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

Nicole W. Brocato
Xinyu Ni
Laura E. Hix

Version date: November 2, 2020

Suggested citation:

Brocato, N. W., Ni, X., & Hix, L. E. (2020). *Technical report: Wellbeing Assessment methods and psychometric properties for the spring 2020 administration*. Wake Forest University. https://www.acha.org/wp-content/uploads/Spring-2020-WFU-WBA-Technical-Report.pdf

Table of Contents

Revision History

April 5, 2021

Swapped the order of the HAPPY_3 and Scoring subsection under the Factor Scores heading. This change was made to clarify the order of the procedures used to generate the factor scores. We also added a *Scoring Procedures Overview* at the start of the Factor Scores heading to provide better orientation to the scoring procedures.

April 29, 2021

Added Table 2 with results of approximate longitudinal invariance testing between spring 2019 and spring 2020.

Document Purpose

The purpose of this document is to describe the methods used in spring 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 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 14 dimensions instead of 18; we provide the list of dimensions in the *Measure* section. Both because of this change in the number of required dimensions and because the 2020 measure's latent model parameter estimates were invariant with respect to the 2019 parameter estimates, we score the 2020 data using the 2019 parameter estimates.
4. HAPPY_3
    a. The HAPPY_3 item changed wording between the 2019 and 2020 administrations, but the change in wording did not significantly impact factor scores. In spring 2019, the item HAPPY_3 was *feeling happy*, and in 2020 the item was *feeling extremely happy*. We changed the item in 2020 to make it

consistent with prior years' administrations and thereby facilitate statistical linking across those years of the survey. Under the *Factor Scores* section, we describe our extensive 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 of wellbeing measured as latent factors: 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. The codebook explains which items from these dimensions were included in the procedures described below.

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. In spring 2020, the following 14 dimensions were required of all participants: life satisfaction, self-esteem, happiness, anxiety, depression, loneliness, social anxiety, optimism, perseverance, activity engagement, academic engagement, belonging, meaning, and purpose. The following 4 dimensions were optionally available at the request of participating schools: friendships, positive coping, civic values - moral, and civic values - political.

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 spring 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. We originally had more than 30 schools enrolled for the spring 2020 administration. Ultimately, 15 schools participated. Of those, one school participated during the mid-March transition from regular to online operations, and 3 schools participated after the transition. We chose the date of March 16, 2020 as the date by which most schools had transitioned to online operations, but the exact date varied from school to school. All 15 of the schools were 4-year institutions. They were public and private institutions varying in size from fewer than 1000 undergraduate students to more than 15,000 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 and made more dimensions optional. As we describe below in the *Data Conditioning* section, however, missing data rates this year were much higher than in 2019 and prior years.

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) and some of the demographic 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 February and May 2019, approximately[1] 62,938 students were invited to participate; 7,789 (12.38%) consented. After removing entirely blank cases and

[1] We say "approximately" because some schools self-administered using an anonymous survey link, and we are relying on their distribution estimates.

graduate students, the final sample is 6,650. Because of unplanned missing data due to attrition, the usable cases in any particular analysis may be lower than this number. Sample descriptives are included with the survey weights in Table 1.

School Characteristics

Of the 15 participating schools:

- Public/private: 8 were private, 7 were public
- Size: 8 schools had undergraduate FTE enrollments of <5,000; 7 had enrollments of >5,000
- Region:
  - Northeast: 1schools
  - Southeast: 8 schools
  - Midwest: 2 schools
  - West: 4 schools

Data Conditioning

Missing Data

For the variables used to generate the factor scores in the 14 dimensions all participants received, unplanned missing data rates range from .33% to 20.26%. The average missingness rate is 11.39%.

For the 4 optional dimensions (positive coping, friends, civic values - moral, civic values - political), a different number of students saw each set of items. A total of 4412 participants were presented with the coping items; missingness range from 28.28% to 28.45%. A total of

910 participants were presented with the coping items; missingness rates range from 34.95% to 35.05%. The two civic dimensions (civic-moral and civic-political) were presented together. A total of 175 participants were presented with the civic items; missingness for all the civic items is 50.86%.

