Technical Report:
Wellbeing Assessment Methods and Psychometric Properties for the Fall 2020 Administration

Xinyu Ni
Nicole W. Brocato
Samantha Shang

Version date: April 12, 2021

Suggested citation: Ni, X., Brocato, N. W., & Shang, S. (2021). *Technical report: Wellbeing Assessment methods and psychometric properties for the fall 2020 administration*. Wake Forest University.

https://www.acha.org/wp-content/uploads/Fall-2020-WFU-WBA-Technical-Report.pdf

Table of Contents

Document Purpose

The purpose of this document is to describe the methods used in fall 2020 to administer the Wellbeing Assessment, condition the data, and score dimensions with *outcome* items. Information about the validity and reliability of the Assessment's factor scores can be found in the Spring 2019 Technical Report.

Major Differences Between 2019 and 2020

Although we might normally focus only on differences between surveys' technical features, the 2019-2020 academic year was unique in its deep disruptions to the typical functioning of higher education. We include those disruptions in this list because they affect longitudinal trends in aggregate scores, relevant interpretations of the scores, and potentially measurement invariance. Although we did not find any non-invariance between this year's scores and the prior year's scores, the possibility exists that future years could be affected by non-invariance due to these contextual effects.

1. Pandemic
    a. By mid-March of 2020, the SARS-coV-2 virus (and the COVID-19 disease it causes) had become a worldwide pandemic. Because social distancing (i.e., remaining at least 6 feet apart from other people and not gathering indoors) was the most important factor in slowing the spread of the pandemic, most higher education institutions were forced to immediately transition all their academic and student engagement activities to online platforms in mid-March. Social distancing also disrupted the economy, resulting in significant stressors for students' families and threats to many students' access to food, housing, and care. Social distancing in and of itself isolated students from normal social activities.

2. Social unrest in reaction to systemic racism
    a. Although deaths in the U.S. of African-American people (particularly men) due to police interactions have a long history of being disproportionately higher relative to the deaths of people who hold other racial and ethnic identities, a string of these violent, police-related deaths caught public attention and social media during the late spring and summer of 2020. Calls for racial equity resonated across higher education, and many institutions began publicly grappling with long histories of slavery and inequitable access.
3. Change in number of required dimensions
    a. This year we required 15 dimensions instead of 18; we provide the list of dimensions in the *Measure* section.
4. ACT2_1-ACT2_3 items
    a. The ACT2_1-ACT2_3 items changed wording between the 2019 and 2020 fall administrations, but the change in wording did not significantly impact factor scores. The details of wording changes were summarized in Table 1. Under the *Factor Scores* section, we describe our evaluations to ensure that this change in item wording did not significantly impact factor scores.

Measure

The Wellbeing Assessment was developed using four rounds of cognitive interviews (Fall 2015 – Spring 2018) and five pilot administrations (two local administrations in Fall 2015 and Spring 2016, and three multisite administrations in 2017, 2018, and 2019). The Wellbeing Assessment includes 18 dimensions: happiness, anxiety, depression, loneliness, social anxiety, life satisfaction, self-esteem, optimism, perseverance, coping, activity engagement, academic engagement, belonging, friends,

meaning, purpose, civic values – moral, and civic values – political. The modeling and scoring procedures in this document include only these dimensions.

Each year, a changing set of dimensions is optional so that we can reduce respondent burden while gathering robust data on our dimensions of substantive interest for upcoming research. For the fall 2020 administration, the following 15 dimensions were administered to all the participants: happiness, anxiety, depression, loneliness, social anxiety, life satisfaction, self-esteem, optimism, perseverance, activity engagement, academic engagement, belonging, meaning, purpose and coping. In addition to the activity engagement dimension, items from all the other 14 dimensions were the same. Three items in the activity engagement dimension were changed which were listed in the Table 1.

Survey Procedures

Recruitment

The Wellbeing Assessment is administered annually at universities and colleges that volunteer to participate, resulting in a large and diverse convenience sample of undergraduate college students.

The fall 2020 administration was unusual because of the pandemic caused by the SARS-Cov-2 virus, which in turn causes the COVID-19 disease. The “coronavirus pandemic” or “covid pandemic” started in late 2019, and by mid-March most higher education institutions had shut down and/or moved classes online. Ultimately, a total of 32 schools participated in the Fall 2020 administration, one of which was a school in Mexico. In this report, we only reported numbers for the schools in the U.S. and excluded the school in Mexico. All 31 U.S. schools were 4-year institutions. They were public and private institutions varying in size from fewer than 1000 undergraduate students to more than 20,000 undergraduate students.

Individual schools provided participation incentives, with some schools providing no incentives, some providing small incentives to all students, and some providing larger lottery items. Schools' incentives were reviewed for appropriateness and IRB compliance by the research team.

Planned Missing Data Design

Because the Wellbeing Assessment is very large (approximately 250 items), it was administered in previous years using a planned missing data design to reduce respondent burden and improve data quality by reducing missingness due to attrition. The Spring 2019 Technical Report provides more detail about that procedure. So that participating schools could have more complete data, we did not use a planned missing data design this year. We instead made fewer dimensions required in fall 2020 administration, but all dimensions were distributed to all schools.

Survey Randomization

To reduce missingness from attrition (i.e., participants not finishing the survey), we randomized many of the substantive sections of the survey. The mood items (happiness, loneliness, anxiety, depression, social anxiety), some of the demographic items, and the work & academic plas items were presented at the start of the survey without randomization. The rest of the item sets used in the factor scoring were presented randomly. The items within the sets were always presented in the same order.

Participants

Between the months of late September and early December 2020, approximately 128,495 students in the U.S. were invited to participate; 14,069 (11.27%) consented. After removing graduate students and students who answered less than two items, the final sample is 10,900. Because of unplanned missing data due to attrition, the usable cases in any particular analysis may be lower than this number. Table 2 (see Appendix) summarized the descriptive statistics with survey weights for the sample.

