THE MEANING IN LIFE QUESTIONNAIRE: A RANDOM EFFECTS RELIABILITY GENERALIZATION

A Thesis

by

BRANDIE SEMMA

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Wen Luo
Committee Members,	Oi-Man Kwok
	Lei-Shih Chen
Head of Department,	Shanna Hagan-Burke

August 2017

Major Subject: Educational Psychology

ABSTRACT

The Meaning in Life Questionnaire (MLQ) is a 10-item self-report measure that assesses perceived meaning in life and search for meaning in life (Steger, Frazier, Oishi, & Kaler, 2006). A reliability generalization was conducted on the Presence and Search subscales of the MLQ to estimate the average reliability, examine the variability among the reliability estimates, and search for moderators. Articles that meet selection criteria were obtained from PsycINFO. Mixed effects analysis was conducted on 152 reliability estimates for the Presence subscale, and 89 reliability estimates for the Search subscale.

Both Presence and Search subscales showed high mean reliability estimates (above .85) and significant heterogeneity in estimates across studies. Language in which scale was administered explained a significant proportion of the variation in the Presence and Search subscales. Region was a significant moderator of reliability of the Search subscale only; regions outside of North America and Europe reported significantly lower reliability. Results indicate that even though the MLQ is highly reliable across samples, reliability varies significantly as a function of language and region. The results have implications of the conceptualization of meaning in life across diverse cultures.

DEDICATION

I dedicate this thesis to my amazing family and supportive friends.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Wen Luo, and my committee members, Dr. Oi-Man Kwok, and Dr. Lei-Sheih Chen. I would also like to thank my amazing EPSY friends, Maria, Lucy, and Mirim, specifically. Finally, thanks to my lovely parents, and my brothers and sister.

CONTRIBUTORS AND FUNDING SOURCES

This work was supported by a committee consisting of Professor Wen Luo, Professor Oi-Man Kwok, Professor Lei-Shih Chen, and Shanna Hagan-Burke of the Department of Educational Psychology.

Maria Henri helped with the coding of the articles All other work conducted for the thesis was completed by the student independently. There are no outside funding contributions to acknowledge related to the research and compilation of this document.

TABLE OF CONTENTS

1. INTRODUCTION: MEANING IN LIFE

Meaning in life is a cornerstone of well-being. It is not a fixed concept, but a dynamic one that transforms according to experience (King, Heintzelman, & Ward, 2016; Steger, Frazier, Oishi, & Kaler, 2006). Meaning in life can be defined as feeling of purposefulness in one's life or a seeking of activities that will contribute to a sense of purpose or coherence to one's life (Reker & Wong, 1988). According to King et al. (2016) the concept of meaning in life has three essential components. The first component is *purpose*, in which individuals have goals and directions. The second component is *significance,* or the amount of value, worth, and importance a person feels they have in their lives. The last component, *coherence*, is characterized by predictability and routine. In other words, coherence occurs when life makes sense to the person living it.

Based on global research, meaning in life seems to be a universal construct, but individual and cultural differences exist in the achievement and expression of meaning in one's life (Frankl, 1965). Additionally, the definition of meaning in life varies; however, what they all have in common is the assumption that life has meaning or purpose. (Steger et al., 2006). Meaning in life has shown to contribute to subjective well-being, and protect against problems including depression, anxiety, poor academic performance, drug and alcohol abuse, negative affect, and other indicators of subjective well-being (Grotberg, 2003).

One of the most commonly used and well-vetted measures for meaning in life is the Meaning in Life Questionnaire (ML; Steger et al., 2006). The MLQ has been widely used over the past decade in multiple disciplines including counseling and clinical psychology, health sciences, occupational sciences, and social work (Schulenberg, Strack, & Buchanan, 2011). Consequently, the psychometric properties of the MLQ, including reliability, have been extensively investigated. Despite its widespread use, no studies have empirically synthesized the reliability estimates and investigated moderators that explain variability in reliability estimates for this measure.

The purpose of this study is to conduct a reliability generalization on the Cronbach's alpha coefficient of the two subscales of the MLQ. This analysis will examine the variability in the reliability estimate and identify potential moderators, such as study characteristics, that may explain the variability.

The research questions explored in this study include:

1) What is the average internal consistency reliability coefficient of the two subscales in the MLQ?