We assumed that unplanned missingness was missing at random (MAR) based on analyses of the Spring 2019 missingness patterns.

If you are reading this report in preparation for conducting analyses with data we have provided you, we strongly recommend you evaluate rates and types of missing data for the variables in your study.

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 eliminates bias in parameter estimates due to missing data only if all variables that explain the missingness under an MAR mechanism (i.e., auxiliary variables) are present in the model. In practice, identifying and including all possible auxiliary variables is nearly impossible (Kline, 2015). We know from prior years' data that at least some portion of missingness in the items used to generate factor scores is associated with other items used to generate factor scores; these items' missingness is interrelated. However, it is unlikely that these items explain all the missingness, which is why we say only that FIML reduces bias due to missingness but does not eliminate it.

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 2017 NCES statistics (National Center for Education Statistics, 2018) for gender and race/ethnicity of undergraduate student enrollment. 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 .40 to 6.85. The upper range of the weights is somewhat extreme, and so we trimmed the upper weights to 3; the resulting weights ranged from .41 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 1 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 = -.33, median = -.61). However, some items did display greater skewness than is typically recommended, with a maximum value of 2.09 and a minimum value of -1.61.

Kurtosis were more varied, although were modest on average (mean = .14, median = -.12). The values ranged from -1.31 to 4.14.

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, only 14 were required of all participants; the remaining 4 were optional and were seen by far fewer participants than the 14 core dimensions. The codebook describes which items from each dimension were included in the latent variable model used to generate the factor scores.

Scoring Procedures Overview

Because the four optional dimensions were not seen by enough participants to generate the sample sizes needed to include them in the full latent variable model, we needed a different strategy for scoring the items than last year's strategy of modeling all 18 dimensions simultaneously and allowing them to correlate. We also needed to account for the changes to the wording of the HAPPY_3 item. This year's scoring strategy therefore involved: (a) testing for measurement invariance across the 2019 and 2020 administrations to ensure approximate longitudinal invariance, (b) evaluating the effects of the changed HAPPY_3 wording, (c) using spring 2019's CFA model parameter estimates to score the spring 2020 data. We describe this strategy in the sections that follow.

A. Measurement Invariance across 2019 and 2020 Administrations

We began by testing for measurement invariance across the two years using the 14 required dimensions. We began with this measurement invariance testing to ensure that the Assessment had the same latent measurement structure across the two administrations, could therefore be scored using comparable methods, and would generate factor scores that could be compared across the years.

Measurement invariance examines whether the same items measure the same latent construct (configural invariance) with similar levels of accuracy (metric invariance) and in the same metric (scalar invariance) across different participant groups (see Vandenberg & Lance, 2000, for a thorough treatment). Although techniques vary across the literature, the basic approach is to systematically constrain (or free) specific types of parameter estimates across the two groups, one type of parameter estimate at a time. We used a method that proceeded from least constrained to most constrained, although this testing can also be performed by proceeding from most to least constrained. Our least constrained model was a configural model in which the item-factor structure was the same across groups, but no other parameter estimates were constrained to equality across the groups. The next most constrained model was a metric model in which item-factor loadings were constrained to equality across groups. For our purposes, the "most constrained" model was one in which item intercepts and loadings were constrained to equality across groups. This type of invariance is often referred to as *scalar invariance*, and it is necessary if extracted factor scores are to be used in between-groups comparisons of average scores, a common use of the Wellbeing Assessment data. Changes in fit indices at each level of invariance were less than recommended guidelines of < .01 (Chen, Curran, Bollen, Kirby, & Pamela, 2008; Cheung & Rensvold, 2002). Full results are available in Table 2. We concluded that the 14 dimensions were equivalent across years. Fit indices for the scalar model met common cutoff guidelines (χ^2 = 18,928.84, df = 1682; CFI = .95, RMSEA = .95, SRMR = .04).

