Running head: IMPORTANCE WEIGHTING

THE OFTEN OVERLOOKED ISSUE OF STATISTICAL POWER: THIS AND OTHER
ISSUES REGARDING ASSESSING IMPORTANCE WEIGHTING IN QUALITY OF LIFE
MEASURES

Chang-ming Hsieh

University of Illinois at Chicago

Author Note

Correspondence concerning this article should be addressed to Chang-ming Hsieh, Jane Addams College of Social Work, University of Illinois at Chicago, 1040 West Harrison Street (MC 309), Chicago, IL 60607-7134. Electronic mail may be sent via Internet to chsieh@uic.edu.

Abstract

In the area of quality of life research, researchers may ask respondents to rate importance as well as satisfaction of various life domains (such as job and health) and use importance ratings as weights to calculate overall, or global, life satisfaction. The practice of giving more important domains more weight, known as importance weighting, has not been without controversy. Several previous studies assessed importance weighting using the analytical approach of moderated regression. This study discusses major issues related to how importance weighting has been assessed. Specifically, this study highlights that studies on importance weighting without considering statistical power are prone to type II error, i.e., failing to reject the null hypothesis of no significant weighting effect when the null hypothesis is actually false. The sample size required for adequate statistical power to detect importance weighting functions appeared larger than most previous studies could offer.

Keywords: subjective well-being, multiplicative scores, moderated regression

THE OFTEN OVERLOOKED ISSUE OF STATISTICAL POWER: THIS AND OTHER ISSUES REGARDING ASSESSING IMPORTANCE WEIGHTING IN QUALITY OF LIFE MEASURES

1. Introduction

The relationship between a psychological construct and individual facets or domains of the construct has long been a topic of study (e.g., James, 1950). In the area of quality of life (QoL) research, for example, it has long been recognized that a noticeable discrepancy exists in the correlation between overall, or global, life satisfaction measures and summed satisfaction scores across life domains, such as job, health and finance (e.g., Hsieh, 2003; Russell & Hubley, 2005). One possible explanation for the discrepancy is that not all life domains are equally important (e.g., Campbell, Converse & Rogers, 1976; Hsieh, 2003, 2004). To account for different domain importance, many researchers include importance ratings as well as satisfaction ratings of life domains to represent global life satisfaction or QoL (e.g., Ferrans & Powers, 1985). The practice of using subjective (i.e., self-report perceived) importance ratings of domains as weights, otherwise known as importance weighting, can be seen in multiple areas of social research, such as self-esteem and job satisfaction (Russell & Hubley, 2005). In their review of empirical evidence regarding importance weighting, Russell and Hubley (2005) concluded that empirical support for importance weighting has been insufficient. More specifically, there does not appear to be sufficient evidence to support that importance weighting improves the prediction of a construct when compared to the use of unweighted (i.e., the total primary) scores. Although Russell and Hubley (2005) discussed several possible reasons leading to these empirical findings, one critical issue, the issue of statistical power, was not included in the discussion. This study extends the discussion of major issues regarding assessing importance

weighting in the context of QoL. Specifically, this study discusses: 1) the limitations of using arbitrarily selected QoL measures, often called criterion variables, to establish criterion validity in assessing importance weighting; 2) the problems associated with using multiplying satisfaction and importance ratings (i.e., multiplicative scores) as the weighting method or weighting mechanism; 3) the questionable assumption that all life domains are included in QoL measures; and 4) the need for considering statistical power, which is the probability that a statistical test correctly rejects the null hypothesis (e.g., Aberson, 2010; Cohen, 1988), in moderated regression analysis used for assessing importance weighting.

In the QoL literature, importance weighting has been accomplished in different ways (Hsieh, 2012). One approach to importance weighting is to assign pre-determined weights to different domains according to a theoretical or conceptual foundation. These weights are used to represent importance of various domains for all respondents without considering any potential individual differences. Another common approach is achieved by using regression coefficients as weights in a regression-type analysis. Given that regression coefficients are obtained based on data at the aggregate level, using them as weights also precludes the consideration of potential individual differences. A third common approach is to apply different domain importance weights, based on responses from different individuals. This approach of importance weighting incorporates relative importance at the individual level and is the focus of this paper.

2. Background

2.1 Debating the Need for Importance Weighting

Debates regarding the need for importance weighting have centered around the fact that quite a few empirical study results showed that domain importance weighting did not seem to increase the variations explained in global QoL measures (e.g., Campbell et al., 1976; Russell et

al., 2006; Wu, 2008a, 2008b). For example, Phillip, Merluzzi, Peterman and Cronk (2009) found that weighting (using domain preference as a proxy of domain importance) did not improve the power of health-related QoL, as measured by the Functional Assessment of Cancer Therapy (FACT), in predicting global QoL, as measured by the Satisfaction with Life scale (SWLS). Therefore, they support the use of unweighted FACT scores to represent health-related QoL.