2) What moderators explain the variability in the reliability estimates?

2. REVIEW OF LITERATURE

The MLQ contains two subscales measuring the Search for meaning and the Presence of meaning. The subscales are only weakly correlated. Each subscale is comprised of five items, and a 7-point Likert response scale, ranging from "absolutely true," to "absolutely untrue." The scores on each subscale can vary from 5 to 35, with higher scores indicating higher presence and search. The Presence subscale includes items such as "I understand my life's meaning" and "My life has a clear sense of purpose". The Search subscale assesses motivation to discover meaning in life. Example items include, "I am searching for meaning in my life" and "I am seeking a purpose or mission for my life". The Presence subscale has been found to be significantly correlated with many well-being, personality, and religiosity variables, such as positive emotions, extraversion, and agreeableness. The Search subscale is positively correlated with depression, negative emotions, and neuroticism (Steger et al., 2006).

The MLQ scores exhibit greater structural stability and better discriminant validity than scores of other popular meaning measures, such as the Purpose in Life Test (Crumbaugh & Maholick, 1964) and the Life Regard Index (Battista & Almond, 1973). The stability of the MLQ has been demonstrated in samples over periods of 2 weeks (Steger et al., 2006), 1 month (Steger, Kawabata, Shimai, & Otake, 2008) and 1 Year (Steger & Kashdan, 2007). Furthermore, a recent review of research has shown the MLQ has demonstrated good internal consistency with alphas ranging between the low .80s and low .90s (Schulenberg et al., 2011).

Research has expanded to include greater diversity in samples and nationality, including clinical samples, veteran samples, and college student samples (Church et al., 2013; Schulenberg et al., 2011). The MLQ has been translated into numerous languages, including Turkish, Mandarin, Portuguese, and Hungarian (Boyraz, Jr, & Can, 2013; Brassai, Piko, & Steger, 2012; Chan, 2014). Given this recent expansion in the global use of the scale, it is essential to assess the reliability of the scale, and to examine what sample characteristics associate with the variance among the alpha estimates.

Reliability, the stability of test scores over repeated administration (Traub & Rowley, 1991), is a property of the data, not the measure, and varies across samples (Thompson & Vacha-Haase, 2000). Reliability represents the ratio of true score to observed score variance (Hess, McNab, & Basoglu, 2014). Classical test theory assumes every participant has a "true" score that is obtained if the measurement does not have any error (Henson & Thompson, 2002). However, instruments are not perfect, and the observed score will differ from true score due to error. Reliability assesses the variance in observed scores attributed to the actual variance in true score.

The error of a score is associated with the inconsistency of the instrument. In this context, measurement error is defined as random variation. This type of variation is influenced by errors in guessing, administration, and scoring, and is considered to be part of the unreliability of a measurement. Consequently, measurement error impacts effect size. This is because measurement error changes the observed effects across studies, and it leads to underestimation of the effects (Baugh, 2002).

Reliability scores range from 0 to 1, with 0 indicating all variability in the observed scores caused by error, and 1 is interpreted as perfect reliability and no error variance. There is no rule regarding the acceptability of a specific coefficient, however current standards indicate a minimum score cutoff of .70, and a maximum of .95. Above .95 may indicate item redundancy in internal consistency. (Panayides, 2013).

One of the most widely used and accepted ways of assessing reliability is through the examination of internal consistency. This refers to the intercorrelations, or homogeneity, between items on the same instrument (Kaplan & Saccuzzo, 2012). Multiple types of internal consistency exist, including Cronbach's alpha, split-half reliability, and Kuder-Richardson test. Cronbach's alpha is one of the most frequently used and conservative estimates of reliability (Dimitrov, 2002). Cronbach's alpha can be defined as:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_i V_i}{V_t}\right)$$

where α is the estimate of reliability, n is the number of items in the instrument, V_i is the variance of item scores after weighting, and V_t is the variance of instrument scores (Cronbach, 1951). An advantage of internal consistency is that the measure only needs to be given once (Thompson, 2002). Alpha is affected by the characteristics of the participants, which influences the total score variance of an instrument (Cronbach, 1951).