B. Evaluating Effects of the Changed HAPPY_3 Wording

In spring 2019, the item HAPPY_3 was *feeling happy*, and in 2020 the item was *feeling extremely happy*. We changed the item in 2020 to make it consistent with prior years' administrations and thereby facilitate statistical linking across those years of the survey. To determine whether we could retain the item in the 2020 scoring and have consistent item sets across survey administration years (i.e., rather than using different combinations of items for different years), we conducted a series of psychometric analyses that are detailed in the remainder of this section. Those analyses can be summarized as finding that the spring 2020 item is slightly higher on the latent trait distribution, but its effects on the factor scores are trivially small. This summary should be interpreted under the methodological limitations 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 design could come to different conclusions than ours.

HAPPY_3 testing methods

We tested the effects of the two different HAPPY_3 items using the following strategy:

1. Tests of partial measurement invariance in this year's 14 required dimensions across the 2019 and 2020 samples treating the HAPPY_3 item as though it was the same item in both years.
2. Comparisons of item and score distributions

1. 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, above) 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 HAPPY_3 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 HAPPY_3. We then used the metric model (loadings constrained to equality across groups) and released the factor loading for HAPPY_3. For both models, releasing the parameter estimate for HAPPY_3 resulted in changes to the goodness-of-fit indices that were lower than recommended cutoff values (see Measurement Invariance procedures, above).

2. Comparisons of item and score distributions

Although measurement invariance procedures help to ensure similar item-performance across participant groups (i.e., 2019 and 2020 administration years), in this particular case they do not guarantee identical scores. It is theoretically possible for the 2020 data to yield different scores if (a) we score the 2020 data using the 2019 model parameter estimates and treat 2020's version of HAPPY_3 (*feeling extremely happy*) as though it was 2019's version (*feeling happy*), than if (b) we score the 2020 using a model recalibrated to the 2020 data and therefore do not assume that the 2020 version of HAPPY_3 is equivalent to the 2019 version.

We used three methods to compare scores derived using 2019's parameter estimates to scores derived from a model recalibrated to the 2020 data: a correlation matrix of both sets of scores; a t-test comparing average score differences across the two sets of scores; and a visual examination of the score distributions.

The correlation between the happiness dimension scores derived using 2019's model parameter estimates and the scores derived using parameter estimates from a model recalibrated to the 2020 data was .9984. Because the score for the happiness dimension was part of an 18-dimension model with correlated factor scores, we also evaluated the correlations for the other 17 dimensions. The average correlation value was .9966, with a range of .9652 to .99997.

A t-test comparing happiness 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 = -.34 - .34, df = 13,294, p = 1.0), indicating that two sets of scores did not differ on average.

To visually evaluate the effects of the two different sets of model parameter estimates, we examined a plot of both models' response distributions in Figure 1, below. The 2019 score distribution (mint green; lightest shade) appears to sit slightly lower on the distribution, but the lowest score in the 2019 distribution is higher than the lowest score in the 2020 distribution (purple; medium 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.