To a large extent, the inconclusive evidence is a result of the ways by which importance weighting is assessed. Multiple issues regarding the ways importance weighting is assessed have been identified (e.g., Hsieh, 2013; Russell & Hubley, 2005). Missing from the discussion regarding assessing importance weighting is the issue of statistical power, however. Many empirical studies assessing importance weighting used moderated regression analysis as the analytical approach (e.g., Hsieh, 2012; Mastekaasa, 1984; Russell et al., 2006; Wu & Yao, 2006a, 2006b). Highlighting the importance of considering statistical power for moderated regression analysis, this paper aims to extend the discussion of the major issues regarding domain importance weighting in QoL measures. Four issues pertaining to assessing domain importance weighting were examined as follows and special attention, including empirical example, was given to the issue of statistical power.

2.2 Major Issues Regarding Assessing Importance Weighting in QoL Measures

As Hsieh (2013) suggested, for most empirical research investigating the performance of importance weighting, the prevailing approach has three major characteristics. First, only a limited number and choice of QoL measures is used as criterion variables. Second, methods or mechanisms used to construct importance weighting are predominately based on multiplicative scores. Third, only a limited number of domains are included in the measures. This prevailing approach brings forth the following three unanswered questions that significantly influence the

current state of knowledge of the topic of importance weighting:

2.2.1 Question One: Is There a “Best” Choice of QoL Measure as a Criterion Variable?

A few measures that are developed to assess overall QoL (i.e., global QoL measures), such as the Satisfaction with Life Scale (Diener, Emmons, Larsen, & Griffin, 1985), and single-item overall QoL questions, have been used in previous studies as criterion variables to assess the performance of importance weighting. There are certainly many more global QoL measures that exist (e.g., Evans, 1994; Hagerty et al., 2001). Although it is reasonable to expect that scores from all global QoL measures reflect global QoL, there does not appear to be any specific standard for the selection of global QoL measures to serve as criteria in evaluating the performance of importance weighting. Different QoL measures should be expected to overlap to some extent. However, there is no reason to presume that all QoL measures must capture identical constructs (Russell & Hubley, 2005; Russell et al., 2006). Given that there is no single agreed-upon “best” QoL measure, the choice of QoL measures as criterion variables becomes arbitrary. For example, Hagell and Westergren (2006) evaluated the value of importance weighting by examining the relationship between weighted and unweighted (satisfaction only) scores of Ferrans and Powers’ (1985) Quality of Life Index and a single-item overall life satisfaction measure as well as a single-item QoL measure. Based on data from a survey of 81 Parkinson’s disease patients, Hagell and Westergren (2006) found that weighted total scores correlated stronger with the single-item overall life satisfaction measure than with the single-item overall QoL measure, and unweighted scores showed the opposite pattern.

Considering the choice of global QoL measures is critical in situations when evidence is used to argue for importance weighting, given that the evidence is heavily dependent on the criterion variables chosen (e.g., Russel et al., 2006). However, considering the choice of global

QoL measures is even more critical in situations when evidence is used to argue against importance weighting because it is unlikely, if not impossible, for any particular study to cover all global QoL measures exhaustively as criterion variables. Arguing against importance weighting using evidence based on a small number of arbitrarily chosen criterion variables without considering that many more other possible criterion variables have been left out can be problematic. Unless there is a widely accepted single “best” QoL measure that can be used as “the” criterion variable, any empirical evidence evaluating importance weighting is clearly not without bias.

2.2.2 Question Two: Is Weighting Equal to Multiplicative Scores?

In the literature on importance weighting, most weighting methods applied so far have been based on the assumption of a linear function of domain importance scores (Hsieh, 2013). More often than not, multiplicative scores (multiplying satisfaction by importance scores) are used to represent the weighting function (e.g., Russell et al., 2006; Trauer & Mackinnon, 2001). As Hsieh (2003, 2004) showed, the performance of importance weighting was dependent upon how importance was measured and how importance was weighted. The popular practice of using rating scales to measure domain importance and using multiplicative scores as weighting is only one of the many possible ways to measure importance weighting. As Kaplan, Feeny, and Revicki (1993) indicated, the use of rating scales was by no means the only method for assessing relative importance. Other methods, including utility assessment, such as the standard gamble, time trade-off and person trade-off methods, as well as economic measurements of choice should not be ignored (Kaplan et al., 1993). In addition to different ways of measuring domain importance, there can be many other options other than simply multiplying satisfaction scores by importance scores to the construction of weighting. Campbell, Converse and Rogers (1976), for example,

suggested various types of weighting function of domain importance, such as “hierarchy of needs,” “threshold,” and “ceiling.” It should be noted that the actual weighting function of domain importance is still unclear and is not necessarily limited to the functions discussed by Campbell, Converse and Rogers (1976). Guardiola and Picazo-Tadeo (2014), for example, used data envelopment analysis (DEA) and multi-criteria-decision-making (MCDM) techniques to construct weighted-domain composite indices of life satisfaction. Based on the so-called “benefit-of-the-doubt” principle (Cherchye, Moesen, Rogge, & van Puyenbroek, 2007), these weighting techniques went beyond the typical multiplicative scores and allowed weights to differ across life domains and individuals (Guardiola & Picazo-Tadeo, 2014).

