in utilizing factor-analytic methods that may lead to misinterpreted measures and drastic clinical implications (mainly, an inability of a measure to actually measure what it is supposed to measure, which could lead to misdiagnosed or incorrectly diagnosed patients. I’m sure that there are other clinical implications too, but I think that meaning in life measures are generally used as a supplement to therapy. If we use measures as a supplement to therapy, like meaning in life measures, then they could inhibit our ability to enhance therapy). Specifically, Preacher & MacMallum (2002 I think – check year) showcase common mistakes made in conducting factor analyses and how they could be avoided. Specifically, researchers will mistakenly use Principal Components Analysis (PCA) and will keep eigenvalues that are greater than 1. Additionally, they use varimax rotation (Kaiser, 1970: explain why using these things is problematic to research. Need to look into this more because I honestly don’t know much about the different types of rotations). Initially, EFA examines factor loadings in order to examine how well items correlate with a latent construct(s). When adequate factor loadings (> .300) have been achieved for each item, CFA assesses model fit by means of fit statistics (cite whoever came up with the .300 criteria). Scales demonstrating adequate model fit are then published for use in psychological assessment. Obviously, scale development is crucial and of utmost importance in accurately measuring constructs of interest.

**Multi-Group Confirmatory Factor Analysis**

A topic of interest to many researchers is whether or not measures yield different results when conducting factor analyses across groups. Generally, it is necessary to examine whether or not the psychometrically-supported factor structure holds across different groups of interest. This is done by examining measurement invariance, which tells us whether or not our measure will yield the same attributes across such groups (Beaujean, 2014). Multi-group Confirmatory Factor Analysis (MGCFA) applies CFA principles to different groups (usually two – more than two would warrant a different type of analyses such as MIMIC) across a measure or measures of interest. As an example of a type of multi-group analysis, this study sought to examine scale-delivery type (random vs. not random) in order to conclude whether or not randomization of different meaning in life measures produces different results in regards to factor structure of said scales. Specifically, we initially 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. Each level of invariance adds certain restrictions to the model.

Configural invariance, also referred to as “equal form invariance,” allows the researcher to understand whether or not factor structure/loadings are identical across specified groups. Regardless of whether or not equal forms was supported across groups, it is then suggested to analyze data using metric invariance. Metric invariance, too, examines factor loadings across groups. This analysis is supported if this test of invariance did not differ significantly from configural invariance. (CFI difference must be less < .01). Assuming our model does not significantly differ in CFI after examining metric invariance, 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. As with metric invariance, if the change in CFI is less than .01, this assumption has been met. Finally, assuming that the metric invariance assumption has been met, researchers will then examine partial invariance. If the model breaks down, researchers will generally debate on whether or not to abandon the measure or remove invariant items for further analyses, as noninvariant items may affect construct validity as well as theory behind the measure of interest (Cheung & Rensvold, 1999). 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.

**Meaning in Life**

Individuals generally experience at least one traumatic event throughout the course of their lifetime (Breslau, Peterson, Poisson, Schultz, & Lucia, 2004; Copeland, Keeler, Angold, & Costello, 2007; Kessler, Sonnega, Bromet, Hughes, & Nelson, 1995). As such, a topic of interest to many researchers is examining responses to traumatic events in order to examine variability in response and clinical implications. For example, some individuals will develop clinical disorders such as Posttrauamtic Stress Disorder (PTSD; Hamblen, Barnett, & Norris, 2012; Williams, McDevitt-Murphy, Fields, Weathers, & Flood, 2011).

According to Frankl (1984), meaning in life occurs when an individual accomplishes personal goals, has positive interactions with others, or experiences meaningful interactions with art and/or nature. What, then, is meaning in life and why is it important? Ryff & Singer (1998) consider meaning to be the dedication of time to achieving personal goals of importance. Battista & Almond (1973) consider meaning in life to be related to a sense of “coherence.” Other definitions, such as that of Crumbaugh & Maholick (1964), define meaning in life as “the ontological significance of life from the point of view of the experiencing individual.” As evidenced by the numerous definitions presented, a coherent, collectivistic definition of meaning in life seems to be lacking. This is not to say, however, that the construct of meaning is unimportant (Steger et al., 2006). Naturally, the variety of definitions lead researchers to measure the construct in different ways.