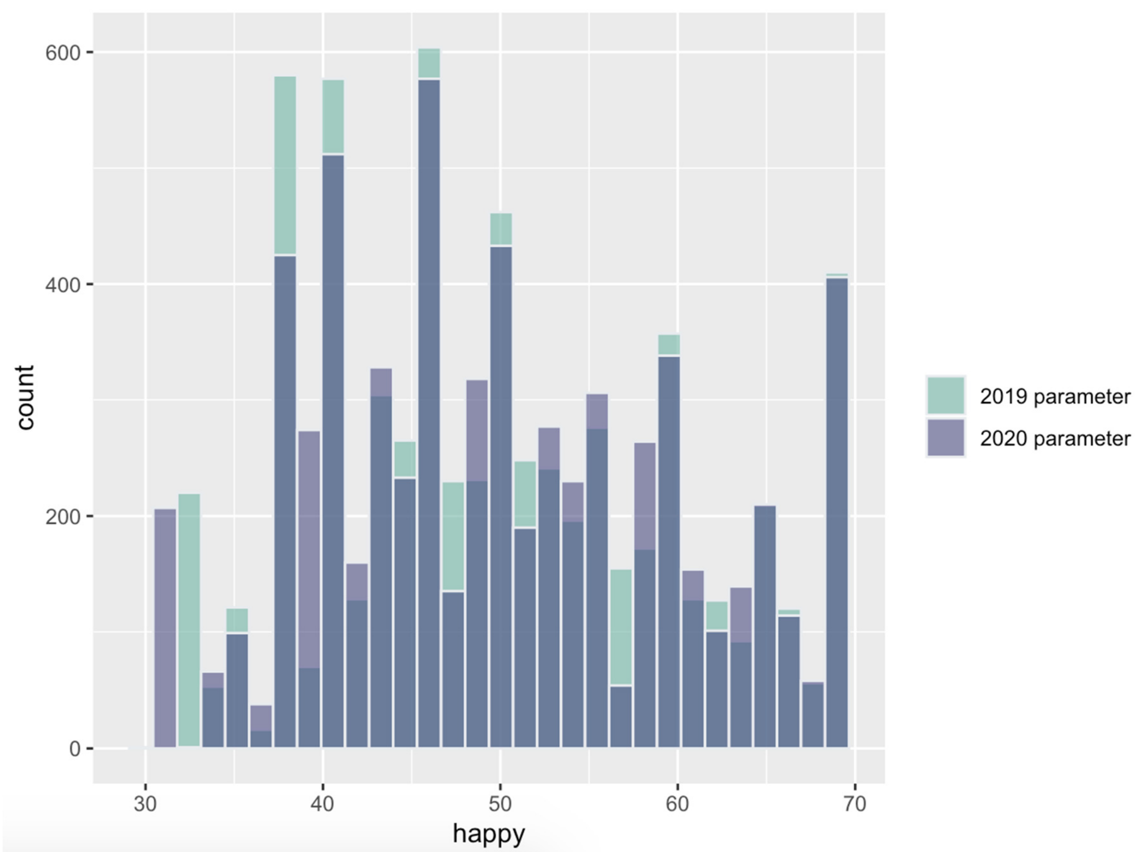

Figure 1. Response distributions for 2020 scores extracted with 2019 model parameter estimates and parameter estimates recalibrated to the 2020 data.

C. Using Spring 2019 CFA Model Parameter Estimates to Score the Spring 2020 Data

Because we concluded that measurement invariance existed across the 2019 and 2020 administration years and because we determined that the changes to the HAPPY_3 item wording did not significantly affect the Happiness dimension scoring, we chose to use 2019's model parameter estimates to score all 18 dimensions in 2020. We then extracted factor scores using the Bartlett method in *lavaan* (Rosseel, 2012). The code associated with this procedure is available in Appendix 1. The model parameter estimates are available in the Spring 2019 Technical Report.

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 be likely to vary across the scales (DiStefano, Zhu, & Mîndrilã, 2009). We scaled the scores (to mean = 0, sd = 1) to correct for those slight variations in scale. To set them in a more usable metric, we then multiplied the scores by 10 and added 50 to give them a mean of 50 and standard deviation of 10.

Reliability and 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

Bollen, K. A. (1980). Issues in the comparative measurement of political democracy. *American Sociological Review*, 45(3), 370–390. doi:10.2307/2095172

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Pamela, 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. http://doi.org/10.1177/0049124108314720

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*(1), 1-27. https://doi.org/10.1177/014920639902500101

Cheung, G. W., & 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

Craney, T. A. (2007). Model-dependent variance inflation factor cutoff values. *Quality Engineering, 14*(3), 391-403. https://doi.org/10.1081/QEN-120001878

DiStefano, S., Zhu, M., & Mîndrilã, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation, 14*(20). https://doi.org/10.7275/da8t-4g52

Enders, C. (2010). *Applied missing data analysis*. The Guilford Press.