Importance weighting should not be limited to only multiplicative scores when evidence is used to argue against importance weighting, especially when evidence is based on multiplicative scores or similar methods. Even if there is no evidence to support a linear function of domain importance based on rating scale scores, it is unclear what the evidence would show with other types of weighting function of domain importance and different types of measurement of domain importance.

2.2.3 Question Three: How Should Importance Weighting be Assessed if a QoL Measure Does Not Cover All Domains of Life Comprehensively?

According to Cummins (1996), there are at least 173 different domain names that have been used in the QoL literature, and it is unlikely that 173 domain names cover all possible domains comprehensively. For any single QoL measure to cover all possible domains of life is extremely unlikely, if not impossible. Measures of global QoL typically follow either the “bottom-up” or “top-down” approach (e.g., Diener, 1984; Feist et al., 1995; Headey et al., 1991; Lance et al., 1989; Scherpenzeel & Saris, 1996). For example, the Satisfaction with Life Scale by

Diener, Emmons, Larsen and Griffin (1985) used the “top-down” approach, while the Quality of Life Index by Campbell, Converse and Rogers (1976) followed the “bottom-up” approach. QoL measures that follow the bottom-up approach are consistent with the “formative-indicator” model, and QoL measures that follow the top-down approach are consistent with the “reflective-indicator” model (Chin & Newsted, 1999; Cohen et al., 1990; Hsieh, 2004). In a formative-indicator model, indicators are viewed as determining or causing the construct. In a reflective-indicator model, indicators are viewed as determined by the construct (Bollen & Lennox, 1991; Cohen et al., 1991; Chin & Newsted, 1999). As Bollen and Lennox (1991) suggested, a general reflective-indicator model can be expressed as:

$$y_i = \lambda_{i1}\eta_1 + \varepsilon_i \quad (1)$$

where y_i is the i th indicator, η_1 is the construct or latent variable that affects it, ε_i is the measurement error for the i th indicator, and λ_{i1} is the coefficient giving the expected effect of η_1 on y_i . For multiple indicators, there will be multiple equations like Equation 1 (one for each indicator). A general formative-indicator model, on the other hand, can be expressed as:

$$\eta_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + \gamma_{13}x_3 + \cdots + \gamma_{1n}x_n + \zeta_1 \quad (2)$$

where x s are the indicators, η_1 is the construct or latent variable affected by them, ζ_1 is the measurement error, and γ s are the coefficients giving the expected effects of x s. In the context of QoL research, Hsieh (2004) provided a conceptual framework for incorporating importance weighting in a formative-indicator model. Specifically, a formative QoL model can be expressed as:

$$Q = \gamma_1d_1 + \gamma_2d_2 + \gamma_3d_3 + \cdots + \gamma_nd_n + \zeta \quad (3)$$

where d s are weighted (by importance) domain satisfactions, Q is global QoL or life satisfaction, ζ is the measurement error, and γ s are the coefficients giving the expected effects of d s.

Consequences of not covering all domains comprehensively vary, depending on the measurement model. For QoL measures that follow the reflective-indicator model in which measure items are considered interchangeable, whether or not domains are covered comprehensively is of no major concern. On the other hand, for QoL measures that follow the formative-indicator model, items in the measure may not be interchangeable. The extent of domains covered in the measure can be a cause for major concern, given that not including items can alter the measurement construct (e.g., Bollen & Lennox, 1991; Diamantopoulos, 2006; Diamantopoulos et al., 2008; Hardin et al., 2011; Jarvis, et al., 2003; MacKenzie et al., 2005).

Given that many QoL measures used to investigate importance weighting followed the formative-indicator model (e.g., Campbell et al., 1976; Hsieh, 2003; Wu, 2008b; Wu et al., 2009) and it is unlikely for any single measure to cover all possible domains comprehensively, the consequences of not including all possible domains comprehensively in formative-indicator QoL measures should be investigated. There has not been any established standard regarding the extent of comprehensiveness or exhaustiveness of the domains for formative-indicator global QoL measures. As a result, the number of domains covered in different QoL measures varies, depending on the focus of the measure. Formative-indicator global QoL measures that cover different subsets of domains are unlikely to measure identical QoL constructs.

Consequently, finding any global QoL measures as criterion variables to assess the performance of importance weighting in formative-indicator QoL measures becomes a complex task. As discussed earlier, it is unrealistic to expect a formative-indicator global QoL measure to cover all possible domains comprehensively. It is, therefore, hard to justify why the formative-indicator QoL measures being assessed (in relation to importance weighting) and the criterion variables must measure the same concept of QoL. It does not seem reasonable to expect a

formative-indicator global QoL measure covering five domains and one covering 20 domains to measure the same concept of QoL. In assessing the performance of importance weighting, it is important to consider the fact that formative-indicator global QoL measures are unlikely to cover all possible domains comprehensively. Specifically, the selection of criterion variables should take into account the scope of domains covered in the formative-indicator QoL measures. In the case when evidence does not support importance weighting, it is important to carefully examine whether the lack of support is a result of inappropriate criterion variables before coming to a concluding argument against importance weighting.

