Validating Assessment Instruments for Use in Engineering Education: A Primer for Conducting and Interpreting Factor Analysis

Dr. Susan L. Amato-Henderson, Michigan Technological University

Susan Amato-Henderson is an Associate Professor Emeritus of Psychology in the Department of Cognitive and Learning Sciences at Michigan Technological University. She received her Ph.D. in Experimental Psychology from the University of North Dakota. Her research interests broadly include STEM education, and focus on individual differences in terms of motivation, self-regulated learning, self-efficacy, grit, resilience, and similar attributes as they can be leveraged to increase academic success indicators. Since retirement she serves as a research consultant.

Dr. Jon Sticklen, Michigan Technological University

Jon Sticklen is an Associate Professor with the Engineering Fundamentals Department (EF) and Affiliated Faculty with the Department of Cognitive and Learning Sciences (CLS). He served as Chair of EF from 2014-2020, leading a successful effort to design an active-learning focused upgrade of the MTU first-year engineering program. His main research interests currently are in early engineering education, particularly remote education, in systems engineering education at the undergraduate level, and in AI. His background has spanned 40 years, with active research in computer science/AI, and in large scale educational reform in CS and engineering. His work has been supported by NSF, NASA, DARPA, and on the Commerical side by McDonnell Douglas (now Boeing), GE Aircraft Engines, and The MathWorks.

Validating Assessment Instruments for Use in Engineering Education: A Primer for Conducting and Interpreting Factor Analysis

Introduction

This paper aims to be a primer for discipline-based educational research (DBER) where there may be a desire to become familiar with confirmatory and exploratory factor analysis (FA) techniques (CFA and EFA) used in multi-dimensional instrument validation. Specifically, our target audience is those in fields such as computer science and engineering who may be on the continuum of dabblers to devotees who focus their scholarly efforts on educational research but have yet to be exposed to FA. CFA is this paper's primary focus, as scales developed within the behavioral sciences, utilizing psychology students as participants, are often used in engineering education. For example, in our work (which we use here to exemplify CFA), we assess first-year engineering students' academic motivation as a predictor of academic engagement and success. The scale we use was developed utilizing students from various fields, although most of the use of this scale has been within colleges of Sciences and Arts. One can imagine that engineering student motivations may differ from non-engineering students. A CFA is warranted to ensure that the instrument we have selected is valid within our population of first-year engineering students because we are applying a scale validated on a different population and perhaps even a different domain (psychology vs engineering). In addition, the scale was validated as measuring motivation in elective courses, and our use was within required courses. Given these differences, ensuring that the original scale factors "fit" under different circumstances from when validated is essential.

Background

FA consists of a group of statistical techniques that simplify more complex variables. Psychologists first used FA to understand the factors underlying the construct of intelligence. It is often used to support theory and develop new social and behavioral science instruments to measure variables that cannot be directly observed. The end goal of FA is to establish construct validity, which is critical to developing quality assessment tools. [1] Construct validity refers to the degree to which a test or measure assesses the underlying theoretical construct it is supposed to measure. For example, construct validity addresses the issues of whether an intelligence test measures intelligence or a test of motivation measures actual motivation.

A construct is a complex variable that cannot be measured directly. Instead, its existence is inferred through the presence of statistical validity (along with other forms of validity as well). For example, take the construct of motivation. Motivation is the process that initiates, guides and maintains goal-oriented behaviors. Construct validity concerns how an instrument or measure accurately assesses what it's supposed to measure. In the case of the motivation example, establishing construct validity allows one to say that the instrument they are using indeed measures motivation.

Other forms of validity include content, criterion, and face validity. Generally, the more types of validity an instrument possesses, the more confidence one can have that it is accurately assessing the construct of interest. While FA is most strongly associated with establishing construct validity, it also plays a role in criterion and content validity. With criterion validity, the focus is on whether the results of the FA can predict some outcome now (concurrent) or in the future (predictive). Thus, statistically speaking, the values from the individual constructs and/or overall "score" on the assessment for an individual are compared to the value of some other outcome variable from the same individual. Across a group of students, these paired-score comparisons would allow a researcher to determine the size and direction of any correlation that may exist. A correlation between the instrument and the second variable suggests that the instrument can predict the second variable. An example of a research question focusing on criterion validity is whether academic motivation predicts GPA (where academic motivation FA results are compared to GPA). Content validity is a question of how representative the individual items in an instrument reflect the whole construct of interest. For example, an instrument to assess motivation may include items regarding interest, success, and usefulness (i.e., motivation would likely increase if one is interested in, experiences success with, and finds the learning helpful material). Suppose the instrument also contained items unlikely to be associated with the construct of interest, such as including items inquiring about socioeconomic status and intelligence on an instrument designed to assess motivation. In that case, we might find that those items do not have content validity. Similarly, if we believe that motivation also includes some measure of the tendency to learn more about a topic, but our instrument did not contain items to assess that component of motivation, we would similarly question the content validity as not representative of the whole motivation spectrum. Lastly, face validity does not result from statistical analysis. Face validity, generally established by asking experts' opinions, is whether an instrument contains items relevant to the measured construct.