Enders, C., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 8*(3),430-457. https://doi.org/10.1207/S15328007SEM0803_5

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50. https://www.jstor.org/stable/3151312

Graham, J., Taylor, B., Olchowski, A., & Cumsille, P. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11*, 323-343. https://doi.org/10.1037/1082-989X.11.4.323

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*(6), 1-31. http://www.jstatsoft.org/v56/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://doi.org/10.1080/01621459.1995.10476488

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

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815-852. https://doi.org/10.1177/1094428106296642

Little, T., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives, 7*(4), 199-204. https://doi.org/10.1111/cdep.12043

Lumley, T. (2019). Survey: Analysis of complex survey sample. R package version 3.35-1. https://cran.r-project.org/web/packages/survey/survey.pdf

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software, 9*(1), 1-19. https://doi.org/10.18637/jss.v009.i08

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Muthén, L. K., and B. O. Muthén. (2001). *Statistical Analysis with Latent Variables*. Los Angeles: Muthén & Muthén.

National Center for Education Statistics. (2018). Table 306.10. https://nces.ed.gov/programs/digest/d18/tables/dt18_105.20.asp

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Muthén & Muthén. https://www.statmodel.com/ugexcerpts.shtml

Raghunathan, T., & Grizzle, J. (1995). A split questionnaire design. *Journal of the American Statistical Association, 90*, 54-63. https://doi.org/ 10.1080/01621459.1995.10476488

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

Rönkkö, M. (2020). Matrixpls: Matrix-based partial least squares estimation. R package version 1.0.9. https://cran.r-project.org/web/packages/matrixpls/index.html

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

U. S. Department of Education (2007). Final guidance on maintaining, collecting, and reporting racial and ethnic data to the U.S. Department of Education, Vol. 72, No. 202 Federal Register 59266. https://www.govinfo.gov/content/pkg/FR-2007-10-19/pdf/E7-20613.pdf

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.

Tables & Appendices

Table 1. Weights

Gen_race indicator	GROUP_COUNT	RAW_WT_Value	TRIM_WT_Value
F_Amind	10	3.1495485	3
F_Asian	293	0.69717344	0.69955299
F_Black	543	0.98061498	0.98299453
F_Hisp	1000	0.80422947	0.80660903
F_NA	31	1	1.00237955
F_Pacific	11	0.98026461	0.98264416
F_Two	286	0.4033061	0.40568566
F_White	2463	0.83117707	0.83355662
M_Amind	3	6.85391488	3
M_Asian	170	1.03673518	1.03911473
M_Black	128	2.36257344	2.364953
M_Hisp	346	1.63382334	1.63620289
M_NA	28	1	1.00237955
M_Pacific	2	4.36555088	3
M_Two	81	0.89907506	0.90145462
M_White	843	1.52691621	1.52929577
NA_Hisp	2	1	1.00237955
NA_NA	269	1	1.00237955
NA_Pacific	1	1	1.00237955
NA_Two	3	1	1.00237955
NA_White	2	1	1.00237955
O_Amind	1	1	1.00237955

O_Asian	15	1	1.00237955
O_Black	6	1	1.00237955
O_Hisp	16	1	1.00237955
O_NA	5	1	1.00237955
O_Two	11	1	1.00237955
O_White	81	1	1.00237955

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 closes 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 2. Results of approximate longitudinal invariance testing between spring 2019 and spring 2020 (n = 18,661)

	χ^2	CFI	ΔCFI	RMSEA	ΔRMSEA	SRMR	ΔSRMR
Model 1 (Configural)	15643.845	0.962	-	0.038	-	0.034	-
Model 2 (Metric)	15747.167	0.962	0	0.037	-0.001	0.034	0
Model 3 (Scalar)	18928.837	0.953	0.009	0.041	0.004	0.036	0.002