2.2.4 Further Consideration

In addition to the three unanswered questions discussed above, our understanding of importance weighting has also been greatly affected by the analytical methods used to assess the performance of importance weighting. The common analytical methods used to assess the performance of importance weighting have been correlation- or regression-based (such as correlation, moderated regression, and partial least squares regression), with moderated regression analysis being most popular (e.g., Hsieh, 2012; Mastekaasa, 1984; Russell et al., 2006; Wu & Yao, 2006a, 2006b). The moderated regression analysis is generally done by selecting a small number of global QoL measures, often one or two popular existing QoL measures or a single-item global measure, as criterion variables, and then comparing weighted and unweighted scores against the criterion variables to assess the performance of importance weighting (e.g., Hsieh, 2003, 2004, 2012; Russell et al., 2006; Wu et al., 2009). The seemingly reasonable logic is predicated upon two major assumptions. One assumption is that the criterion variables chosen represent the “true” global QoL. The other assumption is that the analytical methods used to assess importance weighting are effective. As Hsieh (2013) demonstrated

recently, moderated regression analysis was likely to produce misleading results due to issues related to the weighting mechanisms chosen and the amount of domains covered. Hsieh's (2013) point about the potentially problematic results is worth noting, but it has to do with issues related to the unanswered questions discussed above, not the analytical method itself.

The current paper examines potential issues related to the analytical method itself. More specifically, it is argued here that even in a hypothetical situation where all the issues related to the unanswered questions discussed in 2.1.1 to 2.1.3 are resolved, the use of moderated regression analysis can still be problematic if consideration of statistical power is not given.

Statistical power is by no means a new concept (e.g., Aberson, 2010; Cohen, 1988). For moderated regression analysis, the popular analytical method used to assess the performance of domain importance, statistical power is dependent upon, among other factors, sample size (Aberson, 2010). A close look at the empirical evidence used for or against importance weighting would suggest that most studies had sample sizes that were small or moderate at best. For example, analyzing data from 130 undergraduate students, Wu and Yao (2006a) found that domain importance did not moderate the relationship between domain satisfaction and overall life satisfaction and argued against importance weighting. Russell et al. (2006) investigated the topic of importance weighting based on a sample of 241 subjects. More recently, Phillip et al. (2009) argued against importance weighting based on data from 194 cancer patients.

Sample size affects statistical power. With a limited sample size, statistical power is likely to be limited if the effect size (Cohen, 1998) to be detected is not large. In the next section, a numerical example is used to illustrate this point. This numerical example is based on a hypothetical scenario that all the issues related to the unanswered questions discussed earlier have been resolved. In other words, the numerical example was based on the hypothetical

scenario that 1) the true function of importance weighting was known and appropriate global QoL scores could be obtained as the criterion variable; and 2) the QoL measure being assessed covered all possible domains comprehensively. The hypothetical weighting function used to represent the “true” function of importance weighting was the common practice of multiplicative scores (multiplying satisfaction with importance scores). It should be noted that it is not the intent of this paper to suggest that the true function of importance weighting follow a multiplicative score. Instead, the intent is to show the potential issues related to statistical power of the prevalent analytical method used to assess the performance importance weighting, based on this most popular weighting function (e.g., Hsieh, 2012; Russell et al., 2006; Wu & Yao, 2006a, 2006b).

3. An Example

3.1 Data and Method

To explore the issues related to statistical power, an analysis of survey data from an earlier study was conducted as a numeric example. The telephone survey used random digit dialing to call listed phone numbers from telephone directories in the greater Chicago area. A total 100 telephone interviews were conducted with participants who were at least 50 years old (see Hsieh, 2003 for details). The survey asked respondents to rate their satisfaction for each of the following eight domains: (a) health, (b) work, (c) spare time, (d) financial situation, (e) neighborhood, (f) family life, (g) friendships, and (h) religion. The satisfaction ratings could be completely satisfied (7), somewhat satisfied (6), slightly satisfied (5), neither satisfied nor dissatisfied (4), slightly dissatisfied (3), somewhat dissatisfied (2), or completely dissatisfied (1). Respondents were asked to rate the importance (i.e., domain importance rating) of each of the eight discrete life domains as: not at all important (1), not too important (2), somewhat important

(3), very important (4), or extremely important (5). Three criterion variables were used: a single-item global satisfaction question (satisfaction with life overall) and two multi-item global satisfaction measures, Life Satisfaction Index A (LSIA, Neugarten, Havinghurst, & Tobin, 1961) and Satisfaction with Life Scale (SWLS, Diener et al., 1985). LSIA is a 20-item global subjective well-being measure used and examined quite extensively in social gerontology (George & Bearon, 1980; Liang, 1984, 1985; Liang, Lawrence, & Bollen, 1987). SWLS is a five-item global life satisfaction scale. All its items are measured on a seven-point scale, and it has shown good psychometric properties among both young and older adult populations (Diener et al., 1985).