Numerous theories exist to identify and understand motivation; a whole subfield of psychology is devoted to understanding this process that initiates, guides, and maintains goal-oriented behaviors. Theories differ based on whether they focus on content (describes

needs that drive motivation), processes (describes how a motivational state occurs), or cognitive (describes the role of one's perception and environment in motivation) components of motivation. While we do not intend to review the many theories of motivation in this paper, they all share a common feature: assessment tools have been developed to assess motivation from the many perspectives these theories espouse. When these tools were created, a series of FAs resulted in either a statistically validated tool that researchers are confident measures the more complex construct of motivation, or they did not support the tool as a measurement of motivation, thereby sending the researchers back to the drawing board.

Generally, one of two FA techniques is used to establish an instrument's construct validation. Exploratory factor analysis (EFA) is used to develop new scales or assessment tools. Confirmatory factor analysis (CFA) is used to confirm that an instrument previously validated (through EFA) within a given population, historical period, domain, etc., is valid for use in a different or new situation in which its use is intended.

The following sections present an overview of EFA as used in the initial validation of a new scale and the necessary background on the scale for motivation we chose to utilize in this paper to illustrate CFA. Finally, we provide a more detailed description of how to conduct and interpret the results of a CFA.

Factor Analysis in Instrument or Scale Development and Use

Exploratory Factor Analysis

When using EFA, the goal is to reduce a significant number of individual measurements (i.e., items or statements of agreement) into fewer factors that accurately represent the construct of interest. Hinkin has delineated a detailed model for conducting EFA [2] [3]. His model outlines 12 steps of scale development over two stages of research. We have created a model that can be considered a combination and adaptation of Hinkin's models presented in brief. We add an ideation stage and provide more detail on tasks associated with each stage in our proposed scale development model for EFA (see Figure 1).

In the ideation stage, a researcher conceptualizes a construct by reviewing theories and existing literature. A determination regarding the domain specificity of the construct is also warranted (i.e., academic motivation vs. motivation). Finally, in the ideation stage, the researchers hypothesize factors that likely contribute to their construct (e.g., sense of belonging, interest, and usefulness of material being learned might be three factors contributing to academic motivation).

