Assessing the Precursors and Attainment of Wellbeing in Higher Education

Nicole Brocato, Eranda Jayawickreme

Wake Forest University

Presented at NCME, April 2017

Author Note

Nicole Brocato is with the Office of Institutional Research at Wake Forest University.

Eranda Jayawickreme is with the Department of Psychology at Wake Forest University.

Correspondence concerning this paper should be addressed to Nicole Brocato, Office of Institutional Research, Wake Forest University, PO Box 7427, Winston-Salem, NC 27109. Contact: brocatnw@wfu.edu

Please note that as the measure has evolved, so has our terminology. Most notably, this paper references "precursors," now referred to as "pathways."

Abstract

We propose that measures of wellbeing should assess not only the attainment of wellbeing, but also whether respondents have the necessary precursors for wellbeing. Using structural equation modeling, we present results from an ongoing effort to develop a measure that assesses wellbeing, its precursors, and its correlates.

Assessing the Precursors and Attainment of Wellbeing in Higher Education

The number of wellbeing measures for use in higher education settings is growing rapidly, as are theories about wellbeing content, assessment, and intervention. According to the Engine Model of Well-being[1] (Jayawickreme, Forgeard, & Seligman, 2012), wellbeing attainment is dependent on the extent to which individuals have access to wellbeing precursors that cover a wide range of content, and include both resources outside the individual (e.g., money, social support) and resources within the individual (e.g., values, beliefs, knowledge, and social and behavioral skills). The benefit of measuring these wellbeing precursors is that end-users of the assessment will have more information with which to create targeted, evidence-based interventions to support wellbeing. Creating such a measure presents two psychometric challenges: (a) it is very large and is therefore prone to missing data, and (b) it requires full structural equation modeling rather than just factor analytic techniques. We start this paper by providing a brief conceptual review of the Engine Model of Well-being and its use in Wake Forest University's Wellbeing Assessment, a questionnaire designed to assess wellbeing in college undergraduate students. We then discuss our approach to evaluating the Wellbeing Assessment's reliability and validity before spending the bulk of the *Methods* and *Results* sections of this paper discussing our use of structural equation modeling as a method of establishing structural validity (using Messick, 1995). We conclude the paper with a summary of our structural validity evidence and next development steps for the Wellbeing Assessment.

[1] Authors vary in their use of the terms *well-being* and *wellbeing*. We use the terms as they were used in the cited works.

Introduction to Wake Forest University's Wellbeing Assessment

Comparable measures

The number of wellbeing measures for use in higher education settings is growing, as are theories about wellbeing content, assessment, and intervention. Reviewing those works is beyond the scope of this paper. To get a crude sense of the number of works in this field, we searched the online psychology research database *PsycINFO* for the term "well-being in college." That search returned 815 articles, the first of which was published in 1947. Of those 815 articles, 684 were published in the year 2000 or later.

Our team is aware of five carefully researched wellbeing measures designed specifically for the higher education setting. One focuses solely on physical health (National College Health Assessment (NCHA), American College Health Association); one focuses solely on mental health (Healthy Minds Study (HMS), Healthy Minds Network); and three others cover multiple dimensions of wellbeing (Thriving Quotient (Schreiner); Gallup-Purdue Index (Gallup, 2014); The College Student Subjective Wellbeing Questionnaire (Renshaw & Bolognino, 2014)). Of these five measures, the three measures that cover multiple dimensions of wellbeing are closest in content to Wake Forest University's Wellbeing Assessment. However, all three of those models focus on the attainment of wellbeing across those multiple dimensions without an examination of what we would call *precursors* to wellbeing. In the next section, we discuss how we integrated wellbeing precursors into the Wellbeing Assessment.

The Engine Model of Well-being

Wake Forest University's Wellbeing Assessment was developed using the Engine Model of Well-being (Jayawickreme et al., 2012), which is a conceptual framework for wellbeing theories that advocates for the measurement of both wellbeing and its precursors. The potential precursors to wellbeing are broad in scope, and include both resources outside the individual (e.g., money, social support) as well as resources within the individual (e.g., values, beliefs, knowledge, and social and behavioral skills).

In the Wellbeing Assessment, we have focused on precursors that we think university programming staff (i.e., staff who develop programs to support wellbeing or wellbeing correlates) will be able to address in some form or another. For instance, we have not asked questions about precursors that focus on students' home neighborhoods because university programming staff have no influence over home neighborhoods. However, we have asked questions about students' perceptions of campus resources and faculty and staff because those are resources that programming staff can directly alter, either by providing direct programming or by working with university policy-makers. We have also included items about students' relationships with their family and parents. Although university programming staff are unlikely to be able to address students' relationships with their parents, they can help students negotiate those relationships, and do on some occasions have the ability to interact directly with family.

We hope that by including these precursors in a measure of wellbeing, end-users of the measure will have more information with which to develop targeted, evidence-based interventions. Stated differently, rather than trying to directly improve students' wellbeing, wellbeing might be more easily intervened upon if programming staff address the precursors to wellbeing. As an example, one of the dimensions in the Wellbeing Assessment is Meaning. Having meaning in life is a somewhat abstract concept, and programming staff without training in a clinical psychology approach that addresses meaning in life may not know how to improve meaning. The precursor items in the Meaning dimension include the presence of values-supporting people in the students' lives, opportunities to enact meaning, and self-awareness about meaning in life. In theory, these precursors are somewhat more concrete intervention targets than meaning in life, and interventions can be designed to improve these precursors. Potential examples include: helping students learn how to build conversations about meaning into their close relationships, encouraging university faculty or staff to acknowledge or support conversations about meaning in life, providing organized campus activities that are meaningful to students, and helping students learn how to continually maintain awareness of their values and meaning in life.

Wake Forest University's Wellbeing Assessment

The version of the Wellbeing Assessment that was used to generate the data analyzed in this study included the following 10 dimensions of wellbeing that we believe the higher education setting can and should support: Meaning, Purpose, Engagement in Activities, Positive Attitudes, Relationships (friendships and romantic relationships), Belonging, Identification with all Humanity, Standing up to Discriminatory Behavior, Volunteerism, and Intellectual Curiosity. Since these data were collected, we added an eleventh dimension: Commitment to Lifelong Service; we do not yet have data for that dimension. Of the original 10 dimensions, the methods described in this paper have been applied to all but the Relationships dimension. The Relationships dimension was not analyzed with the methods in this paper because it is being used to screen for a list of risky and healthy behaviors, and it does not include a set of items that measure the attainment of wellbeing in relationships.

In keeping with the Engine Model of Well-being (Jayawickreme, et al., 2012), each wellbeing dimension is associated with a unique set of items that measures (a) the attainment of wellbeing in that dimension (called the *outputs* or *output items*) and (b) the precursors to wellbeing in that dimension. The output items are a small set of reflective indicators (Bollen & Bauldry, 2011) that are intended to measure distinct latent factor of wellbeing achievement in a given dimension. Most dimensions have between 3 and 4 output items. The precursors to wellbeing are measured with causal, single-item indicators (Bollen & Bauldry, 2011) that we treat as manifest variables. We had originally designed the dimensions to each have approximately six precursor items, but as we will see in the *Results* section, we are still working to identify viable precursor items for some of the dimensions. The Positive Attitudes dimension does not have any precursor items because it consists of a larger group of reflective items that measure multiple types of positive attitudes.

In the Engine Model, precursors lead to the attainment of wellbeing. As such, in our statistical models the latent factor that is measured by the output items is regressed on the precursor items. The Meaning dimension items are presented in Table 1 as an example. The statistical modeling results presented later in this paper are based on this version of the items; these items were subsequently edited based on those statistical modeling results as well as results from cognitive process interviews.

Table 1
Items from the Meaning dimension

Precursors
Intra- and extrapersonal resources
1. School offers valued opportunities
2. Family supports meaningful activities
3. Able to talk to faculty/staff
4. Self-awareness of values
5. Family has influenced sense of meaning
6. Friends have influenced meaning
Outputs
Whether wellbeing has been achieved
7. Daily activities are meaningful
8. Life feels meaningful
9. Live life in a meaningful way

Note: These items were the foundation for the statistical modeling results presented later in this paper. These items have been edited based on those results.