Researchers have examined the construct of meaning in life in various ways, as evidenced by opposing definitions mentioned above and the great number of scales factor analyzed in this study. Typically, questionnaires of this nature rely on a subjective participant understanding of the construct. The Meaning in Life Questionnaire (MLQ citation), for example, asks various questions regarding participant perception of whether or not their life has meaning. As an example, one question on the MLQ posits the following statement: “I understand my life’s meaning.” Although this type of question seems transparent, a participant/client who does not have an understanding of a basic definition(s) of meaning in life will struggle to correctly interpret the construct. As a result, meaning in life measures have been criticized for their lack of objectivity in soliciting participant/client responses (Dyck, 1987; Frazier, Oishi, & Steger, 2003; Garfield, 1973; Klinger, 1977; Yalom, 1980). Coming to conclusions from existing meaning in life literature, therefore, can present a difficult challenge to researchers (Cosco et al., 2017). Additionally, meaning-based measures tend to highly correlate with other empirically-supported constructs. For example, Clark & Watson (1995) found that examined items relating to suicide correlated with neuroticism, which may provide evidence for the lack of uniqueness of questions on meaning in life scales. Unfortunately, however, this stems from an inability to answer fundamental questions regarding a collective definition of meaning in life (Steger et al., 2006). Further, Steger and colleagues argue that more effective measurements and empirically validated will better help researchers to understand such constructs, specifically, meaning in life.

In examining factor structures of meaning in life measures, Chamberlain & Zika (1988) analyzed the factor structures of three meaning in life measures: the Purpose in Life test (PIL), the Life Regard Index (LRI), and the Sense of Coherence Scale (SOC). Specifically, this study recruited women with young children in an attempt to examine factor structures of the three aforementioned measures. Analyses did not suggest an underlying general factor. However, an important limitation in this study that needs to be addressed is possible differences across groups in measuring meaning in life (might go into more detail on the results from factor structure and how they contribute to why we are doing what we’re doing with randomization). Chamberlain & Zika (1988) claim that meaning in life may be different for women. Then, they examine different studies conducted that have looked at gender-based differences (PIL; Meier and Edwards, 1974; Reker and Cousins, 1979; Shape and Viney, 1973). It is necessary to further examine group-based differences to identify poor items in meaning in life measures.

**Scale Randomization**

Measurement invariance posits that different modalities of administration in different settings can lead to comparable responses, which is desirable in order to negate confounding of variables (Brown, 2006). Although a multitude of studies have been conducted to examine administration modalities and their effects on responses (T. Buchanan et al., 2006; Meyerson & Tryon, 2003; Ruyter & Wetzels, 2006), an area of concern resulting from these different administration modalities is their effect on factor structure. According to T. Buchanan et al. (2005), factor structure is not replicable in online vs. mailed surveys. As a result of these replicability issues, researchers tend to randomize question orders in order to control for item order effects, as randomizing the item order can counterbalance the scales. This ultimately controls for order effects (Keppel & Wickens, 2004). However, despite controlling for order effects, randomizing items on some measures could lead to changes in scale structure, as some items (suicide questions or death questions) could possibly lead to different answers on subsequent items. Our study sought to determine whether or not the solution of randomizing items to combat the issue of replicability is problematic in and of itself. Meaning in life measures, as discussed above, have the potential to lead to subjective participant responses. Thus, this construct was of interest to the present study. Additionally, these measures tend to be highly correlated with other constructs. For example, as discussed previously, questions related to suicide tend to correlate with neuroticism. These questions may lead to different-than-normal responses on subsequent questions, and this “item reactivity” concept is currently being explored in a separate project (Buchanan et al, 2017). However, an area of concern, aside from differing item responses, is factor structure. Thus, this study sought to examine the effects of scale randomization on across a wide variety of meaning in life measures at the factor level.

**Goals of Present Study**

Our goal for this study is to conduct a MGCFA on a wide variety of meaning in life scales to examine measurement invariance across these scales. We categorized each measure based on model fit by means of fit indices (cite the fit indices used here). Additionally, we hoped to present noninvariant questions for each scale. As stated above, a more collectivistic view of meaning in life measures is needed in regards to scale randomization (Put in more information here on how we categorized these scales and how we presented noninvariant questions).