School Characteristics

Of the 31 participating schools in the U.S.:

- Public/private: 14 were private, 17 were public
- Size: 18 schools had undergraduate FTE enrollments of <10,000; 13 had enrollments of >10,000
- Region:
  - North: 2 schools
  - South: 17 schools
  - Midwest: 8 schools
  - West: 4 schools

Data Conditioning

Missing data

For the variables used to generate the factor scores in the 15 dimensions all participants received, unplanned missing data rates range from 0.10% to 19.62%. If you are reading this report in preparation for conducting analyses with data we have provided you, we strongly recommend you evaluate rates of missing data for the variables in your study.

For the variables used to generate the factor scores, we tested for MCAR using Little's (1988) MCAR test using the TestMCARNormality function in the MissMech package (Jamshidian et al., 2014) in RStudio 3.51 (RStudio Team, 2016). MCAR was not rejected (χ^2 = 17267.42, *df* = 17,177, *p* = 0.312). The sample data of the fall 2020 was missing completely at random.

To reduce bias in the parameter estimates caused by missing data, we used full information maximum likelihood (FIML) estimators to generate the factor scores. FIML reduced bias in parameter estimations for missingness under an MCAR or an MAR mechanism (i.e., auxiliary variables).

Weighting

To improve the generalizability of the data to the general population of undergraduate students, we weighted the data using a raking procedure via the survey package (Lumley, 2004, 2019). We used joint distributions for gender and race/ethnicity per 2018 NCES statistics (National Center for Education Statistics, 2019) for gender and race/ethnicity of undergraduate student enrollment (https://nces.ed.gov/programs/digest/d16/tables/dt16_306.10.asp). Because we used a joint distribution, the raking procedure is effectively a calibration procedure. The NCES data does not capture all the race/ethnicity categories captured in our data, and so we adjusted the national proportions to create the additional categories reflected in our data. The raking procedure provides calibration weights that reduce bias associated with under/oversampling demographic groups in the population, thereby improving generalizability to the general population. This procedure does not correct for all possible sources of survey error. Weights were used in the CFA modeling that provides the dimension factor scores.

The raw weights for the data range from 0.59 to 6.15. The upper range of the weights is somewhat extreme, and so we trimmed the upper weights to 3; the resulting weights ranged from 0.59 to 3. The literature does not provide clear guidance about when weights are "extreme" or which method to use when trimming weights. We somewhat arbitrarily chose an upper cutoff of 3 because it shows up in numerous informal rules-of-thumb and because it makes some general sense: "counting" any individual in the data set as more than 3 times their original record seems like a strong interpretation of the data. Table 2 includes values for both the untrimmed (RAW_WT_Value) and trimmed (TRIM_WT_Value) weights so that you can use whichever weights you think are appropriate. ***We used the TRIM_WT value to conduct all the scoring and other models presented in this document.***

Item Distributions

Skewness for the variables was modest on average (mean = -0.16, median = -0.39). However, some items did display greater skewness than is typically recommended, with a maximum value of 1.76 and a minimum value of -1.32.

Kurtosis were more varied, although were modest on average (mean = -0.32, median = -0.51). The values ranged from -1.37 to 2.00.

In all our modeling, we used a robust maximum likelihood estimator (MLR) to generate standard errors that were robust to non-normalities in the item distributions.

Factor Scores

Included Dimensions

The 18 dimensions measured with latent factor structures were modeled with 57 items. Of those 18 dimensions, 15 dimensions were administered to all participants in the fall 2020 administrations; the remaining 3 dimensions (9 items in Friendships, Civic Orientation - Moral, and Civic Orientation - Political) were not administered to any participants. The codebook describes which items from each dimension were included in the latent variable model used to generate the factor scores.

Scoring Procedures Overview

As stated above, the Fall 2020 administration used only 15 core dimensions instead of 18, and we changed the wording of the Activity Engagement items. To ensure that this year's factor scores could be compared to prior years' scores, we used the following procedures: (1) created a joint dataset (n = 22,821) combining 2019 spring and 2020 fall data; (2) measurement invariance examinations across spring 2019 and fall 2020; (3) evaluating effects of the changed ACT2_1-ACT2_3 wording; (4) concurrent calibration with the joint dataset to generate unscaled fall 2020 factor scores; (5) linking fall 2020 scores calculated from step 4 to the previously released spring 2019 scores. Step 2 was optional because item parameters estimated from a concurrent calibration procedure are in the same scale, and scores

calculated with the estimated parameters from a concurrent calibration are comparable directly. We did it anyway because we aimed to examine whether students having the same latent abilities would be measured equally for both 2019 fall and 2020 fall surveys with an actual 15-common dimension model and an assumed 18-dimension model. Also, the results would provide evidence for the construct validity for the 2020 fall survey.

In the sections that follow, we detail these steps.

1. Create a joint dataset

In the first step, we created a joint dataset combining data collected from 2019 and 2020 fall administrations. Only items that would be analyzed in the model were selected. Demographic and all the other background variables were not included in the joint dataset. The joint dataset was used to conduct a series of measurement invariance analysis and calculate the factor scores with the *lavaan* (Rosseel, 2012) package in RStudio 3.51 (RStudio Team, 2016).

Descriptive Statistics for the joint data

The total number of students in the joint dataset is 22,821. Among them, 11,921 respondents are from the 2019 spring administration and 10,900 are from the 2020 fall administration. Also, there were 59 columns in the joint dataset including 57 survey items, a weight (TRIM_WTS) item and a group item (year). Among the 57 survey items, content of 45 items were exactly the same across two surveys. Twelve items differed across the two surveys: three Activity Engagement items and nine items in the dimensions of Friendships, Civic Orientation - Moral, and Civic Orientation - Political.

The three items in the Activity Engagement dimension (ACT2_1, ACT2_2, ACT2_3) are not exactly the same (described below in the Step 3). We treated the Activity Engagement items as the same across the two years to examine whether measurement invariance could be achieved. In step 3, we further examined the effects of the changed ACT2_1-ACT2_3 wording on the scoring procedure.