Given that reliability fluctuates from one application of the test to another, it is necessary to examine the performance of scores from a number of occasions. Reliability generalization (RG) is a meta-analytic method developed by (Vacha-Haase, 1998) in order to characterize the typical reliability for a test, assess the amount of variability in reliability for a given test across studies, and examine the sources of variability in the reliability coefficient(s) across studies for a given measure. This technique has been used across multiple disciplines using various statistical techniques (Vacha-Haase & Thompson, 2011).

An RG analysis aids researchers in understanding the differences in score reliability across multiple studies. RG analysis is necessary because if reliability is poor then the ability to measure the construct leads to weaker validity, and the attenuation of effect sizes (Thompson, 2002). Using RG analysis, researchers can determine what factors lead to higher reliability estimates by examining sample characteristics, instrument forms, and circumstances under which an instrument is taken. Many variables can influence reliability such as gender, sample type, and language (Thompson, 2002). By examining these factors, we can better understand what contributes to the variability in score quality. For example, Yin & Fan (2000) explain that with other factors being equal, a heterogeneous group of participants will produce higher score reliability than a homogenous group of participants.

RG studies are useful in many ways. They can alter the way researchers think about reliability issues (Henson & Thompson, 2002). They can also help researchers improve their understanding of instruments and what type of information should be gathered (i.e. study relevant futures and expected level of outcome reporting). Maximizing score reliability allows researchers to have additional control over other factors that may influence effect sizes.

This RG investigation is warranted for several reasons. First, the results of the moderation analysis can inform future research of what ranges of reliability estimates can be expected based on the characteristics of the sample. Second, knowing what sample or scale (such as language) characteristics explain variability in reliability estimates may in turn shed light on the interpretation of meaning in life. It may be the case that the measure's items have different meaning for different samples or languages. Third, calculating confidence intervals provides researchers with the limitations of the point estimate. Lastly, recommendation on the importance of reporting reliability estimates and confidence intervals will be offered.

3. METHOD

3.1 Selection Criteria

In order to obtain studies that utilize the MLQ, a search using the keywords [Meaning in Life Questionnaire] and [Reliability OR internal consistency OR Cronbach's alpha] using the EBSCO database was performed. The results were then filtered so that only peer reviewed journal articles that were published in the last 10 years, 2006-2016 (May) were included. The MLQ was published in 2006, for this reason articles published prior to 2006 were excluded. Additionally, articles that were not available in English were excluded. Two hundred thirty-nine (239) unique number of articles remained.

Further inclusion criteria were used for this study. Articles were included if they provided Cronbach's alpha for at least one subscale for the study's sample, and included some information about the sample such as number of observations, mean score on the subscale(s) and standard deviation (articles that contained partial descriptives were included). Ninety-two articles were excluded because they did not administer the MLQ; these articles included literature reviews, commentary/recommendations papers, and studies that incorporated some of the items from the MLQ into a larger scale (did not treat the MLQ items as a separate scale). Thirty-one were excluded for not reporting Cronbach's alpha of the study's sample, while one was excluded because we were unable to obtain the full text. Out of the 239 articles, only 115 met the inclusion criteria.

3.2 Coding Procedure

According to Vacha-Haase & Thompson (2011) it is expected that reliability will be affected by variables such as gender, sample size, age, and ethnicity. To examine possible relationships and moderators, the following study characteristics were coded:

1) Region
2) Mean age
3) Percent female
4) Sample type
5) Sample size
6) Language

Additionally, data was organized in a multilevel structure, with samples nested in studies, studies nested in articles, and articles nested by first author. Studies that did not report some of the information, such as means and standard deviations, were still coded and absent information were treated as missing data. Inter-rater reliability will be calculated as percent agreement for a random sample of the studies included in the analyses.

3.3 Data Analysis

Reliability generalizations can be examined using multiple methods. For this present study, a mixed-effects model will be used. A mixed-effects model is a random-effects model (RE) that includes moderator variables as predictors. RE method assumes that there is an underlying population of alpha coefficient obtained from a population of studies, from which a sample has been drawn. The variability in the sample of coefficient alpha has two sources:

1) The within-study variability, the variability due to the sampling of the

participants in the sample;

2) The between-study variability, assumes that a random sample of studies and coefficient alpha was obtained from a greater population.