In addition to including items associated with the wellbeing dimensions, the Wellbeing Assessment also includes items that assess outcomes which we hypothesize are associated with wellbeing, such as: anxiety, depression, loneliness, feelings about career and financial prospects, engagement in academics, and intent to transfer. In our statistical models, we regressed these items on the latent factors formed by the output items in the wellbeing dimensions. Some of these outcomes were measured with single-item indicators, and some were measured with sets of reflective indicators. We provide more detail about these items in the *Description of items* subsection of the *Methods* section.

A simplified path model for the wellbeing precursor items, the wellbeing output (latent factor) items, and the outcome items is presented in Figure 1. In that figure, the items in the wellbeing dimensions are enclosed by the yellow box. The outcome items (which are not part of the wellbeing dimensions) are in the green hexagon on the right-hand side of the diagram.

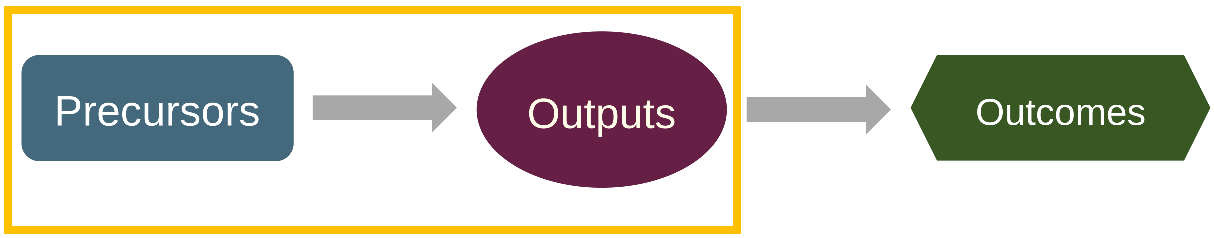

Figure 1. Path model for the elements of the Wellbeing Assessment.

We generated structural equation models for each wellbeing dimension in the Wellbeing Assessment using the basic format provided in Figure 1. A full path model is presented in Figure 2. This is the full path model for the Meaning dimension, presented without any of the parameter estimates. The items on the far left enclosed in the blue box (*meaning1 – meaning4*) are the Meaning dimension's precursors. The central, burgundy box includes is the latent factor of Meaning (the circle labeled *meaning*) and the three items by which it is measured (the boxes labeled *meaning7 – meaning9*). The items on the far right of Figure 2 enclosed in the green box are the outcomes; these outcomes consist of both single-item indicators and several latent factors. We review all the item types in Figure 2 in more detail in the *Description of items* subsection of the *Methods* section. The main goal of Figure 2 is simply to demonstrate the conceptual foundation for the measure and its statistical modeling.

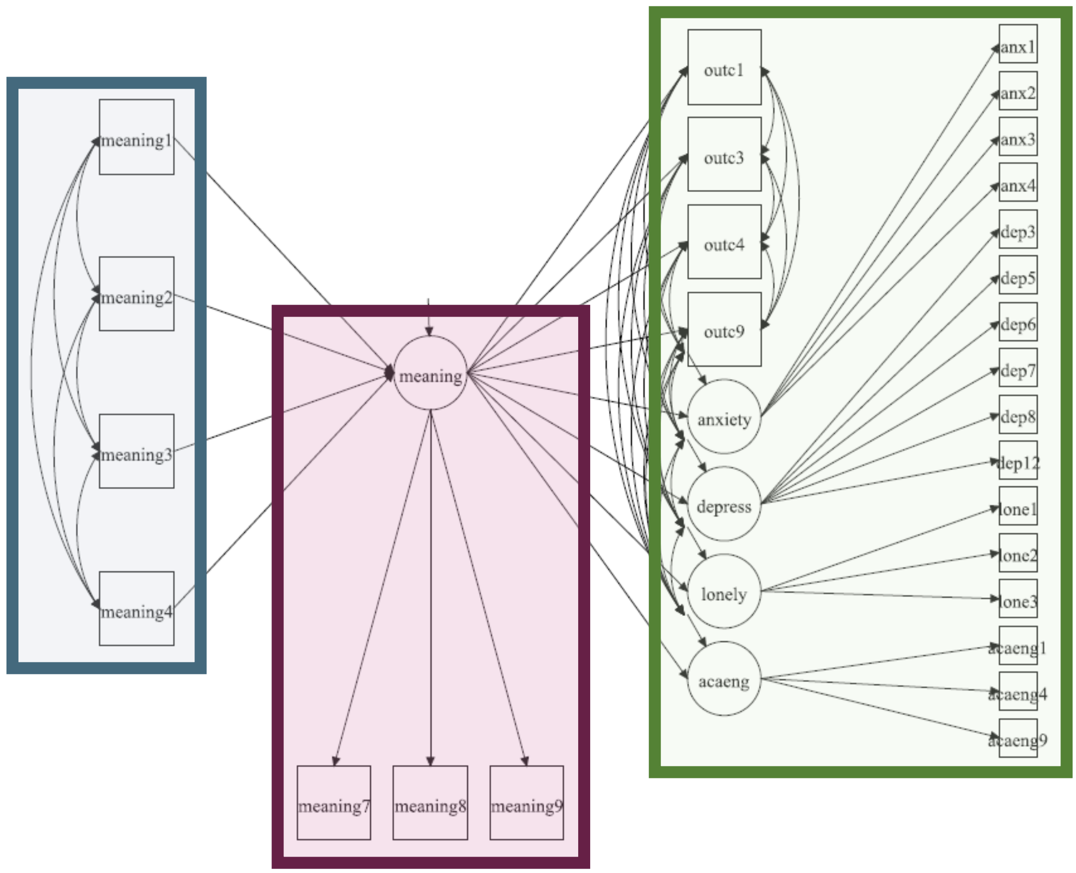

Figure 2. Full path model of Meaning dimension. Items meaning5 and meaning6 are missing from this model because they did not fit the final model well. Those results will be presented and discussed in further detail later in this paper.

Summary of the Introduction

In the *Introduction* section, we described Wake Forest University's Wellbeing Assessment, which was created using the Engine Model of Well-being (Jayawickreme et al., 2012). Using the Engine Model of Well-being makes the Wellbeing Assessment a unique measure of undergraduate student wellbeing because it measures both the attainment of wellbeing and the precursors needed to attain wellbeing. The Wellbeing Assessment encompasses 11 dimensions of wellbeing. Each of those dimensions has a small set of reflective indicator items that assess the attainment of wellbeing in that dimension (called *output* items). Each dimension also has a set of single-item causal indicators that measure the precursors to attaining wellbeing in that dimension. In latent variable models, each dimension's latent wellbeing factor is regressed on the precursors for that dimension.

The Wellbeing Assessment also contains a number of items that assess correlates of wellbeing; those correlates are measured both with reflective items that form a latent factor and with single-item indicators. In latent variable models, those correlates are regressed on the wellbeing latent factors.

In the next section, we describe our methods. We begin with our approaches to reliability, validity, which include a discussion of the role of structural equation modeling. We then present our structural equation modeling methods, which include methods that account for missing data and categorical (ordinal) data.

Methods

In this section, we begin by explaining why we chose to use structural equation modeling (SEM) to help evaluate the measure's reliability and validity. We then briefly discuss how we have used structural equation modeling in those efforts. After we present those methods, we then explain in detail the structural equation modeling methods we chose, which include methods specific to missing data and categorical variables.

SEM, Reliability, and Validity

Over the following several pages, we will introduce our rationale for using structural equation modeling and then explain how we chose to evaluate reliability and validity. We will provide a brief summary of these topics before providing detailed explanations of our SEM methods.

Rationale for using SEM. As we discussed in the *Introduction*, Wellbeing Assessment measures 11 dimensions of wellbeing. Those dimensions are based on the Engine Model of Wellbeing, and therefore contain (a) precursor items that measure the extent to which respondents have the resources necessary to attain wellbeing and (b) output items that measure the attainment of wellbeing. The Wellbeing Assessment also contains (c) outcome items that are meant to measure constructs that are theoretically associated with the presence or absence of wellbeing. Our conceptual model claims that the (a) precursors cause the (b) outputs which in turn partially cause the (c) outcomes. It is because of this directionality in our conceptual model that we chose to statistically model the Wellbeing Assessment using structural equation modeling rather than relying solely on factor analytic methods.