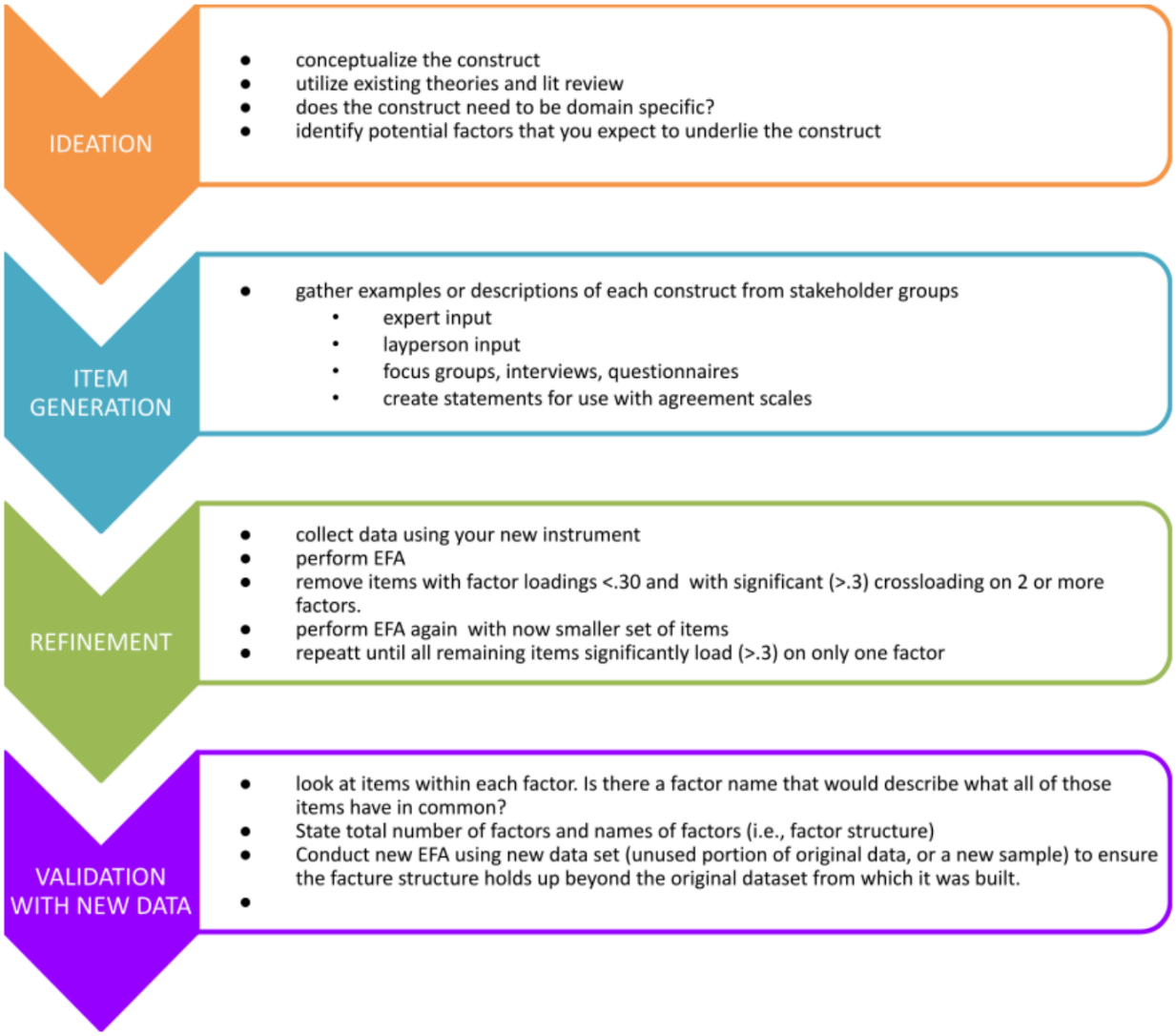

Figure 1. The Process of Dimension Reduction in EFA

During the item generation stage, a series of “questions” or items that will contribute to the factors identified in the ideation stage are created from a review of the literature, theories, expert input, and other means. For example, one scale item that the researcher might believe contributes to a "usefulness" factor might be: *The material presented in this class will be helpful in my future career* and assessed on a Likert-type agreement scale (numerical scale designed to measure agreement with a statement). Generally, at least twice as many items should be generated as you desire to have in the final instrument (i.e., for a 10-item instrument, you should generate 20 items). Another rule of thumb is to have no fewer than three items per factor in your final instrument, while five items per factor are even better [4]. So, if we proposed a 5-factor model in the ideation stage (see Figure 1), we would need a

minimum of 15 - 25 items in the final instrument. Thus, we would generate at least 30 or perhaps even 50 items to test in our initial validation EFA.

The initial validation of an instrument is typically a series of individual EFAs. Following each EFA, items not hitting a specific criterion in terms of "factor loading" are removed, as well as items that load onto multiple factors (i.e., these items don't distinguish between the factors very well). Factor loading refers to the correlation between the item and the factor (anything less than .3 is removed as the item does not contribute significantly to the factor). Once items are removed based upon small factor loadings or an item loading on multiple factors (any item that has a loading value of .3 or greater on more than one factor is removed), conducting the next EFA with the remaining items will result in new factor loading values, cross-loading values, etc. This demonstrates that EFA is a cyclical process of tool refinement through multiple individual EFAs. Once the instrument is at a point where a specific number of factors can be assessed from 3-5 items, and those items only correlate well with one of the factors (i.e., no cross-loadings) the next step is validation. Using new data (or a subset of data reserved from work in the refinement process), one conducts a CFA to ensure that the factor structure is upheld.

Figure 2 demonstrates the dimension reduction process of EFA during scale development's refinement and validation stages. Dimension reduction refers to creating a smaller group of factors from a larger pool of survey items by culling out those items that do not correlate well with one of the factors, correlate with multiple factors, etc.