The assumption under the RE model, $\theta i \sim N(\mu, \tau^2)$, states that the unknown true effect in the outcomes in the population (θ) follows a normal distribution with mean (μ) and variances (τ^2). The mixed-effects model can be seen as

$$y_i=\beta_0+\beta_1 x_{i1}+ \beta_2 x_{i2}+\ldots+\mu_i+e_i,$$

where μ_i is approximately normal with a mean of 0 and variance τ^2 (i.e. $\mu_i \sim N(0, \tau^2)$), and e_i is approximately normal with a mean of 0 and variance v_i (i.e. $e_i \sim N(0, v_i)$). The amount of residual heterogeneity in the true effects can be accounted for by τ^2 (Viechtbauer, 2010). In other words, τ^2 is the amount of variability in the true effects that is not accounted for by the moderators in the model.

The purpose of the analysis is to estimate the mean (u) and variance (τ^2) of the population of alpha coefficients, construct confidence intervals for each alpha coefficient estimate, and determine moderator effects. In order to do so under the assumptions of an RE model, it is recommended that the alpha coefficients be transformed (Sanchez-Mecca, Lopez-Lopez, & Lopez-Pina, 2013). Alpha values will be transformed by using the Hakstian-Whalen transformation method (Aguayo, Vargas, de la Fuente, & Lozano, 2011; Hakstian & Whalen, 1976). This method normalizes the reliability estimates by transforming them into t-scores:

$$\hat{T}_i = \sqrt[3]{1-\widehat{a}_i}$$

with sampling variance of $V(\hat{T}_i) = \frac{18j_i(n_i-1)(1-\hat{\alpha}_i)^{2/3}}{(J_i-1)(9n-11)^2}$, where *J* represents the number of items in the scale, *n* is the study sample size, and α is the reliability estimate.

Under the mixed-effects model, study characteristics become estimate as fixed-effects variables. Additionally, the model requires modifying the weights, defined as

$$\frac{1}{V(y_i) + \hat{T}^2_{Res}}$$

which can be read as the inverse of the sum of the study variance, $V(y_i)$, and the estimate of the residual between studies variance, $\hat{T}^2_{Res}$

DerSimonian & Kacker (2007) outline a method of moments for estimating the between studies variance:

$$\hat{T}^2_{Res} \frac{y'Py - (k - p - 1)}{tr(P)}$$

where y is a vector of k x 1 transformed coefficients, tr(P) is the trace of P, and P is defined as $W-WX(X'WX)^{-1}X'W$, with W as a diagonal k x k matrix with elements from the inverse of the sampling variance of y_i (i.e. the fixed effects weight) (Stephen Raudenbush, 2009). Additional variance components will be estimated when running the model.

Moderators will be analyzed by applying a linear mixed-effects regression model. Lastly, the average reliability estimate (T) will be transformed back to alpha.

4. RESULTS

A total of 152 studies with a combined sample size of 78,673 reported coefficient alpha for the Presence subscale, and 89 studies with a combined sample size of 44,683 reported coefficient alpha for the Search subscale. Means and standard deviations were available for only 152 studies of the presence and 90 studies for the Search. Table 1 shows the descriptive statistics for the reliability coefficients of each subscale. Inter-rater reliability was 90%.

A random-effects model was specified to examine whether there is significant amount of variation between studies. Alpha estimates were adequate for the subscales, Presence and Search, .87 and .88. Table 2 provides the alpha estimates along with their confidence intervals. It was found that for both subscales, there was significant heterogeneity in alpha estimated between studies $Q(145) = 4111.6172$, $p < .0001$ for the Presence subscale, and $Q(82) = 2326.8758$, $p < .0001$ for the Search subscale. The percentage of total variation (I^2) across the studies due to heterogeneity rather than chance also indicated significant heterogeneity for Presence and Search, 96.22% and 95.68%. Figures 1 and 2, funnel plots of the observed coefficient alphas against the standard error, also indicate which studies fall outside of the specified heterogeneity.

Moderators for both subscales included average age, proportion of female, language, geographical region, and sample type. Region was dummy coded with North America as the reference group, and Europe and Other as the comparison groups. The mixed-effects model for the Presence subscale indicates .0031 ($SE = 0.0005$) amount of variability in the true effects is not accounted for by the moderators in the model, and I^2

= 96.63% which indicates significant heterogeneity. Additionally, r^2 = 5.86%, which is the amount of amount of heterogeneity accounted for by the mixed-effects model. Table 3 shows the results of the mixed-effects model for the Presence subscale. Language significantly predicted alpha scores, b = -.0423, z = -2.7730, $p < .01$. No other significant predictors were found.