There were nine items in the three dimensions (items in the dimensions of friends, civic values - moral, civic values - political) that were only administered in 2019. These items were specified as missing data for students participating in the survey in 2020.

Weights

We did not re-calculate the weights for the joint data because the weights for each year in the original datasets were generated through a raking procedure with different joint distributions. In both years, the weights were calculated through a raking procedure that referred to NCES gender and race/ethnicity for undergraduate student enrollment statistics. In 2019, we used NCES 2018 reports that were based on 2017 statistics (National Center for Education Statistics, 2018). In 2020, we used NCES 2019 reports that were based on 2018 statistics (National Center for Education Statistics, 2019). The range of the weights for the joint dataset is from 0.52 to 3.00.

Missingness

For the variables in the joint dataset used to generate factor scores for the 18 dimensions, the missing rate at the item level ranged from 0.29% to 69.55%. For the items that were not administered in the fall 2020 administrations (items in the dimensions of friends, civic values - moral, civic values - political), missingness ranged from 54.01% to 69.55%. For the items that were administered in both years, the missing rate at the item level ranged from 0.29% to 30.99%.

2. Measurement Invariance Examinations

The purpose of the measurement invariance analysis was to examine whether students having the same latent abilities would be measured equally for both spring 2019 and fall 2020 surveys. This measurement invariance testing examined whether the Assessment had the same latent measurement structure across the two administrations, whether the administrations in the two years could therefore be scored using comparable methods, and whether models generated with the two years would

generate factor scores that could be compared across the years. We did two sets of measurement invariance tests: (a) a set with 15 dimensions both years had in common; and (b) a set with all 18 dimensions in which Friendships, Civic - Moral, and Civic - Political were treated as missing for fall 2020.

We started with a 15-dimension model because all items in the 15 common dimensions were administered across two years. The full information maximum likelihood (FIML) estimator was used to reduce bias in parameter estimations caused by missing data. Then we conducted a series of measurement invariance analysis with an "assumed" 18-dimension model using a multiple imputation method to deal with missingness on items that were not administered in fall 2020. Multiple imputation allows for the uncertainty about the missing data by creating several different plausible imputed data sets and appropriately combining results obtained from each of them.

15-dimension Measurement Invariance

We conducted the measurement invariance analyses for the 15 common dimensions of the two surveys administered in 2019 and 2020 with the following procedure. For these models, we used FIML to reduce bias in parameter estimations with missingness under an MCAR (missing completely at random) or an MAR (missing at random) mechanism.

Following previous literature (Vandenberg & Lance, 2000), we first established a baseline model (configural invariance) in which the model structures are the same across two years while all parameters are freely estimated for two different groups. Next, we constrained the intercepts of the measurement model for each race group to be equal and fit a metric-invariant model through an application of a confirmatory analysis (CFA). Finally, we constrained both intercepts and loadings of the measurement model to be equal for each group and fit a scalar-invariant model. For each step, we examined fit indexes including $\chi 2$, CFI, TLI, RMSEA, and SRMR. Considering the large sample size and sensitivity of chi-square tests, we used changes in goodness of fit (GOF) statistics including CFI and RMSEA to determine whether we achieved measurement invariance for each step. Specifically, when the change of CFA and

RMSEA is less than 0.01 and the change of SRMR is less than 0.025, we determine the measurement invariance is established (Chen et al., 2008; Cheung & Rensvold, 2002). Table 3 summarized detailed results.

18-dimension measurement invariance with multiple imputations

The mice function (Buuren & Groothuis-Oudshoorn, 2011) in Rstudio was used to conduct multiple imputations for the joint dataset. Since we have complete missingness on items that were not administered in 2020, we used a multiple imputation method to generate five imputed data sets; computational issues limited us to five data copies. We then used a normal estimation method to conduct measurement invariance analyses after imputations. The cfa.mi function from semtools (Jorgensen et al., 2019) package in R was used to test the measurement invariance for the 18 latent wellbeing dimensions model across two years with five imputed datasets. The results indicated that the metric, loading and scalar measurement invariance was achieved for the 18-dimension model with imputed datasets. Specifically, the change of CFI is 0.007 and the change of RMSEA is 0.003, which are both smaller than the criteria. Table 4 summarized detailed results.

3. Examinations of items of the Activity Engagement dimension

ACT2_1-ACT2_3 items

We changed the Activity Engagement items in the 2020 fall administrations to capture all forms of respondents' activity engagement because on-campus engagement was highly restricted due to the coronavirus pandemic. Table 1 (see Appendix) displays the spring 2019 wording and the fall 2020 wording.

We tested the effects of the two different Activity Engagement items using the following strategy:

A. Tests of partial measurement invariance using the 15 common dimensions across the spring 2019 and fall 2020 samples treating the ACT2_1, ACT2_2 and ACT2_3 items as though it was the same item in both years.
B. Comparisons of item and score distributions across the spring 2019 and fall 2020 samples using the 15 common dimensions.

To conduct these analyses, we began with the 15-dimension FIML model from Step 2. We used this approach rather than attempting multiple imputation with all 18 dimensions because imputing all data for three dimensions (Friendships, Civic - Moral, Civic - Political) in fall 2020 would be unreliable. A drawback to this approach is that we were unable to examine the effects of the wording changes within the original, 18-dimension framework.

A. Partial measurement invariance testing

Partial measurement invariance testing evaluates the impact of constraining/freeing just a few model parameters of a certain type instead of all the parameters of that type (e.g., just a few of the item-factor loadings, just a few of the item intercepts; Cheung & Rensvold,1999). The purpose of this kind of testing is to evaluate the extent to which particular items might be contributing to the overall fit of the model.