Reliability. Reliability can be thought of as the extent to which the score on a measure consistently reflects people's true score on the construct being measured (Anastasi & Urbina, 1997; Peters, 2014; Raykov & Marcoulides, 2011; Revelle & Zinbarg, 2009; Yang & Green, 2011). Evaluations of scale reliability are often made using alpha (Cronbach, 1951), which is an estimate of internal consistency (and is sometimes confused with being an estimate of unidimensionality). However, alpha has been criticized for making unrealistic assumptions and providing underestimates of reliability (Raykov, 1997; Sijtsma, 2009; Yang & Green, 2011).

As an alternative to alpha, we will use Raykov's (1997) composite reliability for congeneric measures model (CRCMM; Raykov, 1997). However, even this approach is somewhat problematic, both because it presumes that item scoring will be based on sum scores, and because it was designed for use with continuous data. To explain a bit further, Raykov and Marcoulides (2011) state that typically when researchers are concerned about the reliability of a scale, they are concerned about the reliability of a summed set of items, which is called a *composite*. Although we do not yet know whether we plan to use sum scores, we are often asked by reviewers to provide a reliability estimate that would be appropriate for a summed score. We are typically asked to provide alpha, which is only an accurate representation of composite reliability under several strict assumptions: no correlated errors, unidimensionality, and high loadings on the latent factor. Otherwise, alpha can either over- or underestimate composite reliability. Unlike alpha, the CRCMM provides an estimate of that composite's reliability using pattern coefficients (i.e., factor loadings) from latent factor measurement models. When data are modeled as continuous, those pattern coefficients provide information about how much observed variance in an item is due to the latent factor, which means those pattern coefficients also provide information about how variance in an item is due to error.

Assuming no correlated errors among the items, that the set of items has been tested for unidimensionality, and that the measurement model has been identified by fixing the variance of the latent factor to 1, the formula for CRCMM is:

$$\hat{\rho}_x = \frac{(\widehat{b_1} + \cdots + \widehat{b_p})^2}{(\widehat{b_1} + \cdots + \widehat{b_p})^2 + \widehat{\theta_1} + \cdots + \widehat{\theta_p}} \quad [1]$$

where $\hat{\rho}_x$is the estimated reliability of a composite, $(\widehat{b_1} + \cdots + \widehat{b_p})$ is the sum of the item-factor loadings (i.e., the pattern coefficients), and $\widehat{\theta_1} + \cdots + \widehat{\theta_p}$ is the sum of the error variances for the items (and not the standard error of the parameter estimates, a mistake sometimes encountered in published works that apply this formula).

In the Wellbeing Assessment, the only items treated as composites are the output items within each wellbeing dimension. The output items are the several items in each dimension that are intended to assess whether wellbeing in that dimension has been attained. As mentioned at the start of the previous paragraph, the CRCMM—and other approaches like it, such as alpha—is problematic for the evaluation of the composite reliability evaluations because it assumes that the data can be treated as continuous in statistical models. As we will explain in later sections of this paper, we treated our data as categorical. As we stated in the previous paragraphs, structural equation modeling methods for evaluating continuous data treat item responses (i.e., the actual, observed data) as having variance that can be explained by an underlying latent construct; the pattern coefficients can be treated as estimates of the amount of variance in an item that is explained by the latent factor. In contrast, structural equation modeling methods for categorical data model the probabilities of respondents providing one response option over the others. The continuous response variable is latent and presumed, rather than observed. As a result, the pattern coefficient for an item is associated with that item's latent, unobserved response variable, rather than with the observed data. Because the point of a reliability estimate to provide information about the observed data, if a composite reliability estimate was calculated for categorical data, that estimate would apply to the latent response variable rather than the observed data. Better explanations of these topics can be found in texts on item response theory, such as: Embretson and Reise (2000), and Samejima (1969), as well as postings on the M*plus* discussion board (Muthén, October 15, 2014 – October 16, 2014; Muthén & Muthén, March 22, 2014 – March 24, 2014, May 2, 2006 – June 25,2013).

Raykov and Marcoulides (2011) suggest parceling categorical items and calculating the CRCMM estimate based on those parcels, but our dimensions did not contain enough items to parcel. We instead used the original CRCMM formula, and we treated the data as continuous when generating CRCMM estimates. The CRCMM reliability estimates were generated in M*plus* software version 7.4 (Muthén & Muthén, 2015) using the MLR estimator (and without the TYPE = categorical option), a maximum likelihood estimator that accounts for missing data under a missing-at-random assumption. Instructions for generating the syntax can be found in the Raykov and Marcoulides (2011) text (Chapter 7).

Validity. We chose to use Messick's (1995) conceptualization of validity because it is thorough and broad in scope. It covers the topics found in most other conceptualizations of validity in one unified framework. Messick's validity framework includes six components (a summary of them can be found on page 745 of the original article):

1. Content: evidence for the extent to which a measure's items are relevant to and representative of the construct being measured. Messick describes most forms of content validity evidence as coming from experts and theory.
2. Substantive: theoretical rationales regarding which processes participants should use to respond to measure items, and empirical evidence indicating whether they actually do. Messick

describes substantive evidence as coming from studies such as cognitive process interviews, task observations, and latent variable (i.e., factor) modeling.

3. Structural: evaluations of the appropriateness of the scoring method given the construct being measured
4. Generalizability: the extent to which score interpretations are valid across groups, settings, and tasks
5. External: multi-trait, multimethod evidence for convergence with and discrimination from alternative measures; also, evidence for appropriateness and utility of any scoring paradigms
6. Consequential: evaluation of the extent to which a measure's use creates any injustices via bias, fairness, etc.

The Wellbeing Assessment is still in development, and not all areas of validity have been evaluated. Only some of these areas can be evaluated using structural equation modeling methods. In the list below, we briefly explain how we have evaluated each area of validity. If we have not yet evaluated that area of validity, we explain our plans for doing so.

1. Content: We have conducted multiple and ongoing rounds of conferring with substantive experts in multiple fields, both applied and theoretical.
2. Substantive: Three rounds of cognitive process interviews have been conducted to evaluate whether participants respond to items using the information and cognitive processes we designed the items to capture. Additional evidence comes from structural equation modeling results we provide in this paper.
3. Structural: In progress. We have already consulted with theoretical experts who have stated that Likert-type response options (and, therefore, associated scoring methods) are more appropriate than other structures such as Thurstone scaling or test-type questions. Once the current pilot is complete, we will have a sufficiently large sample size to trial multiple

scoring methods (e.g., factor scores, summed scores) and ask applied experts for their opinions on the impacts and utility of those various methods.

4. Generalizability: In progress. We are currently collecting a data set that will be sufficiently large to allow for measurement invariance studies using structural equation modeling methods (e.g., between gender groups, race/ethnicity groups, class standing groups, and more as applied and theoretical experts suggest).
5. External: Pending. After content, structural, and generalizability validity are evaluated, additional studies will be needed to assess external validity. These studies have not yet been designed. Depending on the nature of the studies, they may be evaluated using structural equation modeling methods.
6. Consequential: Ongoing. We began addressing consequential validity at the start of the measure development process by creating clear guidelines for the contexts under which the measure and its data could be used. The measure development team determined that the measure is designed to assess typically-aged undergraduate students and that the data are designed for use in the anonymous aggregate by university staff for the purposes of improving wellbeing-related programming. We are exploring other potential uses of the data, such as providing feedback to respondents about their answers. We plan to continue examining consequential validity throughout the lifespan of the measure to ensure that it is not used in a way that could propagate injustices.

Summary of SEM, reliability, and validity methods. Reliability and validity are complex concepts that can be evaluated using numerous methods. We have chosen to evaluate composite reliability using the CRCMM, a method that is based on structural equation modeling item-factor loadings. We have chosen to evaluate validity using Messick’s (1995) six-part conceptualization. Of those six parts, three can be evaluated using structural equation modeling methods: structural, generalizability, and external.

In this paper, we present results pertaining to structural validity. Results for the other three are in progress or pending.

Starting with the next paragraph, we will detail our SEM methods, including discussions of missing data and categorical variable methods.