Following the work of Mastekaasa (1984) and Evans (1991), moderated regression analysis was conducted. The analysis was based on the three-step process outlined by Evans in 1991 (also see Hsieh, 2012). The first step was to regress the criterion variable on satisfaction scores of all eight domains. The second step was to add the importance scores of the eight domains. The third step was to add to the second step the eight interaction terms of domain satisfaction by importance. A significant increase in R^2 (indicated by the by F -test) from step two to step three would indicate the need for including the interaction terms, and therefore, support importance weighting. Based on the assumption made in this study, one would expect to see a significant increase in R^2 from step two to step three. The same moderated regression analysis was conducted with each of the three criterion variables as the dependent variable. Changes in R^2 from step two to step three would be the effect sizes of the importance by satisfaction terms (Cohen, 1988).

3.2 Results

Table 1 shows the results from the moderated regression analysis with each of the three

criterion variables. As shown in Table 1, none of the importance by satisfaction interaction terms in these models was statistically significant. However, there were noticeable differences in the contribution of the interaction terms with different criterion variables. That is, size of the effect of importance by satisfaction terms on the criterion variables appeared to differ. For the model with the single-item measure as the criterion variable, the importance by satisfaction interaction terms contributed to an R^2 change of .108 (R^2 increased from .339 in step 2 to .447 in step 3). For the model with LSIA as the criterion variable, the importance by satisfaction interaction terms contributed to an R^2 change of .027 (R^2 increased from .349 in step 2 to .375 in step 3). For the model with SWLS as the criterion variable, the importance by satisfaction interaction terms contributed to an R^2 change of .017 (R^2 increased from .468 in step 2 to .485 in step 3).

[Table 1 about here]

Assuming the effect sizes shown in Table 1 accurately reflect the relationships between the interaction terms and the criterion variables, the sample size provided by the survey data did not appear to offer adequate statistical power to detect any interaction effects. Based on the effect sizes shown in Table 1 with α at .05 and the sample size of 90 in the survey data, statistical power was calculated using the software Power Analysis and Sample Size (PASS) 2008 (NCSS, 2008) for each of the criterion variables. For the single-item measure, the survey data had a statistical power of about 82%; for LISA, the survey data had a statistical power of about 21%; and for SWLS, the survey data had a statistical power of 16%.

4. Further Illustration

4.1 Data and Method

Since empirical data sets available usually have a limited sample size and cover a limited amount of life domains, further illustration was provided, using computer-generated, simulated

data to overcome the limitations. It should be noted that examining statistical power to detect interaction effects with simulated data is by no means new (e.g., McClelland & Judd), the intent here is to emphasize the importance of considering statistical power in the context of assessing importance weighting in the QoL measures. A total of 10,000 observations, representing 10,000 individuals, were generated as the study population. For each observation, there were 25 satisfaction score variables (S_1 to S_{25}) and 25 importance score variables (I_1 to I_{25}). These variables were generated to represent satisfaction and importance scores for 25 life domains. The (satisfaction as well as importance) scores were generated randomly, each with a uniform distribution ranging between one and seven to resemble the commonly used seven-point Likert-type scale scores.

Although the number of all possible life domains is likely to be much larger than 25 (Cummins, 1996; Cummins et al., 1994), most studies investigating the topic of importance weighting included no more than 25 domains (e.g., Campbell et al., 1976; Hsieh, 2003, 2004; Russell et al., 2006; Wu & Yao, 2006a, 2006b). To overcome the limitation of using only a small number of life domains as presented in 3.1 and 3.2 while still consistent with the QoL literature to date, this study used 25 domains for this numeric example.

In addition, five global QoL variables were generated as criterion variables to represent scores from other global QoL measures. Given different QoL measures are likely to have varying degrees of overlap, these criterion variables were created to be significantly correlated with the multiplicative scores at different degrees. The correlations between the criterion variables (criterion 1, criterion 2, criterion 3, criterion 4 and criterion 5) and the multiplicative scores were set at $r = .80$, $r = .70$, $r = .60$, $r = .50$ and $r = .40$ respectively. These criterion variables were used as the dependent variables for the later moderated regression analysis.

The same three-step moderated regression analysis described earlier was conducted with each of the five criterion variables as the dependent variable. Changes in R^2 from step two to step three were obtained as effect sizes (Cohen, 1988). Based on the effect size with α at .05 and β at .90, required sample size was calculated using PASS 2008 (NCSS, 2008).

4.2 Results

Table 2 shows the results from the moderated regression analysis with each of the five criterion variables. Given that 1) the criterion variables were significantly correlated with the multiplicative scores, and 2) the large sample size, it was reasonable to expect that the moderated regression analysis results would indicate a significant contribution of the importance by satisfaction interaction terms as shown in the change in R^2 (and F -test) in all five models. As shown in Table 2, the contribution of importance by satisfaction interaction terms was statistically significant in all five models. For the model with criterion variable 1, the importance by satisfaction interaction terms contributed to an R^2 change of .056 (R^2 increased from .596 in step 2 to .652 in step 3). For the model with criterion variable 2, the importance by satisfaction interaction terms contributed to an R^2 change of .047 (R^2 increased from .451 in step 2 to .498 in step 3). For the model with criterion variable 3, the importance by satisfaction interaction terms contributed to an R^2 change of .038 (R^2 increased from .331 in step 2 to .369 in step 3). For the model with criterion variable 4, the importance by satisfaction interaction terms contributed to an R^2 change of .021 (R^2 increased from .235 in step 2 to .256 in step 3). For the model with criterion variable 5, the importance by satisfaction interaction terms contributed to an R^2 change of .015 (R^2 increased from .146 in step 2 to .158 in step 3).