Typically, partial measurement invariance testing is conducted by starting with the most constrained model and then releasing parameter estimates until acceptable model fit is achieved. Because our scalar model (i.e., from the 2019-2020 testing) already meets acceptable fit criteria, our purpose in conducting these tests was to evaluate the magnitude of the impact on model fit if we released parameter estimates for ACT2_1, ACT2_2 and ACT2_3 items across the 2019 and 2020 administrations while leaving all other parameter estimates constrained across the 2019 and 2020 administrations.

Beginning with the scalar model, we released the intercept for ACT2_1, ACT2_2 and ACT2_3. We then used the metric model (loadings constrained to equality across groups) and released the factor loading for ACT2_1, ACT2_2 and ACT2_3. For both models, releasing the parameter estimate for ACT2_1, ACT2_2 and ACT2_3 resulted in changes to the goodness-of-fit indices that were lower than recommended cutoff values (see Measurement Invariance procedures).

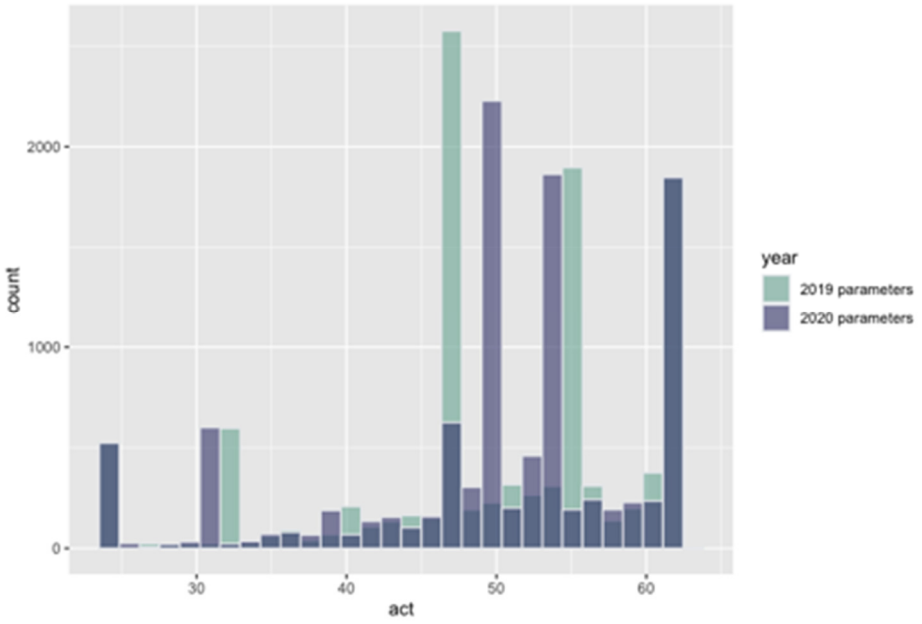
B. Comparisons of item and score distributions

Although measurement invariance procedures help to ensure similar item-performance across participant groups (i.e., spring 2019 and fall 2020 administration years), in this case they do not guarantee identical scores. It is theoretically possible for the fall 2020 data to yield different scores under either of the following two scoring scenarios available to us: (a) we score the fall 2020 data using the spring 2019 model parameter estimates, thereby treating the fall 2020 data as though it had the same measurement structure as the spring 2019 data; (b) we score the fall 2020 data using a model recalibrated to the fall 2020 data and therefore do not assume that the fall 2020 data has the same measurement structure as the spring 2019 data.

We used three methods to compare scores derived using spring 2019's parameter estimates to scores derived from a model recalibrated to the fall 2020 data: 1) a correlation matrix of both sets of scores; 2) a t-test comparing average score differences across the two sets of scores; 3) and a visual examination of the score distributions. We used the same 15-dimension FIML model used in Step 2 and in Step 3.A (immediately above). Those analyses can be summarized as finding that the two sets of items do not perform identically, but the differences between them are small. This summary should be interpreted with the methodological limitation that we did not have repeat participant samples responding to both items at the same time point; instead, we had two different samples responding to the items across a one-year time span. More robust testing strategies could come to different conclusions than ours.

1) ***Correlations.*** The correlation between the Activity Engagement dimension scores derived using spring 2019's model parameter estimates and the scores derived using parameter estimates from a model recalibrated to the fall 2020 data was 0.992. Because the scores for the Activity Engagement dimension were part of an 15-dimension model with correlated factor scores, we also evaluated the other 14 dimensions' correlations between the scores extracted from the model generated with the spring 2019 parameters and the model that was recalibrated to the fall 2020 data. The average correlation value was 0.995, with a range of 0.977 to 1.000. The detailed results were summarized in Table 5.
2) ***T-tests.*** A t-test comparing activity engagement scores derived using 2019's parameter estimates and the scores derived using estimates from a recalibrated model yielded a non-significant t-value of 0 (95% CI = -0.27 - 0.27, df = 21,796, p = 1.0), indicating that two sets of scores did not differ on average.
3) ***Score distributions.*** Figure 1 virtually displays the two score distributions. The 2019 score distribution (mint green; lightest shade) appears to sit slightly higher on the distribution, but the lowest score in the 2019 distribution is identical to the lowest score in the 2020 distribution (dark blue; darkest shade). Most of the distributions overlap (dark blue; darkest shade), which is consistent with the t-test and correlation values reported in the prior paragraphs.

Figure 1. Response distributions for 2020 scores extracted with 2019 model parameter estimates and parameter estimates recalibrated to the 2020 data.



4. Concurrent calibration and factor score calculations

We next extracted factor scores for the fall 2020 data using an 18-dimension concurrent calibration model with the joint dataset (i.e., spring 2019 and fall 2020). We chose this method to estimate the fall 2020 scores because the analyses in step 2 found that the Assessment was measurement-invariant between spring 2019 and fall 2020 and because the analyses in step 3 showed that the Activity Engagement dimension items in the fall 2020 administration could be used interchangeably with the Activity Engagement dimension items in spring 2019 administration. The prior analyses from steps 2 and 3 indicated that a joint-calibration model with FIML would be appropriate for the current study because joint calibration with FIML would reduce computational burden (relative to multiple imputation) while preserving the original 18-dimension model structure and without violating the assumptions of FIML.