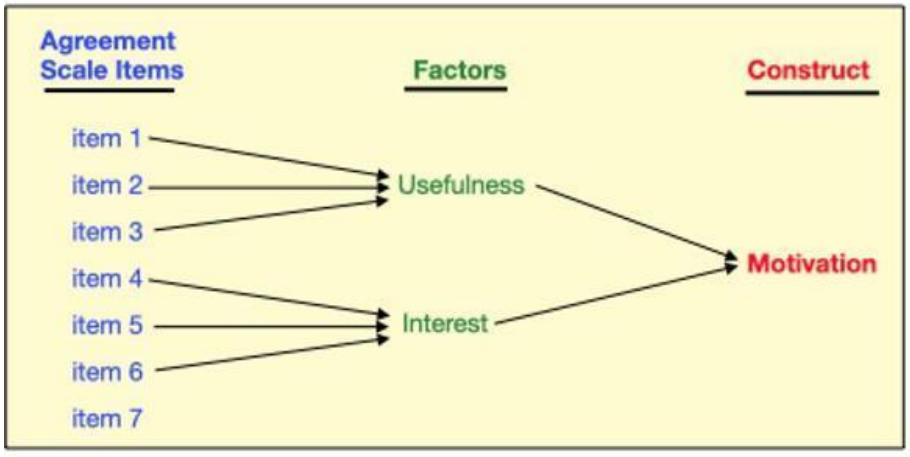

Figure 2: Relationship between Items, Factors and Constructs in hypothetical Factor Analysis

Next, we will briefly explain our motivation scale to model the CFA. We used it in previous

work, confirming the validity of the instrument to assess first-year engineering students' academic motivation. We discuss confirmatory factor analysis (CFA) and the underlying assumptions necessary to perform CFA with ordinal data, such as normality and missing data. We will describe the various model options when conducting CFA and summarize how to assess a model's fit. If model fit is lacking, we demonstrate returning to exploratory factor analysis (EFA) to assist in interpreting the model. We will use our work with the MUSIC inventory as a case study to illustrate the CFA methodology. Of note, our case study underscores the central importance of replication studies in science.

Replication is the core method used to validate original study conclusions, utilize assessment tools within new contexts, apply conclusions, or utilize cognitive models within varying contexts (e.g., external validation). Through replication, we build confidence in our scientific knowledge and the merit of our results [5][6].

Academic Motivation

The MUSIC model of motivation, developed by Jones [7], presents a framework to organize teaching strategies to motivate students to engage in learning. The accompanying MUSIC Model of Academic Motivation Inventory [8] was developed to assess student perceptions of the learning environment to aid teachers in assessing the motivational climate in their class. According to Jones and colleagues [9], the motivational climate is "the aspects of the psychological environment that affect students' motivation and engagement within a course." The MUSIC model suggests that five perceptions of the motivational climate affect students' motivations to engage in activities. MUSIC is the acronym for eMpowerment, Usefulness, Success, Interest, and Caring perceptions. According to Jones, figure 3 models the factors and demonstrates the relationship between the MUSIC factors, motivational climate, and student engagement [10]. While Jones does make the connection between motivation and engagement, his work in this area is minimal. Engagement was not his main focus. Thus, minimal attention was given to this construct in Jones' research. However, we aim to examine engagement as MUSIC and other variables predicted. Therefore, we will use a broad definition of engagement, as provided in Figure 3. In reality, researchers studying engagement have utilized a wide range of definitions, something that Wong and colleagues [11] address in their systematic review and meta-analysis of student engagement by concluding that "...while the study of student engagement is valuable due to its positive associations with desirable student outcomes, this study shows that the construct is, at the current point in time, overgeneralized".