[Table 2 about here]

The next step was to determine the adequate sample size using PASS 2008 (NCSS,

2008), based on the effect size obtained from each of the criterion variables. For criterion variable 1 (a correlation of $r = .80$ with the multiplicative score), a sample size of 204 is needed to achieve 90% power to detect an R^2 of .056 attributed to 25 independent variables using an F-test with a significance level of .05 (after adjusting for 50 independent variables with an R^2 of .594). For criterion variable 2 (a correlation of $r = .70$ with the multiplicative score), a sample size of 326 is needed to achieve 90% power to detect an R^2 of .047 attributed to 25 independent variables using an F-test with a significance level of .05 (after adjusting for 50 independent variables with an R^2 of .451). For criterion variable 3 (a correlation of $r = .60$ with the multiplicative score), a sample size of 493 is needed to achieve 90% power to detect an R^2 of .038 attributed to 25 independent variables using an F-test with a significance level of .05 (after adjusting for 50 independent variables with an R^2 of .331). For criterion variable 4 (a correlation of $r = .50$ with the multiplicative score), a sample size of 1,025 is needed to achieve 90% power to detect an R^2 of .021 attributed to 25 independent variables using an F-test with a significance level of .05 (after adjusting for 50 independent variables with an R^2 of .235). For criterion variable 5 (a correlation of $r = .40$ with the multiplicative score), a sample size of 1,599 is needed to achieve 90% power to detect an R^2 of .015 attributed to 25 independent variables using an F-test with a significance level of .05 (after adjusting for 50 independent variables with an R^2 of .15).

5. Discussion

Although importance weighting in QoL measures is not a new topic of study (e.g., Hagerty et al., 2001; Hagerty & Land, 2007; Hsieh, 2012; Russell & Hubley, 2005; Russell et al., 2006; Skalli et al., 2008; Wu & Yao, 2007), how to best assess the performance of importance weighting remains an unsettled issue (e.g., Hsieh, 2012; Russell et al., 2006). The use of

moderated regression analysis as a method of assessing the performance of importance weighting has gained popularity in recent years (e.g., Wu & Yao, 2006a, 2006b). However, the adequacy of using moderated regression analysis to assess the performance of domain importance remains unclear. As Hsieh (2013) recently pointed out, results produced by moderated regression analysis as a method of assessing the performance of importance weighting could be strongly influenced by factors such as the weighting approach selected and the number of domains included.

The current study examined the often neglected issue of statistical power of moderated regression analysis as a method of assessing the performance of importance weighting. This study showed that the sample size required for reasonable statistical power might be larger than many of the previous studies could offer. The actual sample size required for any study to evaluate the performance of importance weighting will depend on the criterion variable(s) selected and the number of domains included. However, if the criterion variable(s) selected and the “true” scores of importance weighting function have only a moderate level of correlation (for example, $r = .60$ or less), a sample size of 200 or less, which is easily found in many previous studies, is unlikely to reach reasonable statistical power to assess the performance of importance weighting. Without adequate statistical power, chances of a type II error (false negative decision) occurring are likely to increase. Importance weighting studies that do not have adequate statistical power are prone to fail to reject the null hypothesis of no significant weighting effect when the null hypothesis is, in fact, false.

According to Shultz and Whitney (2004), most validity coefficients used to assess concurrent validity do not reach .5. In other words, it would be unrealistic to expect high associations/correlations between different measures of QoL. Under the circumstances, unless there is clear evidence to suggest otherwise, it would be prudent not to assume that any criterion

variable selected to assess importance weighting will represent the “true” global QoL with a very high degree of overlap. As a matter of fact, the association or correlation between any criterion variable and the “true” global QoL is unknown. Selection of criterion variables is often subjective. Treating the criterion variables as absolute standards for assessing importance weighting without considering whether or not as well as the extent to which these criterion variables really represent the “true” global QoL is by no means ideal.

Evidence provided in this study suggests the following. First, for the popular analytical method of moderated regression analysis to be effective in assessing importance weighting, statistical power must be considered. For studies on importance weighting, considerations of statistical power should be given with regard to sample size as well as the degree to which criterion variables selected actually represent the “true” global QoL. Unfortunately, the “true” global QoL of any measure under investigation for importance weighting is unknown (Hsieh, 2013). The degree to which any criterion variable represents this true global QoL is, therefore, unknown. Under the circumstances, it would be helpful for future studies to include statistical power information, especially when analysis results show no significant weighting effect, so the potential false positive rate can be understood. Second, evidence presented here was based on the assumption that importance weighting functioned perfectly following multiplicative scores. It is unrealistic to expect a perfect weighting function of multiplicative scores or any other type of weighting in real-life situations. That is, the size of moderation/interaction effects under investigation in moderated regression analysis may be smaller than what was presented based on the simulated data. With smaller size of moderation/interaction effects, sample size requirements presented in this study are likely under-estimated. Third, the validity as well as sensitivity of moderated regression analysis as a method to assess the performance of importance weighting is

heavily affected by the criterion variables selected. Since there is no single universally agreed-upon “best” QoL measure out there, it is important for future studies to provide clear rationale for the choice of the criterion variables if moderated regression analysis is to be used to assess the performance of importance weighting. Efforts to develop analytical methods that are more robust and less dependent on criterion variables selected for assessing importance weighting should be encouraged.