In concurrent calibration, item parameters are estimated simultaneously using a combined data set. Items that were unique to each year were filled with missing data for the years in which they were

not presented; the items for Friendships, civic-moral and civic-political are specified as missing values in the concurrent calibrations.

When concurrent calibration is used, item parameters for the operational items in both the new and the old forms are estimated simultaneously in a single calibration run (Hanson, 1999) . Because the new and old forms have items in common, the resulting item parameters for all items in the concurrent calibration run are on the same scale.

After we have the estimated parameters, we calculated the unscaled factor scores for the joint dataset with the *lavPredict* function in the lavaan (Rosseel, 2012) package.

5. Observed score scaling and linking

Although factor score extraction theoretically results in factor scores that are normally distributed on a latent trait continuum ranging from -3 to +3 with a mean of 0 and standard deviation of 1, in practice those scores are on slightly different scales: the means, standard deviations, and scale continuum ranges may be slightly different from the values listed above, and those differences may vary across the scales (DiStefano & Hess, 2005). To make meaningful comparisons between scores calculated with the concurrent calibrated model and the scores generated in 2019, we conducted the following scaling and linking procedure.

First, we conducted a scaling procedure to transform all the scores (i.e., the joint dataset of spring 2019 and fall 2020) calculated using the concurrent calibration model to a scale with a mean of 50 and a standard deviation of 10. It seems at this point that we should be finished because the original spring 2019 scores also were scaled to have means of 50 and standard deviations of 10. However, when we add the spring 2019 and fall 2020 datasets together, we change the relative position of the spring 2019 participants to the overall distribution. We can use the Happiness dimension scores as an example. The Happiness dimension's original spring 2019 mean is 50 with a standard deviation of 10. From the

concurrent calibration model with the joint dataset, the spring 2019 Happiness dimension's mean and standard deviation are 52.78 and 9.87. The spring 2019 participants' *responses* did not change, but because the *overall* distribution of factor scores changed when we combined the spring 2019 and fall 2020 data, the spring 2019 scores' relative mean and standard deviation values changed.

To get the scores back into the same scale as the original spring 2019 scores (i.e., mean 50, standard deviation of 10), we used a linear linking and equating procedure (Kolen et al., 2014, p31). The first step in this procedure is to develop a formula that links (a) the spring 2019 scores calculated from the concurrent calibration to (b) the original spring 2019 scores. The linear linking and equating procedure is defined by setting standard deviation scores (z-scores) on the two forms to be equal. After some algebra, the resultant formula is a linear regression in which we solve for *y*:

$$l_{Y(x)} = y = ax + b$$

In this formula, $l_{Y(x)}$ (or y) is the converted scores we need, and $x$is the scores from the concurrent calibration model. Using the Happiness dimension as an example, $l_{Y(x)}$is the spring 2019 Happiness scores after they have been converted from the concurrent calibration model scoring (*x* scores, mean = 52.87) to the original spring 2019 scale (*y* scores, mean = 50). The slope (a) for this formula is $\frac{\sigma(Y)}{\sigma(X)}$, and the intercept (b) is $\left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\right]$. If we substitute those values into our regression equation, we get: σ

$$l_{Y(x)} = y = \frac{\sigma(Y)}{\sigma(X)}x + \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)\right]$$

$\sigma(Y)$ is the standard deviations of each dimension's original spring 2019 scores, and $\sigma(X)$ is the standard deviation of each dimension's spring 2019 scores calculated from the concurrent calibration model. $\mu(Y)$ is the means of the original spring 2019 dimension scores, and $\mu(X)$ is the means of the spring 2019 dimension scores calculated with the concurrent calibration model.

Continuing with our Happiness dimension example, Happiness dimension scores for all 2019 participants calculated from the concurrent calibration model have a mean of 52.78 ($\mu(X)$) and a standard deviation of 9.87 ($\sigma(X) = 9.87$). The original spring 2019 scores for the Happiness dimension (Y) have a mean of 50 ($\mu(Y) = 50$) and a standard deviation of 10 ($\sigma(Y) = 10$). Using the formula above, our regression formula would be:

$$l_{Y(x)} = y = 1.013x - 3.48$$

As proof that this formula works to set scores from the 2019 concurrent calibration model into the same scale as the original 2019 scores, we can use for x our example's mean for the Happiness dimension calculated from the concurrent calibration model (52.78); our formula should turn this value into the value of the original 2019 Happiness scores' mean, which was 50:

$$l_{Y(x)} = y = 1.013(52.78) - 3.48$$
= 49.99 (which is close enough to 50!)

Now that we know this formula works, we can apply the formula to each participant's fall 2020 scores calculated under the concurrent calibration model to set the fall 2020 scores to the same scale as the original spring 2019 scores. For example, let's pretend someone in fall 2020 received a happiness score of 47.38. If we apply our linking and equating formula, that person's score would be 44.52:

$$l_{Y(x)} = y = 1.013x - 3.48$$
= 1.013(47.38) – 3.48
= 44.52

We replicated this procedure across all 15 dimensions administered in fall 2020. We calculated a linking and equating formula for each dimension and applied it to each participant's fall 2020 scores calculated with the concurrent calibration model to get the final scores.

Reliability & Validity

Because we used 2019 model parameter estimates to generate the factor scores, the Spring 2019 Technical Report can be used for estimates of the Wellbeing Assessment's reliability and validity.

References

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(1), 1–67. https://doi.org/10.18637/jss.v045.i03

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, *36*(4), 462–494. https://doi.org/10.1177/0049124108314720

Cheung, G., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

DiStefano, C., & Hess, B. (2005). Using Confirmatory Factor Analysis for Construct Validation: An Empirical Review. *Journal of Psychoeducational Assessment*, *23*(3), 225–241. https://doi.org/10.1177/073428290502300303

Hanson, B. A. (1999). *Separate Versus Concurrent Estimation of IRT Item Parameters in the Common Item Equating Design*. 44.