The mixed-effects model for the Search subscale shows .0030 (*SE* = 0.0006) amount of variability in the true effects is not accounted for by the moderators in the model. Significant heterogeneity was indicated, I^2 = 94.32%, while the amount of heterogeneity accounted for by the model (r^2) was 4.48%. Table 4 shows the results of the mixed-effects model. Language was also a significant predictor in alpha scores for the Search subscale, b = -.0356, z = -1.9975, $p < .05$. Additionally, Other Region significantly predicted alpha scores, b = -.0366, z = -2.0094 $p < .05$. No other significant predictors were found.

Overall it was found language predicted alpha estimates in both the Presence and Search Subscale. For both subscales, languages other than English had lower reliability estimates. Additionally, compared to North America, Other Regions had significantly lower estimates in the Search subscale. However, as indicated by the I^2 values, a significant amount of heterogeneity still existed in the model, with the moderators in this model predicting less than 6% of the heterogeneity in both subscales.

5. CONCLUSIONS

The goals of this reliability generalization study were: 1) to provide a typical score reliability estimate for the Presence and Search subscale of the MLQ and 95% confidence intervals, 2) determine is there is significant heterogeneity in reliability estimates beyond what can be explained by sampling variability and 3) to determine if sample characteristics contribute to the score variability. Overall, it was found the MLQ was a very reliable scale, with Cronbach's alpha values greater than .85 for both subscales and narrow 95% confidence intervals. Given the results showed significant heterogeneity among the reliability estimates, several moderators were examined to determine whether they could explain the variability. It was found the reliability coefficients for both subscales differed as a function of language, non-English versions had lower reliability estimates for the Presence and Search subscale. The reliability estimates for the Search subscale were lower in regions other than North America and Europe.

The findings have many practical implications. First, the results provide evidence for the strength of the reliability in the MLQ subscales across diverse samples and various translations. Second, it is important for researchers to be aware that although the MLQ tends to have high reliability estimates, language spoken and specific regions moderates the strength of the reliability estimate. It may be the case the quality of translations from English to other languages needs to be examined. Not all the articles reviewed utilized translated versions reported whether back translation was done. Additionally, researchers should be aware of this implication when sampling outside of

Europe and the North America. The conceptualization meaning in life may be dependent on the cultural context. Although the wording of the MLQ seems very straightforward, when administered in diverse cultures, the items may have different meanings. The differences in interpretation of meaning in life are consistent with Frankl's argument (1965) that there are individual and cultural differences in the achievement and expression of meaning in one's life.

However, this study does have a few limitations. First, the mixed-effects model for both subscales did not fully explain the heterogeneity found in the samples. Each model explained less than 6% of the heterogeneity in the sample for both subscales. There may be other potential moderators that explain the heterogeneity that

were not accounted for in this model. Second, only peer-reviewed articles available on PsycINFO were included. This limit the search to only studies completed in psychology and psychology-related disciplines. Typically research in psychology uses college student samples. This led to a third limitation in which most of the studies included administered the survey to college students.; the specialized comparison group was smaller than expected and may have negated any true difference in reliability estimate.

Future research should further examine the scale's reliability by including additional moderators that may explain the large heterogeneity found, such as GPD or collectivism/individualism scores. Additionally, a literature search in multiple databases should be conducted to find articles that have used the MLQ in more diverse samples. Overall, the RG analysis provided preliminary evidence of high reliability across diverse samples of both the Presence and Search subscales of the MLQ.

REFERENCES

Aaguayo, R., Vargas, C., de la Fuente, E.I., & Lozano, L.M. (2011). A meta-analytic reliability generalization study of the Maslach Burnout Inventory. *International Journal of Clinical and Health Psychology, 11*, 343-361.

Battista, J., & Almond, R. (1973). The Development of Meaning in Life. *Psychiatry*, *36*(4), 409–427. https://doi.org/10.1521/00332747.1973.11023774

Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement*, *62*(2), 254–263.