Structural Equation Modeling Methods

In this section of the paper, we describe our structural equation modeling methods. The primary goal of the structural equation modeling was to evaluate structual validity. To achieve that goal, we needed to demonstrate that the items in each dimension could be modeled using the structural equation conceptual model in Figure 1, which meant determining (a) that we had a set of output items that formed a latent factor, and (b) that the latent factor could be regressed on a set of single-item indicators that explained a meaningful portion of the latent factor's variance. As a lesser goal, we also needed to demonstrate that the single-item and latent factor outcome items could be regressed on the wellbeing latent factor, which would explain a meaningful portion of the outcome items' variances.

As we will discuss further in the *Discussion* section, the absence of relevant intervention literature made defining the phrase "meaningful variance" impossible. We instead evaluated variance amounts within and relative to our own results, a shortcoming we hope to ameliorate through intervention and other applied research. That intervention research will also help us explore external validity.

For each of the dimensions in the Wellbeing Assessment, we created structural equation models based on Figure 1. As we explained in the section *Wake Forest University's Wellbeing Assessment* above, we applied those models to nine dimensions in the measure: meaning, purpose, engagement in activities, positive attitudes, belonging, identification with all humanity, standing up to discriminatory behavior, volunteerism, and intellectual curiosity.

We start the description of our methods by describing the items in the Wellbeing Assessment in more detail and explaining why we chose to analyze the data as ordinal (or, in the language of M*plus*: *categorical*) rather than as continuous. We then talk about our assumption of the missing data mechanism as missing at random, and why we used multiple imputation to account for missing data and then analyzed that data using a diagonally-weighted least squares estimator. After these descriptions are complete, we present the results of our analyses.

Description of items. The items within our wellbeing dimensions (i.e., the box and circle on the left of Figure 1, which are equivalent to the left and center boxes of Figure 2) were structured using 6-point, Likert-type response options of the form: *strongly agree, agree, slightly agree, slightly disagree, disagree, strongly disagree.*

In addition to the wellbeing dimension items, we also measured and modeled outcomes that we hypothesized would be associated with wellbeing (i.e., the hexagon on the right side of Figure 1, which is equivalent to the right-most box in Figure 2). We began the structural equation modeling process for each wellbeing dimension using the same set of outcome items, although we sometimes had to drop outcome items from the process if there were problems with imputation convergence or with model-fitting. We used two types of outcome items: single-item indicators, and multi-item reflective indicators of latent constructs. The single-item indicators were structured using 4-point, Likert-type response options of the form: *very likely, somewhat likely, not very likely, and not likely at all*. The multi-item reflective indicators of latent constructs were structured using 4-point, relative frequency response options of the form: *not at all, several days, over half the days, nearly every day*. The exception to these structures was the set of Academic Engagement items, which were originally a wellbeing dimension. These items were structured using the same 6-point, Likert-type response option structure as the wellbeing dimension items. A summary of the outcome items is presented in Table 2.

Data recoding. The response options for our wellbeing dimension items were ordered *strongly agree* to *strongly disagree*. As a result, our wellbeing dimension items were structured so that lower scores on an item were associated with higher levels of wellbeing (unless the item was negatively valenced). Within our outcome items, the item sets measuring anxiety, depression, and loneliness were also coded so that low scores reflected lower levels of the anxiety, depression, and loneliness. Thus, a positive structural coefficient between these item sets and the wellbeing latent factors actually indicates that higher levels of wellbeing are associated with lower levels of anxiety, depression, and loneliness. However, this reading is counter-intuitive, so we reverse coded these items. In the *Results* section, negative structural coefficients can be interpreted to mean that higher scores on the latent wellbeing factors are associated with lower levels of anxiety, depression, and loneliness. The remaining outcomes items did not need recoding.

Table 2
Summary of outcome items used for each wellbeing dimension's structural equation model

Single-item indicators
Response options: Very likely – not likely at all
Outc1: How likely are you to transfer to another 4-year college
Outc3: How likely are you to leave this school before graduation and without further plans
Outc4: How likely are you to graduate from this school
Outc9: How likely are you to have a job within 6 months of graduation
Multi-item, reflective indicators (4-point, relative frequency response options)
Response options: Not at all – nearly every day
*Anxiety**
7 items requesting symptom frequency over a 2-week period
*Depression**
13 items requesting symptom frequency over a 2-week period
*Loneliness**
3 items requesting symptom frequency over a 2-week period
Multi-item, reflective indicators (6-point, Likert-type response options)
Response options: Very likely – not likely at all
Academic engagement
3 items requesting strength of endorsement

Note: Items marked with an asterisk (*) were reverse-coded in analyses

Decision to treat items as categorical. We chose to model our data using methods designed for ordinal data for three reasons. First, a number of our items had only four response options. Second, the data for the items with six response options tended to be highly asymmetrically distributed. Under both of these conditions, categorical estimation methods outperform continuous methods, as we describe in the next paragraph. Third and finally, part of the purpose of our analyses was to determine which items in the Wellbeing Assessment could be retained and which should be rejected or edited; given the high-stakes goals for these analyses, we wanted to be sure that we minimized parameter estimation bias.

It is fairly common to see Likert-type and other ordinal items with five or more response options treated as continuous in structural equation modeling efforts, including confirmatory and exploratory factor analysis models. However, a recent and thorough simulation study by Rhemtulla, Brosseau-Liard, and Savalei (2012) demonstrated that categorical estimation methods outperform robust maximum likelihood methods (when the data are treated as continuous) when ordinal items have between five and seven categories. The more asymmetric the data are distributed, the truer this becomes. Although the bias produced by continuous-method robust maximum likelihood methods was fairly small (<=10%), in studies such as the current study where decisions are being made about whether and which items to retain, minimizing bias is critical.

Missing data methods. Within each of our wellbeing dimensions, data was missing at rates of up to 63% for individual variables. This high rate of missingness was due to a partial planned missing data design. The version of the survey we administered to collect the data being reported on here was nearly 330 items in length, a length that we thought was far too long. To reduce respondent burden, we asked participants to complete only some of the dimensions in the measure. Data were also missing above and beyond this planned missingness. We chose to assume that the additional missing data were missing at random. We chose to presume that the data were missing at random for two reasons. First, our data were highly unlikely to have been missing completely at random. Second, sensitivity analyses to assess whether the data were missing not at random would have been complex, inconclusive, and made stringent assumptions about that were not appropriate for our data (e.g., that they were normally distributed; Enders, 2010). We talk more about planned missing data designs in the *Discussion* section.

We used the multiple imputation method for data missing at random suggested by Enders (2010) to impute missing observations using M*plus* software version 7.4 (Muthén & Muthén, 2015). We chose multiple imputation over the use of full information maximum likelihood estimator because multiple imputation would allow us to analyze the data with a diagonally-weighted least squares estimator (WLSMV), which produces less biased parameter estimates than maximum likelihood when data are ordinal and non-normal (Asparouhov & Muthén, 2010; Li, 2015).

To conduct the multiple imputation procedure, we used an unrestricted variance/covariance model (Asparouhov & Muthén, 2010), and imputed 50 data sets to maximize statistical power (Enders, 2010). We conducted separate imputations for each wellbeing dimension. Those imputations included the items in the wellbeing dimension plus all the outcome items.

For models analyzed over imputed data sets using the WLSMV estimator (i.e., the measurement and structural models in the *Results* section), M*plus* provides parameter estimates that are averaged over the data sets and standard errors that are based on Rubin's (1987; as cited in Muthén & Muthén, 2015) formula for combining variances within and between imputation sets.

Structural equation modeling steps. We used a multi-step approach to the structural equation modeling process. Those steps were based on the recommendations of (in alphabetical order): Asparouhov and Muthén (2009); Hayduk and Glaser (2000); Kline (2015); Marsh, Morin, Parker, and Kaur (2014); the M*plus* supplemental literature (e.g., discussion board postings and website supplemental materials); and Mulaik and Millsap (2000).

In our multi-step approach, we first analyzed measurement models before we analyzed structural models. Unlike most guidelines for conducting multi-step analyses, we started with confirmatory rather than exploratory factor models because we had clear hypotheses about how the items should be modeled. We conducted those confirmatory factor models on the *output* items in each dimension. If those pre-determined items did not form a well-fitting measurement model, we added or removed items based on theory to identify alternative item sets. Once a measurement model was identified, the rest of the model was then added, including: the single-item precursor items, the single-item outcome items, and the 4 latent outcome factors.