Jamshidian, M., Jalal, S., & Jansen, C. (2014). MissMech: An R Package for Testing Homoscedasticity, Multivariate Normality, and Missing Completely at Random (MCAR). *Journal of Statistical Software*, *56*(1), 1–31. https://doi.org/10.18637/jss.v056.i06

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2019). *SemTools: Useful tools for structural equation modeling. R package version 0.5-2*. https://CRAN.R-project.org/package=semTools

Kolen, M. J., Brennan, R. L., & Kolen, M. J. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*. Springer New York. http://ebookcentral.proquest.com/lib/wfu/detail.action?docID=1636450

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.

Lumley, T. (2004). Analysis of complex survey samples. *Jounral of Statistical Software*, *9*(1), 1–19.

Lumley, T. (2019). *Survey: Analysis of complex survey samples. R version 3.35-1*. https://cran.r-project.org/web/packages/survey/index.html

National Center for Education Statistics. (2018). Table 306.10. Retrieved May 13, 2019 from Digest of education statistics 2018 website: https://nces.ed.gov/programs/digest/d18/tables/dt18_306.10.asp

National Center for Education Statistics. (2019). *Table 306.10. Total fall enrollment in degree-granting postsecondary institutions, by level of enrollment, sex, attendance status, and race/ethnicity or nonresident alien status of student: Selected years, 1976 through 2018*. Digest of Education Statistics. https://nces.ed.gov/programs/digest/d19/tables/dt19_306.10.asp

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

RStudio Team. (2016). *RStudio: Integrated development for R*. RStudio, Inc. http://www.rstudio.com/

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations of organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Tables

Table 1. Content differences for items in the Activity Engagement dimension

ACT2_1	Spring 2019	I am involved in at least one activity **at my school** that I enjoy.
	Fall 2020	I am involved in at least one activity that I enjoy.
ACT2_2	Spring 2019	I am involved in at least one activity **at my school** that has expanded my skills.
	Fall 2020	I am involved in at least one activity that has expanded my skills
ACT2_3	Spring 2019	I am involved in at least one activity **at my school** that is meaningful to me.
	Fall 2020	I am involved in at least one activity that is meaningful to me.

Table 2. Weights

Gen_race Indicator	RAW_WTS Value	TRIM_WTS Value	Count
F_Amind	2.500	2.503	18
F_Asian	0.621	0.623	571
F_Black	1.090	1.092	750
F_Hisp	1.102	1.105	1089
F_NA	1.000	1.003	70
F_Pacific	3.794	3.000	4
F_Two	0.590	0.593	388
F_White	0.680	0.683	4342
M_Amind	5.794	3.000	5
M_Asian	1.438	1.440	219
M_Black	3.023	3.000	163
M_Hisp	1.807	1.809	478
M_NA	1.000	1.003	32
M_Pacific	6.156	3.000	2
M_Two	1.070	1.072	158
M_White	1.409	1.411	1690
NA_Amind	1.000	1.003	2
NA_Asian	1.000	1.003	28
NA_Black	1.000	1.003	30
NA_Hisp	1.000	1.003	74
NA_NA	1.000	1.003	451
NA_Pacific	1.000	1.003	1
NA_Two	1.000	1.003	34
NA_White	1.000	1.003	285
O_Amind	1.000	1.003	1
O_Asian	1.000	1.003	3
O_Black	1.000	1.003	1
O_Hisp	1.000	1.003	2
O_White	1.000	1.003	9

Note: NA = not answered (missing)
Gen_race indicator abbreviations:

- The abbreviations are structured as *gender_race/ethnicity*
- Gender abbreviations
    - These abbreviations match the categories used in the GENDER item
    - F = female
    - M = male
    - O = other
    - NA = not answered (missing)
- Race/ethnicity abbreviations
    - These abbreviations match the categories in the calculated RACETHN variable
    - Amind = American Indian or Alaska Native, not Hispanic

- Asian = Asian, not Hispanic
- Black = African American or Black, not Hispanic
- Hisp = Hispanic/Latino/a of any race
- NA = not answered (missing)
- Pacific = Native Hawaiian or other Pacific Islander, not Hispanic
- White = White, not Hispanic
- Two = Two or more races, not Hispanic

Asking about race and ethnicity using this structure is somewhat controversial. We use this method because it is the closest match to the NCES data, which are the best available data on undergraduate student enrollment.

The RAW_WT_Value column includes the untrimmed weights.
The TRIM_WT_Value column includes the trimmed weights. We used these weights in all our modeling.

Table 3. Measurement invariance tests using a 15-dimension model for the 2019 and 2020 survey administration (*n* = 22,821)

	χ^2	df	CFI	ΔCFI	RMSEA	ΔRMSEA	SRMR	ΔSRMR
Model 1 (Configural)	14180.751	1840	0.964	-	0.027	-	0.034	-
Model 2 (Metric)	14320.306	1872	0.964	0	0.027	0	0.034	0
Model 3 (Scalar)	14834.574	1904	0.962	0.002	0.027	0	0.035	0.01

Table 4. Measurement invariance tests using a 18-dimension model for the 2019 and 2020 survey administration (*n* = 22,821)

	χ^2	df	CFI	ΔCFI	RMSEA	ΔRMSEA	SRMR	ΔSRMR
Model 1 (Configural)	19767.349	1858	0.961	-	0.033	-	0.034	-
Model 2 (Metric)	19783.805	1890	0.961	0	0.033	0	0.034	0
Model 3 (Scalar)	23155.387	1922	0.954	0.005	0.036	0.003	0.035	0.01