References

- Aberson, C. L. (2010). *Applied power analysis for the behavioral sciences*. New York: Psychology Press.
- Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, 110, 305-314.
- Campbell, A., Converse, P. E., & Rogers, W. L. (1976). *The quality of American life: Perceptions, evaluations, and satisfactions*. New York: Russel Sage.
- Cherchye, L., Moesen, W., Rogge, N., & van Puyenbroek, T. (2007). An introduction to ‘benefit of the doubt’ composite indicators. *Social Indicators Research*, 82, 111–145.
- Chin, W. W., & Newsted, P. R. (1999). Structural equation modeling analysis with small samples using partial least squares, in R. H. Hole (Ed.), *Statistical strategies for small sample research* (pp. 307-341). Thousand Oaks, California: Sage.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Cohen, P., Cohen, J., Teresi, J., Marchi, M., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equations casual models. *Applied Psychological Measurement*, 14, 183-196.
- Cummins, R. A. (1996). The domains of life satisfaction: An attempt to order chaos. *Social Indicators Research*, 38, 303-328.
- Cummins, R. A., McCabe, M. P., Romeo, Y., & Gullone, E. (1994). The comprehensive quality of life scale: Instrument development and psychometric evaluation on tertiary staff and students. *Educational and Psychological Measurement*, 54, 372-382.

- Diamantopoulos, A. (2006). The error term in formative measurement models: Interpretation and modeling implications. *Journal of Modeling in Management, 1*, 7-17.
- Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research, 61*, 1201-1302.
- Diener, E., Emmons, R.A., Larsen, R.J. & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*, 71–74.
- Diener, E. (1984). Subjective well-being. *Psychological Bulletin, 95*, 542-575.
- Evans, D. R. (1994). Enhancing quality of life in the population at large. *Social Indicators Research, 33*, 47–88.
- Evans, M.G. (1991). The problem of analyzing multiplicative composites: Interactions revisited. *American Psychologist, 46*, 6–15.
- Ferrans, C., & Powers, M. (1985). Quality of Life Index: Development and psychometric properties. *Advances in Nursing Science, 8*, 15-24.
- Feist, G. J., Bodner, T. E., Jacobs, J. F., Miles, M., & Tan, V. (1995). Integrating top-down and bottom-up structural models of subjective well-being: A longitudinal investigation. *Journal of Personality and Social Psychology, 68*, 138-150.
- George, L. K., & Bearon, L.B. (1980). *Quality of life in older persons: Meaning and measurements*. New York: Human Sciences Press.
- Guardiola, J., & Picazo-Tadeo, A. J. (2014). Building weighted-domain composite indices of life satisfaction with data envelopment analysis. *Social Indicators Research, 117*, 257-274.
- Hagell, P., & Westergren, A. (2006). The significance of importance: An evaluation of Ferrans and Powers' quality of life index. *Quality of Life Research, 15*, 867-876.

- Hagerty, M. R., Cummins, R. A., Ferris, A. L., Land, K. C., Michalos, A. C., Peterson, M., Sharpe, A., Sirgy, J., & Vogel, J. (2001). Quality of life indexes for national policy: Review and agenda for research. *Social Indicators Research*, 55, 1-96.
- Hagerty, M. R., & Land, K. C. (2007). Constructing summary indices of quality of life: A model for the effect of heterogeneous importance weights. *Sociological Methods and Research*, 35, 455-496.
- Hardin, A. M., Chang, J. C., Fuller, M. A., & Torkzadeh, G. (2011). Formative measurement and academic research: In search of measurement theory. *Educational and Psychological Measurement*, 71, 281-305.
- Headey, B., Veenhoven, R., & Wearing, A. (1991). Top-down versus bottom-up theories of subjective well-being. *Social Indicators Research*, 24, 81-100.
- Hsieh, C. M. (2003). Counting importance: The case of life satisfaction and relative domain importance. *Social Indicators Research*, 61, 227-240.
- Hsieh, C. M. (2004). To weight or not to weight: The role of domain importance in quality of life measurement. *Social Indicators Research*, 68, 163-174.
- Hsieh, C. M. (2012). Should we give up domain importance weighting on QoL measures? *Social Indicators Research*, 108, 99-109.
- Hsieh, C. M. (2013). Issues in evaluating importance weighting in quality of life measures. *Social Indicators Research*, 110, 681-693.
- James, W. (1950). *The principles of psychology* (Vol. 1). New York: Dover
- Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, 30, 199-218.