Boyraz, G., Jr, O. R. L., & Can, A. (2013). The Turkish version of the Meaning in Life Questionnaire: Assessing the measurement invariance across Turkish and American adult samples. *Journal of Personality Assessment*, *95*(4), 423–431. https://doi.org/10.1080/00223891.2013.765882

Brassai, L., Piko, B. F., & Steger, M. F. (2012). Existential attitudes and Eastern European adolescents' problem and health behaviors: Highlighting the role of the search for meaning in life. *The Psychological Record*, *62*(4), 719.

Chan, W. C. H. (2014). Factor Structure of the Chinese Version of the Meaning in Life Questionnaire among Hong Kong Chinese ccaregivers. *Health & Social Work*, hlu025. https://doi.org/10.1093/hsw/hlu025

Church, A. T., Katigbak, M. S., Locke, K. D., Zhang, H., Shen, J., de Jesus Vargas-Flores, J., … Ching, C. M. (2013). Need satisfaction and well-being: Testing self-determination theory in eight cultures. *Journal of Cross-Cultural*

Psychology, *44*(4), 507–534. https://doi.org/10.1177/0022022112466590

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Crumbaugh, J. C., & Maholick, L. T. (1964). An experimental study in existentialism: The psychometric approach to Frankl's concept of noogenic neurosis. *Journal of Clinical Psychology*, *20*(2), 200–207.

DerSimonian, R., & Kacker, R. (2007). Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials*, *28*(2), 105–114. https://doi.org/10.1016/j.cct.2006.04.004

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement*, *62*(5), 783–801. https://doi.org/10.1177/001316402236878

Frankl, V. E. (1965). The concept of man in logotherapy. *Journal of Existentialism*, *6*(21), 53–58.

Grotberg, E. H. (2003). Resilience and tragedy. Resilience for today: Gaining strength from adversity (pp. 241–264). Westport, CT.: Praeger Publishers/Greenwood Publishing Group.

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

Henson, R. K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "Reliability Generalization" studies. *Measurement & Evaluation in Counseling &*

Development (American Counseling Association), *35*(2), 113.

Hess, T. J., McNab, A. L., & Basoglu, K. A. (2014). Reliability generalization of perceived ease of use, perceived usefulness, and behavioral intentions. *MIS Quarterly*, *38*(1), 1-A29.

Kaplan, R. M., & Saccuzzo, D. P. (2012). *Psychological Testing: Principles, Applications, and Issues* (8 edition). Belmont, CA: Cengage Learning.

King, L. A., Heintzelman, S. J., & Ward, S. J. (2016). Beyond the search for meaning A contemporary science of the experience of meaning in life. *Current Directions in Psychological Science*, *25*(4), 211–216. https://doi.org/10.1177/0963721416656354

Panayides, P. (2013). Coefficient alpha: Interpret with caution. *Europe's Journal of Psychology*, *9*(4), 687–696. https://doi.org/10.5964/ejop.v9i4.653

Reker, G. T., & Wong, P. T. P. (1988). Aging as an individual process: Toward a theory of personal meaning. In *Emergent theories of aging* (pp. 214–246). New York, NY, US: Springer Publishing Co.

Schulenberg, S. E., Strack, K. M., & Buchanan, E. M. (2011). The meaning in life questionnaire: psychometric properties with individuals with serious mental illness in an inpatient setting. *Journal of Clinical Psychology*, *67*(12), 1210–1219. https://doi.org/10.1002/jclp.20841

Steger, M. F., Frazier, P., Oishi, S., & Kaler, M. (2006). The Meaning in Life Questionnaire: Assessing the presence of and search for meaning in life. *Journal of Counseling Psychology*, *53*(1), 80–93. https://doi.org/10.1037/0022-

0167.53.1.80

Steger, M. F., & Kashdan, T. B. (2007). Stability and specificity of meaning in life and life satisfaction over one year. *Journal of Happiness Studies*, *8*(2), 161–179. https://doi.org/10.1007/s10902-006-9011-8

Steger, M. F., Kawabata, Y., Shimai, S., & Otake, K. (2008). The meaningful life in Japan and the United States: Levels and correlates of meaning in life. *Journal of Research in Personality*, *42*(3), 660–678. https://doi.org/10.1016/j.jrp.2007.09.003

Sterne, J.A., Sutton, A.J., Loannidis, J.P.A., Terrin, N., Jones, D.R., Lau, J.,...&Higgins, J.P.T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomized controlled trials. *The BMJ*, 343,

Raudenbush, S. (2009). Analyzing effect sizes: Random effects models. In *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 295–315). New York: Russell Sage Foundation.