Table 5 Correlations between the scores extracted from the model generated with the spring 2019 parameters and the model that was recalibrated to the fall 2020 data

	happiness	anxiety	depression	loneliness	social anxiety	life satisfaction	self-esteem	optimism	perseverance	activity engagement	academic engagement	belonging	meaning	purpose	coping
happiness	0.998														
anxiety	-0.525	1.000													
depression	-0.540	0.604	1.000												
loneliness	-0.390	0.411	0.570	0.997											
social anxiety	-0.343	0.491	0.561	0.562	0.999										
life satisfaction	0.505	-0.365	-0.518	-0.391	-0.349	0.996									
self-esteem	0.454	-0.372	-0.545	-0.394	-0.447	0.638	0.997								
optimism	0.484	-0.392	-0.469	-0.342	-0.362	0.597	0.614	0.984							
perseverance	0.283	-0.194	-0.297	-0.188	-0.219	0.405	0.374	0.450	0.997						
activity engagement	0.283	-0.180	-0.216	-0.205	-0.177	0.332	0.268	0.284	0.225	0.992					
academic engagement	0.322	-0.243	-0.263	-0.212	-0.146	0.338	0.234	0.285	0.238	0.246	0.977				
belonging	0.369	-0.263	-0.343	-0.383	-0.288	0.452	0.385	0.397	0.274	0.337	0.346	0.999			
meaning	0.519	-0.350	-0.509	-0.383	-0.359	0.651	0.640	0.626	0.465	0.336	0.316	0.425	0.997		
purpose	0.248	-0.132	-0.256	-0.177	-0.174	0.418	0.351	0.412	0.423	0.244	0.224	0.275	0.439	0.999	
coping	0.344	-0.448	-0.392	-0.313	-0.381	0.340	0.368	0.378	0.265	0.176	0.166	0.200	0.332	0.176	0.997