Because these analyses are being conducted early in the life of the project to develop the Wellbeing Assessment model, we took an exploratory approach to the structural equation models. When models were not fitting well, we added or removed items or sets of items in an attempt to improve model fit. As the *Results* section will show, we were not always able to achieve good model fit. We discuss the possible reasons for these challenges in the *Discussion* section.

By default, M*plus* correlates the residuals of endogenous variables (Muthén & Muthén, 2015). The residuals are correlated for the purposes of reducing structural coefficient misfit; high residual correlations may indicate a need to correct the model or add further latent factors (Muthén & Muthén, June 23, 2010 – November 11, 2016). We retained these correlated residuals for two reasons. First, we have reason to believe that many of our endogenous variables (our *outcome* items and factors) will be associated with each other. For instance, the mood factors are likely to be associated with each other, and the items measuring intent to transfer are likely to be associated with each other. We also thought that confidence in future job prospects was likely to be associated with mood, as was academic engagement. Second, we retained these correlated residuals because our modeling process was exploratory, and the correlated residuals would provide us with information about where our structural equation models (and perhaps our underlying conceptual model) might need further refinements. We provide brief interpretations of these correlated residuals in the *Results* and *Discussion* sections, but focus more discussion on the structural coefficients.

Model identification. In models where only measurement models were generated, the models were identified by setting the variance of the latent factor to 1. This allowed us to examine the pattern coefficients (i.e., factor loadings) of all the items in the measurement model. When full structural models were generated, model identification was achieved by setting the pattern coefficient for the first item in each factor to 1. This allowed us to examine the amount of variance explained in those latent factors.

Evaluation of model fit. As we stated above in the *Structural Equation Modeling Methods* introductory paragraphs, the primary goal of the structural equation modeling was to evaluate structural validity. To achieve that goal, we needed to demonstrate that the items in each dimension could be modeled using the structural equation conceptual model in Figure 1, which meant determining (a) that we had a set of output items that formed a latent factor, and (b) that the latent factor could be regressed on a set of single-item indicators that explained a meaningful portion of the latent factor's variance. As a lesser goal, we also (c) needed to demonstrate that the single-item and latent factor outcome items could be regressed on the wellbeing latent factor, which would explain a meaningful portion of the outcome items' variances.

To achieve those four goals for the structural equation models, we used multiple approaches for evaluating model fit, which we will present here somewhat out of order so that the rationales for those approaches are clear.

Model fit for goal (b). To evaluate goal (b) of having a set of output items that formed a latent factor, we used standard model-fitting techniques. Consistent with Kline's (2015) recommendations, we evaluated model fit at the level of both overall model fit and item-level fit. We began by evaluating global fit using Yu's (2002) guidelines for non-normal, categorical items analyzed in non-imputed data: RMSEA should be .05 or lower; CFI should be at least .95, and WRMR should not be much more than 1.0. In keeping with Kline's (2015) recommendation, we also determined that chi-square tests of model fit should be non-significant, although chi-square *p*-values are not available in M*plus* for analyses conducted with imputed data. For the same reason, we do not present RMSEA confidence intervals for any of our models.

We evaluated fit at the item level by examining correlation residuals (i.e., the difference between the observed and model-estimated item correlations). M*plus* can be used to generate correlation residuals when both items in a pair are ordinal (i.e., categorical). Using Kline's (2015) suggestion, correlation residuals greater than .10 were considered possible indicators of model misfit for the items in those correlations. Because there is no established cutoff for how many correlation residuals can exceed .10, we used our judgment and the patterns of residuals to make determinations about model fit.

We also added the requirement that the output items should have standardized pattern coefficients of at least .71. In models analyzed using methods for continuous data, standardized pattern coefficients of at least .71 mean that the latent factor explains at least 50% of the variance in the items; in turn, items with at least 50% of their variance explained by the latent factor are less likely to be complex indicators (Kline, 2015). Although we used methods for categorical rather than continuous data, no such guidelines for categorical data exist. We therefore retained this guideline so that we would have some method for evaluating the pattern coefficients, which are likely to have been lower had we estimated the models using methods for continuous data (Bandalos, 2014; Li, 2015; Rhemtulla et al., 2012).

Model fit for remaining three goals (a), (c), and (d). Although we could have used the guidelines in the preceding paragraphs to evaluate the remaining three goals for our structural equation modeling efforts, we thought those guidelines failed to capture intent of these three goals to make judgments about the utility of the structural equation models. Because the fit methods described for evaluating goal (b) are largely determined by the fit of the measurement models, we also thought those methods would not be sufficiently sensitive to the structural components of the models.

For goal (a)'s purpose of determining whether we could model the wellbeing dimensions using the conceptual structural model in Figure 1, we considered this goal met if a measurement model with acceptable fit could be generated, and if there was some suggestion that the precursor and outcome items had some meaningful associations with the latent wellbeing factor. We somewhat arbitrarily set that value as a standardized structural coefficient of .2, simply because it is large enough that it is unlikely to reflect only error.

For goals (c) and (d) in which we tried to determine whether "meaningful amounts" of latent factor variance were explained, we again selected a somewhat arbitrary value of 30% for the wellbeing dimension latent factors (measured by the output items), although implementation and intervention studies may later demonstrate that more or less variance explained is a better floor. We expected the outcome latent factors to be far more multiply-determined, and so we set the minimum amount of variance explained to 10%, another arbitrary value that needs confirmation from implementation and dissemination studies. Ultimately, we also made judgments based on the relative amounts of variance explained within our own results; for example, we will see in the *Results* section that the amount of variances explained in the latent wellbeing dimension constructs measured by the output items ranged from 14% to 88%, and that the dimensions with robust models all had levels of variance explained that exceeded 50%. Even though our predetermined floor was 30%, the fact that our robust models all had 50% or more of variance explained led us to view lower values and potential signs of problems.

Summary of model fit. In summary, our structural equation models needed to meet *all* of the following fit criteria: (a) pattern coefficients for the output items of at least .71, (b) RMSEA of .05 or lower, (c) CFI of .95 or higher, (d) WRMR of not much more than 1, (d) standardized structural coefficients of at least .2, and (e) wellbeing dimension latent factor variances explained of at least 30%.

Summary of Methods

In this *Methods* section, we described our approaches to reliability (CRCMM; Raykov, 1997), validity (Messick, 1995), and a multi-step structural equation modeling approach to evaluate structural validity. We explained that our structural equation modeling approach was exploratory. We used multiple imputation to account for high rates of missing data, which are both missing completely at random and also presumed missing at random. We analyzed the imputed data with a diagonally weighted least squares estimator because our data were ordinal and their distributions were highly asymmetric. Finally, we used an eclectic approach to evaluating model fit, one that was based not only on guidelines for model fit indices, but one that also attempted to account for the goals of our modeling efforts.

In the next section, we present the results of structural equation modeling efforts based on an April 2016 pilot survey at Wake Forest University. These results will demonstrate that for approximately half of the Wellbeing Assessment's dimensions, our modeling approach: is feasible overall, supports structural validity, and provides early evidence supporting Wellbeing Assessment's use of the Engine Model of Well-being.

Results

We have thus far submitted the measure to two rounds of cognitive interviews and two rounds of pilot administrations to an undergraduate population. A third round of cognitive interviewing and piloting is currently underway. The results presented here are based on the most recent pilot, which was conducted in April 2016 at Wake Forest University. That pilot yielded a usable sample size of 661 participants. We considered records unusable when participants did not consent or when the record was entirely blank outside of the consent items.

We start the presentation of the results by providing some descriptive information about the data, including respondent demographics, skew and kurtosis, and missingness. We next begin presentating model results by presenting overall model fit for the final structural equation model generated for each dimension. We then provide details of the pattern coefficients and CRCMM reliability estimates for the output and outcome items. We finish with the structural coefficient estimates before concluding with the *Discussion* section.

Descriptives

Demographics. The sample was unbalanced on gender and class standing (Table 3), and was therefore weighted in subsequent analyses. The sample's race and ethnicity demographics were roughly equivalent to those of the overall Wake Forest University undergraduate population: 76% White (population = 73%), 14% Asian (population = 11%), 9% Black (population = 6%), 1% other (population = 10%). We did not weight the sample for race/ethnicity demographics.