Appendix 1. 2020 Scoring Code

Note: This R code requires the *lavaan* and *semTools* packages. This code applies the 2019 model parameter estimates to the 2020 data. The resultant item-factor loadings can be found in the 2019 technical report.

```
###recode coping variables
dat_2020_nongrad <- dat_2020 %>%                #dat_2020 contains the raw 2020 data
  mutate(coping_n1 = 6 - COPING_1) %>%
  mutate(coping_n2 = 6 - COPING_2) %>%
  mutate(coping_n3 = 6 - COPING_3)


#::::::::Using 2019 parameter estimates to calculate 2020 factor scores - Bartlett method
#::::::=> Based on the assumptions that all the items are not drift since 14-core dimension
model achieve measurement invariance (MI)
#1. Fit 2019 model using new weights and common items for 14 dimensions;
model2019 <-"
HAPPY_FS=~HAPPY_1+HAPPY_2+HAPPY_3+HAPPY_5
ANX_FS=~ANX_1+ANX_2+ANX_5
DEP_FS=~DEP_1+DEP_6+DEP_7
LONE_FS=~LONE_2+LONE_3+LONE_4+LONE_5
SOCANX_FS=~SOCANX_1+SOCANX_2+SOCANX_3
LIFESAT_FS=~LIFESAT_1+LIFESAT_2+LIFESAT_4
SELFEST_FS=~SELFEST_1+SELFEST_3+SELFEST_4
OPT_FS=~OPT_2+OPT_3+OPT_5
PERS_FS=~PERS_1+PERS_2+PERS_3
ACT_FS=~ACT2_1+ACT2_2+ACT2_3
ACAENG_FS=~ACAENG_1+ACAENG_2+ACAENG_3
BELONG_FS=~BELONG_1+BELONG_2+BELONG_3
MEANING_FS=~MEANING_1+MEANING_2+MEANING_3
PURP_FS=~PURP_1+PURP_2+PURP_3
COPING_FS=~coping_n1+coping_n2+coping_n3
FRIENDS_FS=~FRIENDS_1+FRIENDS_2+FRIENDS_3
CIVICMORAL_FS=~CIVIC_1+CIVIC_2+CIVIC_3+CIVIC_4
```

```
CIVICPOL_FS=~CIVIC_5+CIVIC_6+CIVIC_7
"

# Read in the 2019 data
dat_full_2019<- read.csv(**SOURCE OF 2019 DATA FILE**)

fit_newweights2019<-cfa(model2019,
missing='fiml',data=dat_full_2019,sampling.weights="TRIM_WTS_JUNE",
                estimator="MLR",std.lv=TRUE )
summary(fit_newweights2019,fit.measures=TRUE)

#2.change item name in 2020 data to what it was in 2019 data to calculate factor scores.
##select 2020 variables for analysis
var_list2020_18dimen<-c("HAPPY_1","HAPPY_2","HAPPY_3","HAPPY_4"
                ,"ANX_1","ANX_2","ANX_4"
                ,"DEP_1",  "DEP_4", "DEP_5"
                ,"LONE_2","LONE_3","LONE_4","LONE_5"
                ,"SOCANX_1","SOCANX_2","SOCANX_3"
                ,"LIFESAT_1","LIFESAT_2","LIFESAT_3"
                ,"SELFEST_1","SELFEST_2","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",
                "FRIENDS_1","FRIENDS_2","FRIENDS_3",
                "coping_n1","coping_n2","coping_n3",
                "CIVIC_1","CIVIC_2","CIVIC_3","CIVIC_4",
                "CIVIC_5","CIVIC_6","CIVIC_7",
                "TRIM_WTS")
##create a new dataset only containing analyzed items
dat_2020_temp<-dat_2020_nongrad[,var_list2020_18dimen]
##:::::Common 2020 item names to 2019 item names; For example Happy_4 to Happy_5 for
score calculating
colnames(dat_2020_temp) <- c("HAPPY_1","HAPPY_2","HAPPY_3","HAPPY_5"
```

```
                    ,"ANX_1","ANX_2","ANX_5"