Appendix

2020 fall administration Scoring Code

###Prepare data

##select items from 15 common dimensions from 2019 data

```
dat_2019<-read.csv(file="2019data.csv",header = TRUE)

dat_2020<-read.csv(file="2020data.csv",header = TRUE)
var_list2019<-c("HAPPY_1","HAPPY_2","HAPPY_3","HAPPY_5"
          ,"ANX_1","ANX_2","ANX_3","ANX_5"
          ,"DEP_1", "DEP_2","DEP_3","DEP_6", "DEP_7"
          ,"LONE_1","LONE_2","LONE_3","LONE_4","LONE_5"
          ,"SOCANX_1","SOCANX_2","SOCANX_3"
          ,"LIFESAT_1","LIFESAT_2","LIFESAT_4"
          ,"SELFEST_1","SELFEST_3","SELFEST_4"
          ,"OPT_2","OPT_3","OPT_5"
          ,"PERS_1","PERS_2","PERS_3"
          ,"ACT2_1","ACT2_2","ACT2_3"
          ,"ACAENG_1","ACAENG_2","ACAENG_3"
          ,"BELONG_1","BELONG_2","BELONG_3"
          ,"MEANING_1","MEANING_2","MEANING_3"
          ,"PURP_1","PURP_2","PURP_3",
          "COPING_1","COPING_2","COPING_3",
          "TRIM_WTS")
var_list2020_15dimen<-c("HAPPY_1","HAPPY_2","HAPPY_3","HAPPY_4"
                ,"ANX_1","ANX_2","ANX_3","ANX_4"
                ,"DEP_1", "DEP_2","DEP_3","DEP_4", "DEP_5"
                ,"LONE_1","LONE_2","LONE_3","LONE_4","LONE_5"
                ,"SOCANX_1","SOCANX_2","SOCANX_3"
                ,"LIFESAT_1","LIFESAT_2","LIFESAT_3"
                ,"SELFEST_1","SELFEST_3","SELFEST_3"
                ,"OPT_1","OPT_2","OPT_3"
                ,"PERS_1","PERS_2","PERS_3"
                ,"ACT2_1","ACT2_2","ACT2_3"
                ,"ACAENG_1","ACAENG_2","ACAENG_3"
                ,"BELONG_1","BELONG_2","BELONG_3"
                ,"MEANING_1","MEANING_2","MEANING_3"
                ,"PURP_1","PURP_2","PURP_3",
```

```
                "COPING_1","COPING_2","COPING_3",
                "TRIM_WTS")
dat_concu_2019<-dat_2019[,var_list2019]
dat_concu_2019$year<-paste("2019")
dat_concu_2020<-dat_2020[,var_list2020_15dimen]
dat_concu_2020$year<-paste("2020")
dim(dat_concu_2020)
###change dat_concu_2020 colnames the same as dat_concu_2019
colnames(dat_concu_2020) <- c("HAPPY_1","HAPPY_2","HAPPY_3","HAPPY_5"
                    ,"ANX_1","ANX_2","ANX_3","ANX_5"
                    ,"DEP_1", "DEP_2","DEP_3","DEP_6", "DEP_7"
                    ,"LONE_1","LONE_2","LONE_3","LONE_4","LONE_5"
                    ,"SOCANX_1","SOCANX_2","SOCANX_3"
                    ,"LIFESAT_1","LIFESAT_2","LIFESAT_4"
                    ,"SELFEST_1","SELFEST_3","SELFEST_4"
                    ,"OPT_2","OPT_3","OPT_5"
                    ,"PERS_1","PERS_2","PERS_3"
                    ,"ACT2_1","ACT2_2","ACT2_3"
                    ,"ACAENG_1","ACAENG_2","ACAENG_3"
                    ,"BELONG_1","BELONG_2","BELONG_3"
                    ,"MEANING_1","MEANING_2","MEANING_3"
                    ,"PURP_1","PURP_2","PURP_3",
                    "COPING_1","COPING_2","COPING_3",
                    "TRIM_WTS","year")

####Add back 10 items in friendship civi1 civic2 dimensions back to the dataset and calculate scores
using 2019 18 dimension model parameters with a 2019&2020 joint data set
dat_2020_nongrad_tem<-
dat_concu_2020%>%add_column(FRIENDS_1=NA,FRIENDS_2=NA,FRIENDS_3=NA,CIVIC_1=NA,CIVIC_2=
NA,CIVIC_3=NA,CIVIC_4=NA,CIVIC_5=NA,CIVIC_6=NA,CIVIC_7 =NA)
dat_2020_nongrad_tem$year<-paste("2020")
otheritem_2019<-c( "FRIENDS_1","FRIENDS_2","FRIENDS_3",
            "CIVIC_1","CIVIC_2","CIVIC_3","CIVIC_4",
            "CIVIC_5","CIVIC_6","CIVIC_7")
dat_otheritem_2019<-dat_2019[,otheritem_2019]
dat_2019_nongrad_tem<-cbind(dat_concu_2019,dat_otheritem_2019)
dat_2019_nongrad_tem$year<-paste("2019")
dim(dat_2019_nongrad_tem)
dim(dat_2020_nongrad_tem)

###Create a joint dataset
```

```
dat_score_joint<-rbind(dat_2019_nongrad_tem,dat_2020_nongrad_tem)
dim(dat_score_joint) #22821 63

##recode coping items for the new jointed data (2019 &2020) all items in eighteen dimensions
dat_score_joint<- dat_score_joint %>%
  mutate(coping_n1 = 6 - COPING_1) %>%
  mutate(coping_n2 = 6 - COPING_2) %>%
  mutate(coping_n3 = 6 - COPING_3)
dat_2019<- dat_2019 %>%
  mutate(coping_n1 = 6 - COPING_1) %>%
  mutate(coping_n2 = 6 - COPING_2) %>%
  mutate(coping_n3 = 6 - COPING_3)
##fit 2019 model with 2019 data only
model2019<-"
happy=~HAPPY_1+HAPPY_2+HAPPY_3+HAPPY_5
anx=~ANX_1+ANX_2+ANX_5
dep=~DEP_1+DEP_6+DEP_7
lone=~LONE_2+LONE_3+LONE_4+LONE_5
socanx=~SOCANX_1+SOCANX_2+SOCANX_3
lifesat=~LIFESAT_1+LIFESAT_2+LIFESAT_4
selfest=~SELFEST_1+SELFEST_3+SELFEST_4
opt=~OPT_2+OPT_3+OPT_5
pers=~PERS_1+PERS_2+PERS_3
act=~ACT2_1+ACT2_2+ACT2_3
acaeng=~ACAENG_1+ACAENG_2+ACAENG_3
belong=~BELONG_1+BELONG_2+BELONG_3
meaning=~MEANING_1+MEANING_2+MEANING_3
purp=~PURP_1+PURP_2+PURP_3
cop=~coping_n1+coping_n2+coping_n3
friend=~FRIENDS_1+FRIENDS_2+FRIENDS_3
civic1=~CIVIC_1+CIVIC_2+CIVIC_3+CIVIC_4
civic2=~CIVIC_5+CIVIC_6+CIVIC_7
"
fit_newweights2019<-cfa(model2019, missing='fiml', data=dat_2019,sampling.weights="TRIM_WTS",
                        estimator="MLR",std.lv=TRUE )

###calculate scores for 2019 dataset only
score2019<-lavPredict(fit_newweights2019,method="Bartlett")

###concurrent calibrations and calculate scores
```

```
fit_newweightsconcu<-cfa(model2019, missing='fiml',
data=dat_score_joint,sampling.weights="TRIM_WTS",
                    estimator="MLR",std.lv=TRUE )
score_concu<-lavPredict(fit_newweightsconcu,method="Bartlett")

###Linear equating to the original scaled 2019 scores
##First scaling two scores to a scale with a mean of 50 and a standard deviation of 10
score_concu<-as.data.frame(score_concu)
for (i in (1:18)) {
  score2019[,i]<-scale(score2019[,i],scale=TRUE)*10+50
  score_concu[,i]<-scale(score_concu[,i],scale=TRUE)*10+50
}
###Adjusted scales so that the  scores calculated from concurrent calibrations could be comparable to
the original 2019 scores
score_concu_2019scale<-score_concu_2019
for (i in (1:18)) {
  score_concu_2019scale[,i]<-scale(score_concu_2019[,i],scale=TRUE)*10+50
}
##put 2020 scores (calculated from concurrent calibration to original 2019 score scale)
score_concu_2020<-score_concu[11922:22821,]
score_concu_2020scale<-score_concu_2020
for (i in (1:18)) {
  for  (j in (1:10900)){
    score_concu_2020scale[j,i]<-(score_concu_2020[j,i]-
mean(score_concu_2019[,i],na.rm=TRUE))/sd(score_concu_2019[,i],na.rm = TRUE)*10+50
  }
}

###Change score names to "xx.FS"
colnames(score2020_2019parameters_18dimension)<-
c("HAPPY_FS","ANX_FS","DEP_FS","LONE_FS","SOCANX_FS","LIFESAT_FS","SELFEST_FS","OPT_FS","PER
S_FS",
"ACT_FS","ACAENG_FS","BELONG_FS","MEANING_FS","PURP_FS","COPING_FS","FRIEND_FS","CIVIC1_F
S","CIVIC2_FS")
colnames(score_concu_2020scale)<-
c("HAPPY_FS","ANX_FS","DEP_FS","LONE_FS","SOCANX_FS","LIFESAT_FS","SELFEST_FS","OPT_FS","PER
S_FS",
"ACT_FS","ACAENG_FS","BELONG_FS","MEANING_FS","PURP_FS","COPING_FS","FRIEND_FS","CIVIC1_F
S","CIVIC2_FS")

###combine with original 10900 cases;exclude friendship civic1 and civic2 dimensions
```

```
dat_2020_scores<-cbind(dat_2020,score_concu_2020scale[,c(1:15)])
##save the file 2020 fall scores calculations
write.csv(dat_2020_scores,"2020fall_factorscores_originaldatiswba_fall_2020_power_bi_dummy_v2no friedncivic1&2_new scaling.csv",row.names = FALSE）
```