Skew and kurtosis. We used the software *SAS Enterprise Guide version 6.1* (SAS Institute, Inc., 2013) to evaluate the skew and kurtosis of our data. As we stated in the *Methods* section above, our data were highly non-normally distributed. The skew of our variables ranged from -3.80 to 4.61, with an average of .68. Kurtosis ranged from -2.00 to 24.95 with an average of 1.04.

Missingness. Data for individual variables was missing at rates of up to 63%. As we discussed in the *Methods* section, this high rate of missingness was due in part to a partial planned missing data design. At most, however, data should have been missing at no more than 50%; thus, not all of our missing data was due to planned missing data. We did not conduct tests of whether the data were missing completely at random such as t-tests or Little's (1988) omnibus test; we did not conduct these tests because the high number of variables being analyzed and the length of survey made it unlikely that the data were missing completely at random. We did not conduct sensitivity analyses for missing not at random because those analyses make strong assumptions that are not appropriate for our data, such as a normal distribution; further, many analyses conducted using methods for data that are missing at random are robust to the presence of data that are missing not at random (Enders, 2010).

Table 3
Descriptive statistics for pilot sample

	Count	Proportion	Wake Population
Gender			
Male	192	29.05%	47.4%
Female	454	68.68%	52.6%
Other	2	0.30%	
Missing	13	1.97%	
Total possible	*661*		
Class standing			
Freshman	190	28.74%	27.16%
Sophomore	168	25.42%	24.80%
Junior	115	17.40%	24.51%
Senior	175	26.48%	23.42%
Missing	13	1.97%	.10%
Total possible	*661*		

Note: Proportions may not sum to 100% due to rounding error.

Model Results

Overall model fit. As we discussed in the *Methods* section, we generated structural equation models nine of the dimensions in the Wellbeing Assessment measure. Each of those models included (a) single-item precursors, (b) latent wellbeing factor indicators, and (c) outcomes that were measured by single items and by latent factors. In our structural equation models, (c) was regressed on (b) was regressed on (a).

We start this section by presenting the fit indices for each of our 9 final structural equation models. For the sake of space, we do not present the results from the earlier steps in the structural equation modeling process, namely the measurement models. The results of those fit indices are presented in Table 4.

Table 4

Fit indices and statistics for final structural equation models

	RMSEA	CFI	WRMR	Chi-square (df)	n	Variance explained
Meaning	.05	.98	1.22	714.46 (286)	612	55%
Purpose	.05	.98	1.38	630.79 (274)	593	66%
Activity Engagement	.06	.97	1.64	919.95 (290)	612	83%
Positive Attitudes	.05	.99	.99	731.90 (318)	612	--
Belonging	.05	.97	1.36	798.52 (308)	612	66%
Identification with all Humanity	.07	.98	1.47	753.15 (211)	612	37%
Standing up to Discrimination	.12	.91	1.96	377.28 (39)	612	14%
Volunteerism	.06	.97	1.24	191.37 (60)	612	45%
Intellectual Curiosity	--	--	--	--	--	

Note: All fit indices and statistics represent averages over the set of 50 imputed data sets. The column (n) represents the average number of cases in each data set. The Variance Explained column represents the average amount of variance explained in each dimension's latent factor by the precursor items. The Purpose dimension had very high levels of missing data, leading to a smaller average number of cases. The Positive Attitudes dimension does not have precursor items.

As can be seen in Table 4, we were unable to achieve acceptable model fit for the last four dimensions, although the problems with Volunteerism are less obvious than the problems for the remaining three. In all four of those models, we had to reduce the number of model components because of convergence problems with imputation process. Further explanations of that problem and why they threaten the seemingly good fit of the Volunteerism model are provided in the *Structural Coefficients* section when we discuss Table 7. The reduced number of model components is the reason the chi-square degrees of freedom are so much smaller in these models than they are in the first five models. Among the poor-fitting models, we were able to identify a set of items measuring a latent wellbeing factor for that dimension in all but the Intellectual Curiosity dimension, but those measurement models were formed by items with poor pattern coefficients (as seen in Table 5). We were unable to impute the variables for the Intellectual Curiosity dimension; imputations would not converge even after 10,000 iterations. The final four dimensions in Table 4 also have lower rates of variance explained in the latent wellbeing factors compared to the first five dimensions, indicating that the precursor items have lower associations with the latent wellbeing factors. In some cases, such as Standing up to Discrimination, there were only a few precursor items, which also led to low amounts of variance explained. Full details about the structural coefficients and the number of precursor items can be found in Table 7 and its associated discussion.

Pattern coefficients and CRCMM estimates for output items. As we stated in the *Methods* section, these reliability estimates were generated by treating the data as continuous and using a robust maximum likelihood estimator (MLR). The pattern coefficients (factor loadings) are a component of the CRCMM, and so we include them with the CRCMM estimates in Table 5. For the sake of comparison, we also provide the pattern coefficients from the structural equation models generated using the WLSMV estimator and the imputed data sets.

We were not surprised to see that the pattern coefficients for the MLR models were lower than those for the WLSMV models for all but one of the 32 pattern coefficients provided in Table 5 (for further discussion, see: Bandalos (2014) and Li (2015)). We were surprised to see how much the parameter estimates differed in the poor-fitting models. In the models that fit well (Meaning – Belonging), the parameter estimates differed by no more than about 15%. Among the poor-fitting models (Identification with All Humanity – Intellectual Curiosity), some WLSMV estimates were nearly 50% larger than their MLR counterparts.

Table 5

CRCMM estimates, MLR pattern coefficients, and WLSMV pattern coefficients for the output items in each dimension of the Wellbeing Assessment

	CRCMM estimate	MLR pattern coefficients	WLSMV pattern coefficients
Meaning	.91	.79	.80
		.94	.98
		.94	.96
Purpose	.90	.84	.87
		.83	.86
		.89	.96
		.78	.88
Activity Engagement	.88	.77	.82
		.89	.90
		.85	.95
Positive Attitudes	.93	.78	.84
		.70	.76
		.75	.82
		.80	.85
		.66	.74
		.81	.85
		.91	.95
		.81	.85
Belonging	.83	.87	.94
		.82	.89
		.72	.82
Identification with all Humanity	.67	.58	.61
		.62	.63
		.50	.77
		.64	.79
Standing up to Discrimination	.21	.95	.86
		.61	.74
		.15	.41
		.21	.38
Volunteerism	.86	.76	.85
		1.00	1.00
		.69	.76
Intellectual Curiosity	--	--	--

Note: All pattern coefficients are presented in STDYX standardization, an approach that standardizes parameter estimates using all variables in the model. The dimensions vary in the number of output items used to represent the latent wellbeing construct for that dimension. As previously noted, the model for the Intellectual Curiosity dimension could not be estimated.

Pattern coefficients and CRCMM estimates for outcome items. In Table 2 in the *Methods* section, we listed our outcome items as being of two types: (a) four single-item indicators that we treated as manifest variables, and (b) four sets of outcome items that measured the latent constructs of: anxiety, depression, loneliness, and academic engagement. In Table 6 we present CRCMM estimates for those four latent constructs, and we also provide the MLR and WLSMV pattern coefficients. These four sets of items all had well-fitting measurement models, high pattern coefficients, and high CRCMM estimates. As we saw in Table 5, the difference between the MLR and WLSMV parameter estimates for these well-fitting models varied little, although they were consistently higher in the WLSMV models.

Table 6
CRCMM estimates, MLR pattern coefficients, and WLSMV pattern coefficients for the outcome items

	CRCMM estimate	MLR pattern coefficients	WLSMV pattern coefficients
Anxiety	.92	.83	.87
		.91	.95
		.89	.93
		.80	.87
Depression	.91	.84	.89
		.86	.91
		.82	.93
		.74	.82
		.74	.82
		.78	.87
Loneliness	.91	.80	.89
		.88	.93
		.94	.98
Academic engagement	.87	.84	.87
		.86	.90
		.78	.86

Note: All pattern coefficients are presented in STDYX standardization, an approach that standardizes parameter estimates using all variables in the model.