                    ,"DEP_1", "DEP_6", "DEP_7"
                    ,"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",
                    "FRIENDS_1","FRIENDS_2","FRIENDS_3",
                    "coping_n1","coping_n2","coping_n3",
                    "CIVIC_1","CIVIC_2","CIVIC_3","CIVIC_4",
                    "CIVIC_5","CIVIC_6","CIVIC_7","TRIM_WTS")

######::::::Using 2019 model parameter to estimate 2020 dataset and get factor scores.
score2020_2019parameters<-
lavPredict(fit_newweights2019,newdata=dat_2020_temp,method="Bartlett")
summary(score2020_2019parameters)

###calculate 2019 model old scores,not scaled
score2019<-lavPredict(fit_newweights2019,method="Bartlett")
####create a new matrix for further use to calculate scaled 2019 scores
scale_score2019<-score2019
#####change optional module scores to NA for schools not participating the module

dat_score_temp<-cbind(dat_2020_nongrad,score2020_2019parameters)
dat_score_temp <- dat_score_temp %>%
  mutate(
    cop_sum=coping_n1+coping_n2+coping_n3,
    COPING_FS=ifelse(is.na(cop_sum),NA,COPING_FS),
    friend_sum=FRIENDS_1+FRIENDS_2+FRIENDS_3,
    FRIENDS_FS=ifelse(is.na(friend_sum),NA,FRIENDS_FS),
    civic1_sum=CIVIC_1+CIVIC_2+CIVIC_3+CIVIC_4,
```

```
    CIVICMORAL_FS=ifelse(is.na(civic1_sum),NA,CIVICMORAL_FS),
    civic2_sum=CIVIC_5+CIVIC_6+CIVIC_7,
    CIVICPOL_FS=ifelse(is.na(civic2_sum),NA,CIVICPOL_FS)
  )

dat_score_temp<-
dat_score_temp%>%select(HAPPY_FS,ANX_FS,DEP_FS,LONE_FS,SOCANX_FS,LIFESAT_FS,
                        SELFEST_FS,OPT_FS,PERS_FS,ACT_FS,ACAENG_FS,BELONG_FS,
                        MEANING_FS,PURP_FS,COPING_FS, FRIENDS_FS, CIVICMORAL_FS,
                        CIVICPOL_FS)

summary(dat_score_temp)
####:::three scores; score2019 is the scores calculating using last-year (2019) method, no
scaled first. scale_score2019 is sclaed first;
###:::score2020_2019parameters using 2019 parameters to calculate 2020 scores
for (i in (1:18)) {
  score2019[, i] <- score2019[, i]*10+50
  scale_score2019[, i] <- scale(scale_score2019[, i],scale=TRUE)*10+50
  dat_score_temp[, i] <- scale(dat_score_temp[, i],scale=TRUE)*10+50
}

##Summary and check
summary(dat_score_temp)
summary(score2019)
summary(scale_score2019)

names(dat_score_temp)
summary(dat_score_temp$FRIENDS_FS)
summary(dat_score_temp$COPING_FS)
summary(dat_score_temp$CIVICMORAL_FS)
summary(dat_score_temp$CIVICPOL_FS)
sum(!is.na(dat_2020_temp$HAPPY_1))
sum(!is.na(dat_2020_temp$CIVIC_1))
sum(!is.na(dat_2020_temp$CIVIC_4))
sum(!is.na(dat_2020_temp$coping_n1))
sum(!is.na(dat_2020_temp$FRIENDS_1))
```

```
###combine factor scores to previous data set for further use.
###:::::::::::::::::::::::::
dat_2020_nongradscore<-cbind(dat_2020_nongrad,dat_score_temp)
names(dat_2020_nongradscore)
dat_2019_scalescore<-cbind(dat_full_2019,scale_score2019)
dat_2019_rawscore<-cbind(dat_full_2019,score2019)
```