Terrin, N., Schmid, C.H., & Lau, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology, (58)* 894-901

Thompson, B. (2002). *Score Reliability: Contemporary Thinking on Reliability Issues*. SAGE Publications.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not Reliable. *Educational and Psychological Measurement*, *60*(2), 174–195. https://doi.org/10.1177/0013164400602002

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice*, *10*(1), 37–45. https://doi.org/10.1111/j.1745-3992.1991.tb00183.x

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*(1), 6–20. https://doi.org/10.1177/0013164498058001002

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development*, *44*(3), 159–168. https://doi.org/10.1177/0748175611409845

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement*, *60*(2), 201–223. https://doi.org/10.1177/00131640021970466

APPENDIX A

Table 1

Subscale reliability coefficients descriptive statistics.

	Mean	SD
N	531.57	1064.13
Mean Age	31.08	15.80
Prop. Female	0.64	0.200
P	0.87	0.051
S	0.88	0.050
MP	14.31	10.170
SDP	4.14	4.20
MS	16.60	8.7
SDS	5.65	2.82

Note. MP denotes mean of Presence for each sample study, SDP is the standard deviation of P for each sample study, MS is the mean of search, and SDS is the standard deviation of search.

Table 2

Reliability coefficients, Confidence Intervals, and Q statistics.

Scale	K	Estimated α	CI	Q
Presence	152	.87	(.872, .887)	5725.20***
Search	89	.88	(.879, .897)	2492.06***

Note. ****p* < .0001

Table 3

Mixed-effects results for Presence subscale

Variable	Estimate	SE	Zval	Pval	Ci lb	Ci ub
Mean age	-0.0005	.0004	-1.283	.199	-.0012	.0002
Proportion Female	-.002	.028	-.059	.953	-.057	.054
Language	-.042	.0152	-2.773	*.005***	-.072	-.012
Europe	-.003	.018	-.150	.881	-.038	.032
Other Region	-.011	.014	-.760	.4472	-.038	.019
Sample Type	-.014	.015	-.951	.3414	-.042	.015

Note. ***p* < .01

Table 4

Mixed-effects results for Search subscale

Variable	Estimate	SE	Zval	Pval	Ci lb	Ci ub
Mean age	.0005	.0005	1.0267	.3046	-.0004	.0014
Proportion Female	.0157	.0353	.4452	.6562	-.0535	.0850
Language	-.0356	.0178	-1.9975	*.0458**	-.0706	-.0007
Europe	-.0048	.0203	-.2385	.8115	-.0447	.0350
Other Region	-.0366	.0182	-2.0094	*.0445**	-.0722	-.0009
Sample Type	-.0095	.0190	-.4982	.6183	-.0466	.0277

Note. **p* < .05

APPENDIX B

Figure 1

Funnel plot of Presence Subscale

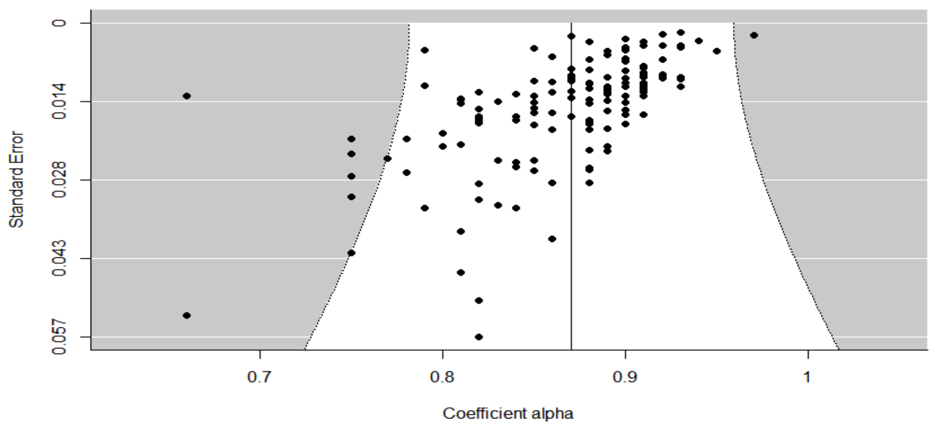

Figure 2

Funnel plot of the Search Subscale