Structural coefficients. Although we intended to include the same outcome items in each dimension's structural equation model, we were unable to do so because of estimation problems either in the imputation process or the structural equation modeling process. In Table 7 we present the structural coefficient estimates for each dimension's structural equation model. When we presented the model fit information in Table 5, we noted that some models did not fit well because we had to remove model components.

In Table 7, we indicated when model elements were removed to improve the imputation process (and therefore not available for inclusion in the modeling process) and when they were removed to improve model fit. We do not supply tests of statistical significance for these coefficient estimates because the large size of our sample meant that tests of most coefficient estimates returned statistically significant results, even when those coefficient estimates were low (e.g., <.20). Another reasons we did not provide statistical significance information is that there is no existing literature against which to interpret the magnitude of the results. In other words, we know that the results are statistically significant, but we do not know if they are meaningful. For those more applied interpretations, we need a body of intervention research to provide context. We return to this topic in the *Discussion* section.

The results of the structural models showed that improved mood was associated with higher levels of wellbeing in the Meaning, Purpose, Activity Engagement, Positive Attitudes, Belonging, and Identification with All Humanity dimensions. Because these measures are cross-sectional, we cannot know the direction of the causality, only that these constructs are associated with each other. For the Discrimination and Volunteerism dimensions, the mood items were not associated with these dimensions' latent wellbeing factors; we were not able to even impute the mood items in these dimensions' imputation sets. A review of the raw correlation matrix confirmed that the mood items were not associated with the items measuring these dimensions' latent wellbeing factors (i.e., the output items).

Table 7

WLSMV structural coefficient estimates for each wellbeing dimension

	Meaning	Purpose	Activity engagement	Positive attitudes	Belonging	Identification all humanity	Discrimination	Volunteerism	Intellectual curiosity
<u>Precursors</u>									
Precursor 1	.19	.19	.17		.23	.61	-.17	.03	xx
Precursor 2	.21	-.05	.15		.32	xx	.37	.08	xx
Precursor 3	.16	.32	.23		-.32	xx		.33	xx
Precursor 4	.39	.06	.25		-.32			.44	xx
Precursor 5		.02	.25		.17				
Precursor 6		.38	.22						
<u>Outcomes</u>									
Outcome 1	-.03	-.16	-.27	-.36	-.50	xx	-.11	<.01	xx
Outcome 3	.15	-.10	--	-.31	-.18	xx	--	-.20	xx
Outcome 4	.16	.21	--	.40	.27	.53	--	.07	xx
Outcome 9	.14	.01	.10	.23	.19	.11	-.16	.06	xx
Anxiety	-.48	-.17	-.30	-.52	-.18	-.31	xx	xx	xx
Depression	-.62	-.25	-.39	-.67	-.31	-.33	xx	xx	xx
Loneliness	-.52	-.22	-.43	-.56	-.63	-.28	xx	xx	xx
Academic engagement	.37	.30	.51	.53	.35	.56	.23	.14	xx
<u>Model fit</u>									
%Variance explained	55%	66%	88%	(no precursors)	66%	37%	14%	45%	xx
Acceptable model fit?	Yes	Yes	Yes	Yes	Yes	No	No	Marginal	No

Note: All pattern coefficients are presented in STDYX standardization, an approach that standardizes parameter estimates using all variables in the model. The dimensions vary in the number of precursor items. As previously noted, the model for the Intellectual Curiosity dimension could not be estimated. The row *%Variance explained* is the amount of variance explained in the latent wellbeing construct (measured by the output items) by the precursor items in that dimension. Precursor item cells are blank when no such precursor item existed for that dimension. Cells that contain dashes indicate that items (or sets of items) were removed from the model in an effort to improve model fit. Cells that contain “xx” indicate that items were removed to make the imputation process possible.

For the sake of space, we do not present the residual correlation values for the endogenous variables, nor do we present the latent factor correlations. Instead, we provide a brief summary of them in this paragraph. The values for the correlated residuals among the endogenous variables typically ranged between absolute values of .4 - .6 among the mood latent factors, and .4 - .6 among the three outcome items assessing intent to transfer. The remainder tended to have absolute values of .25 or less. Patterns for the correlations among the latent factors were similar, with the absolute values for the mood factors ranging from approximately .6 - .8, and the absolute values for the outcome items assessing intent to transfer ranging from approximately .4 - .6. Absolute values for the remaining correlations were typically less than .3. Taken together, we believe these patterns of correlations among the endogenous variable residuals and latent factors indicates a need for reducing or combining the items that assess intent to transfer because these items all measure a very similar construct. We are less inclined to argue for combining the items that measure the mood factors (e.g., into a larger “negative mood” factor) because the mood factors are associated with different wellbeing factors in different ways. For instance: in the model for the Meaning dimension, the structural coefficient for Depression is larger than the coefficient for Loneliness; this pattern is reversed in the model for the Belonging dimension.

Summary of Structural Equation Modeling Results

The purpose of our structural equation modeling efforts was to assist in our evaluation of the Wellbeing Assessment’s structural validity. We were attempting to determine the structural validity of both our application of the Engine Model of Well-being (Jayawickreme, et al., 2012) and the extent to which the latent wellbeing traits were associated with a range of outcome items and latent factors. We were also attempting to assess reliability, which we did using CRCMM estimates (Raykov, 1997).

We used a multi-step structural equation modeling approach in which we first tested measurement models before generating structural coefficient estimates. Because of the early nature of this project, these structural models were exploratory in nature; we added or removed components in order to maximize model fit.

Our data were collected from a sample of Wake Forest University undergraduate students. While they were representative of the larger Wake Forest population in terms of race and ethnicity, the sample was not representative for gender or class standing. We therefore weighted subsequent analyses for gender and class standing.

We chose to analyze our data using the WLSMV estimator in M*plus* because our data had high levels of skew and kurtosis, and also because our data were ordinal. To accommodate high levels of missing data, we first imputed 50 data sets using an unrestricted variance/covariance model. We conducted separate imputations for each of the nine dimensions in the Wellbeing Assessment for which we conducted structural equation modeling.

For five of the nine modeled wellbeing dimensions, we were able to generate models that matched our conceptual model (Figures 1 and 2) and had acceptable levels of fit and sufficiently large pattern coefficients as gauged against our pre-determined guidelines. Those dimensions were: Meaning, Purpose, Activity Engagement, Positive Attitudes, and Belonging. For these five dimensions, the CRCMM estimates and also provided evidence of reliability.

For four of the nine modeled wellbeing dimensions, we were unable to generate models that matched our conceptual model while having acceptable levels of fit and sufficiently large pattern coefficients. In the Identification with All Humanity and the Standing up to Discrimination dimensions, pattern coefficients were low, model fit was poor, and we had to remove items from the imputation process. In the Volunteerism dimension, model fit was acceptable as were pattern coefficients for the outputs, but the items had little to no association with many of the outcomes; thus, while we had evidence for the structural validity of the Engine Model portion of the Volunteerism dimension, we did not have evidence for our associated theory that Volunteerism would be associated with our predetermined set of outcomes. Finally, the Intellectual Curiosity items were so poorly interrelated that we were not able even to generate imputations for this set of items.

Discussion

Summary of Results and Implications

We are in the early stages of developing the Wellbeing Assessment. We have completed two rounds each of pilot administrations and cognitive interviews, and are in the process of conducting a third of each. The data we analyzed for this paper were collected during our second pilot administration, in the Spring 2016 academic semester.

For five of the nine dimensions in the Wellbeing Assessment, we have evidence of structural validity and reliability using structural equation modeling methods: Meaning, Purpose, Activity Engagement, Positive Attitudes, and Belonging. In those dimensions, the output items form latent factors and the precursor items explain what we think might be a meaningful amount of variance in those factors. We hope that those precursor items will provide campus programming staff with more concrete intervention points than the latent wellbeing factors themselves. One of the next steps in our research project is to begin examining intervention possibilities and the extent to which the precursor items are valid indicators of intervention targets. We plan to do this by conducting intervention research with other higher education institutions.

For these dimensions with well-fitting structural models, we also have evidence that the dimensions' latent factors were associated with a set of outcome items that assessed intent to transfer, optimism about job prospects, anxiety, depression, loneliness, and academic engagement. These outcomes are important in and of themselves because in addition to having implications for student wellbeing, they have implications for services use and needs, retention, graduation rates, and persistence in securing post-graduation employment. As with the other applications of this measurement project, these potential implications must be assessed in longitudinal research before they can be confirmed.