- Kaplan, R.M., Feeny, D., & Revicki, D. A. (1993). Methods for assessing relative importance in preference based outcome measures. *Quality of Life Research*, 2, 467-475.
- Lance, C. E., Lautenschlager, G. J., Sloan, C. E., & Varca, P. E. (1989). A comparison between bottom-up, top-down, and bidirectional models of relationships between global and life facet satisfaction. *Journal of Personality*, 57, 601-624.
- Liang, J. (1984). Dimensions of the Life Satisfaction Index A: A structural formulation. *Journal of Gerontology*, 39, 613-622.
- Liang, J. (1985). A structural integration of the Affect Balance Scale and the Life Satisfaction Index A. *Journal of Gerontology*, 40, 552-561.
- Liang, J., Lawrence, R.H., & Bollen, K. (1987). Race differences in the factorial structure of two measures of subjective well-being. *Journal of Gerontology*, 42, 426-428.
- Locke, E.A. (1969). What is job satisfaction? *Organizational Behavior and Human Performance*, 4, 309-336.
- Locke, E. A. (1976). The nature and causes of job satisfaction. In M. D. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology* (pp. 1297-1349). Chicago: Rand McNally.
- MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology*, 90, 710-730.
- Mastekaasa, A. (1984). Multiplicative and additive models of job and life satisfaction. *Social Indicators Research*, 14, 141-163.
- McClelland, G. H. & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114, 376-390.

- NCSS (2008). *Power analysis and sample size 2008*. Kaysville, Utah: NCSS.
- Neugarten, B.L., Havinghurst, R.J., & Tobin, S.S. (1961). Measurement of life satisfaction. *Journal of Gerontology, 16*, 134–143.
- Philip, E. J., Merluzzi, T. V., Peterman, A., & Cronk, L. B. (2009). Measurement accuracy in assessing patient's quality of life: To weight or not to weight domains of quality of life. *Quality of Life Research, 18*, 775-782.
- Rojas, M. (2006). Life satisfaction and satisfaction in domains of life: Is it a simple relationship? *Journal of Happiness Studies, 7*, 467-497.
- Russell, L. B., & Hubley, A. M. (2005). Importance ratings and weighting: Old concerns and new perspectives. *International Journal of Testing, 5*, 105-130.
- Russell, L. B., Hubley, A. M., Palepu, A., & Zumbo, B. D. (2006). Does weighting capture what's important? Revisiting subjective importance weighting with a quality of life measure. *Social Indicators Research, 75*, 146-167.
- Scherpenzeel, A. & Saris, W. (1996). Causal direction in a model of life satisfaction: The top-down/bottom-up controversy. *Social Indicators Research, 38*, 161-180.
- Shultz, K. S., & Whitney, D. J. (2004). *Measurement theory in action: Case studies and exercises*. Thousand Oaks, California: Sage.
- Skalli, A., Theodossioub, I., & Vasileioua, E. (2008). Jobs as Lancaster goods: Facets of job satisfaction and overall job satisfaction. *Journal of Socio-Economics, 37*, 1906-1920.
- Trauer, T., & Mackinnon, A. (2001). Why are we weighting? The role of importance ratings in quality of life measurement. *Quality of Life Research, 10*, 579-585.

- Wu, C. H. (2008a). Examining the appropriateness of importance weighting on satisfaction score from range-of-affect hypothesis: Hierarchical linear modeling for within-subject data. *Social Indicators Research*, 86, 101–111.
- Wu, C. H. (2008b). Can we weight satisfaction score with importance ranks across life domains? *Social Indicators Research*, 86, 468–480.
- Wu, C. H., Chen, L. H., & Tsai, Y. M. (2009). Investigating importance weighting of satisfaction scores from a formative model with partial least squares analysis. *Social Indicators Research*, 90, 351–363.
- Wu, C. H., & Yao, G. (2006a). Do we need to weight satisfaction scores with importance ratings in measuring quality of life? *Social Indicators Research*, 78, 305–326.
- Wu, C. H., & Yao, G. (2006b). Do we need to weight item satisfaction by item importance? A perspective from Locke's range-of-affect hypothesis. *Social Indicators Research*, 79, 485–502.

Table 1

Results of Moderated Regression on Global QoL Criterion Variables with Survey Data (N = 90)

	Step 1 Satisfaction R^2	Step 2 Importance ΔR^2	Step 3 Satisfaction x Importance ΔR^2
Single-item measure	.237**	.103	.108
LSIA	.248**	.101	.027
SWLS	.385***	.083	.017

*** $p < .001$; ** $p < .01$

Table 2

Results of Moderated Regression on Criterion Variables with Simulated Data (N = 10,000)

	Step 1 Satisfaction R^2	Step 2 Importance ΔR^2	Step 3 Satisfaction x Importance ΔR^2
Criterion 1	.309***	.287***	.056***
Criterion 2	.230***	.221***	.047***
Criterion 3	.163***	.168***	.038***
Criterion 4	.118***	.117***	.021***
Criterion 5	.080***	.070***	.015***

*** $p < .001$