The remaining four dimensions continue to prove problematic: Identification with All Humanity, Standing up to Discrimination, Volunteerism, and Intellectual Curiosity. To further explore the possible causes for these challenges, we are in the process of conducting qualitative cognitive process interviews. The most immediate next step in our research program is to review the pending results from a recent qualitative cognitive process interviewing study to examine whether and how we can restructure those dimensions. We also have evidence from the structural models presented in this paper that the mood items may not be relevant outcomes for all these dimensions.

Taken together, the results from the structural equation models presented here suggest that there are sufficient theoretical grounds for treating wellbeing as a construct that can be modeled in dimensions that are founded on the Engine Model of Well-being (Jayawickreme et al., 2012).

Strengths, Limitations, and Future Directions

We present this project's strengths, limitations, and future directions together because these topics are highly intertwined for this project.

One of this project's mutual strengths and weaknesses is its use of a planned missing data design. The planned missing data design we used to collect the data analyzed for this paper was problematic because students received either the entirety of a dimension or none of it. As a result, the missing data methods we used to analyze the data could not account for responses from students who did not take the portion of the measure being modeled. In the version of the questionnaire currently being piloted, we used a more refined planned missing data design in which all students answer at least a few items in each dimension (Graham, Taylor, Olchowski, & Cumsille, 2006; Harel, Stratton, & Aseltine, 2015; Little & Rhemtulla, 2013; Raghunathan & Grizzle, 1995). Our analyses should therefore face fewer estimation challenges, more accurately reflect our overall sample, and be more generalizable to the overall population of undergraduate college students.

Another of this project's mutual strengths and challenges is its reliance on the Engine Model of Well-being (Jayawickreme et al., 2012). Our use of that model is a strength because it makes this project unique in its intention to directly inform interventions. However, because our use of this model is so new, there is not existing intervention literature against which we can gauge the magnitude of our structural coefficients or the amount of variance explained in our latent wellbeing constructs (i.e., the constructs measured by the output items). Thus, while we have many significant statistical tests of our models' parameter estimates (which we did not report), we do not have any way of knowing if those parameter estimates are also meaningful in applied settings. Stated differently: the tests of our models' parameter estimates are statistically significant, but do those parameter estimates provide meaningful guidance to measure end-users such as campus programming staff? For that, we need intervention research, which is a next component in our overall research program. We also argue that these concerns can be expressed about most measures of wellbeing: because we have so little intervention research in the field of wellbeing, we do not know how large structural coefficients must be to indicate meaningful associations between interventions and results; similarly, we simply do not know how much change in latent wellbeing variables (i.e., our output items) is needed for participants to experience meaningful change in their lives.

A limitation of the current project is its newness. We have yet to explore several features of validity, including structural, generalizablility, or external validity. Plans are in place for evaluating structural and generalizability validity: for both, we can conduct additional statistical analyses with the results from our in-progress pilot effort, which includes 10 partner schools outside of Wake Forest and a potential participant pool of 29,000 students. From that pilot, we should have a large enough sample to evaluate scoring methods and conduct between-groups analyses, including differential item functioning studies.

The project’s newness all means that the modeling efforts presented here were largely exploratory and, thus, potentially capitalized on chance. We found evidence, for instance, that we may need to combine or eliminate items measuring intent to transfer. However, these exploratory and iterative results may not generalize to a new data sample. In our next modeling effort, we need to test the models we generated here and compare those new results to those in this paper.

To summarize this project’s strengths, limitations, and future directions, we have preliminary evidence that the Engine Model of Well-being (Jayawickreme et al., 2012) can be applied to the creation of a reliable and valid wellbeing measure intended for use in the higher educational setting, which may allow university programming staff to more easily create targeted interventions to support undergraduate student wellbeing. We also have evidence that wellbeing in a variety of dimensions is associated with a number of other important outcomes that could have implications for services use, retention, and persistence in pursuing post-graduation employment. Taken together, we believe the Wellbeing Assessment’s application of the Engine Model of Well-being provides a strong and unique foundation for evaluating and intervening upon undergraduate student wellbeing.

References

American College Health Association. *National College Health Assessment*. http://www.acha-ncha.org/

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, N.J.: Prentice Hall.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multipdisciplinary Journal, 16,* 397-438.

Asparouhov, T., & Muthén, B. (2010). Multiple imputation with Mplus. Retrieved from https://www.statmodel.com/download/Imputations7.pdf

Bandalos, D. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal, 21*, 101-116.

Bollen, K., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods, 16*, 265-284.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L, & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists.* New York: Psychology Press.

Enders, C. (2010). *Applied missing data analysis.* New York, NY: The Guilford Press.

Gallup (2014). *Gallup-Purdue Index Inaugural National Report.* http://products.gallup.com/168857/gallup-purdue-index-inaugural-national-report.aspx

Graham, J., Taylor, B., Olchowski, A., & Cumsille, P. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11*, 323-343.

Harel, O., Stratton, J., & Aseltine, R. (2015). Designed missingness to better estimate efficacy of behavioral studies—application to suicide prevention trials. *Journal of Medical Statistics and Informatics, 3*, Article 2. http://www.hoajonline,com/journals/pdf/2053-7662-3-2.pdf. DOI: 10.7243/2053-7662-3-2.

Healthy Minds Network. *Healthy Minds Study*. http://healthymindsnetwork.org/

Jayawickreme, E., Forgeard, M., & Seligman, M. (2012). The engine of well-being. *Review of General Psychology, 16*, 327-342.

Kline, R. (2015). *Principles and practice of structural equation modeling, 4th edition.* New York: The Guilford Press.

Li, C.-H. (2015). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*. doi: 10.3758/s13428-015-0619-7.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198-1202.

Little, T., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives, 7*, 199-204.

Marsh, H., Morin, A., Parker, P., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10,* 85-110.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Muthén, B.O. (October 15, 2014 – October 16, 2014). M*plus* discussion board post: Ordinal variables, estimators, and missing data. Retrieved from: http://www.statmodel.com/discussion/messages/9/20451.html?1413504738

Muthén, B.O., & Muthén, L.K. (March 22, 2014 – March 24, 2014). M*plus* discussion board post: Composite reliability for categorical variables. Retrieved from: http://www.statmodel.com/discussion/messages/23/18432.html?1395686232

Muthén, B.O., & Muthén, L.K. (May 2, 2006 – June 25,2013). M*plus* discussion board post: How to calculate composite reliability. Retrieved from: http://www.statmodel2.com/discussion/messages/9/258.html?1372284067

Muthén, B.O., & Muthén, L.K. (June 23, 2010 – November 11, 2016). M*plus* discussion board post: Covariance among endogenous factors. Retrieved from: http://www.statmodel.com/discussion/messages/11/256.html?1479949515

Muthén, L.K., & Muthén, B.O. (1998-2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Peters, G-J. (2014) The alpha and the omega of scale reliability and validity. *The European Health Psychologist, 16*, 56-69.

Raghunathan, T., & Grizzle, J. (1995). A split questionnaire survey design. *Journal of the American Statistical Association, 90*, 54-63.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*, 173-184.

Raykov, T, & Marcoulides, G. (2011). *Introduction to psychometric theory*. London: Taylor and Francis.

Renshaw, T., & Bolognino, S. (2014, December 17). The College Student Subjective Wellbeing Questionnaire: A brief, multidimensional measure of undergraduate's covitality. *Journal of Happiness Studies*. DOI 10.1007/s10902-014-9606-4

Revelle, W., & Zinbarg, R. (2009). Coefficients alpha, beta, omega, and the GLB: Comments on Sijtsma. *Psychometrika, 74*, 145-154.

Rhemtulla, M., Brosseau-Liard, P.E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*, 354-373.

Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from: http://www.psychometrika.org/journal/online/MN17.pdf

SAS Institute, Inc. (2013). SAS Enterprise Guide Version 6.1.

Schreiner, L. *The Thriving Quotient.* http://www.thrivingincollege.org/

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120.

Yang, Y., & Green, S. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*, 377-392.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* (Doctoral dissertation). Retrieved from the M*plus* website: http://www.statmodel.com/download/Yudissertation.pdf