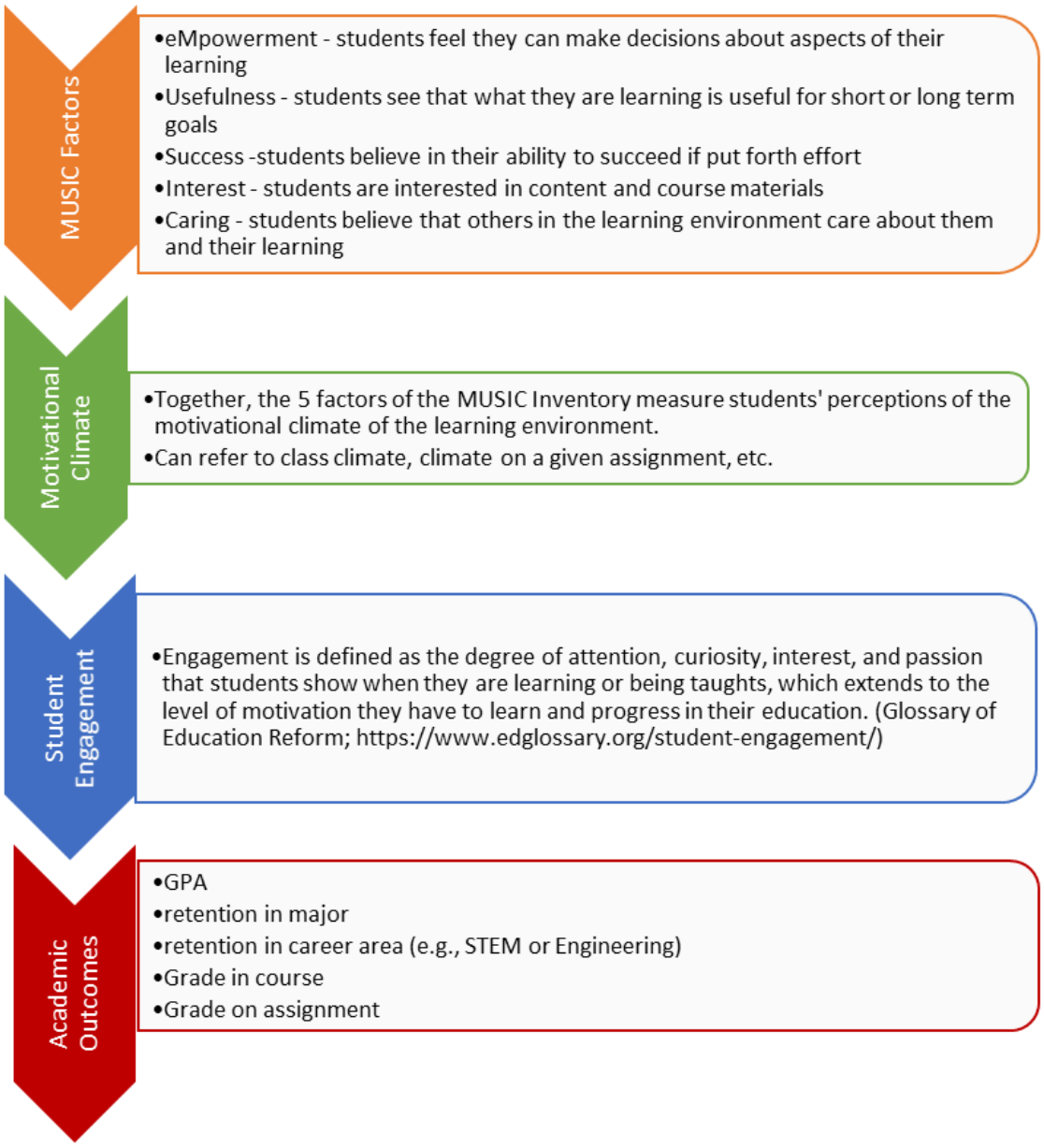

Figure 3. The Relationship between MUSIC Inventory Factors and Academic Outcomes

Confirmatory Factor Analysis

The MUSIC Inventory has been utilized across the world by numerous researchers. Of this body of work, most of those who sought to validate the initial inventory (using CFA) within their sample found that the 5-factor model of motivational climate, as identified through EFA by Jones and colleagues, is a good fit for their data. However, a handful of studies, including one with which graduate medical students were assessed, found that only 4 of the 5 scales could be validated, citing that the interest factor needed to fit the data. In our early work [12,

13], we encountered the same finding – suggesting that the interest factor overlapped with the usefulness factor in our first-year engineering students. Jones provides a nice bibliography of work conducted using the MUSIC model and inventory [14]. We will discuss this finding when interpreting our CFA below.

Unlike using EFA, in which all “items" are thrown into the black box of statistical programming language and out pops a list of items and their contribution to one or more factors, in the CFA, the researcher begins by identifying which items are intended to contribute to each of the five factors that had been found initially through the EFA. We utilize JASP open-source software [15] to demonstrate and conduct CFA analysis. The J in JASP stands for Sir Harold Jeffreys, a pioneer of Bayesian statistics; however, most users don’t know this well-hidden fact. JASP stands for Jeffreys's Amazing Statistics Program and was developed at the University of Amsterdam. JASP’s CFA is built on lavaan (lavaan.org), an R package for performing structural equation modeling (SEM).

SEM, which is a combination of FA and multiple regression, is a multivariate statistical technique used to analyze the structure of relationships between variables. CFA is the measurement part of SEM, which shows relationships between factors (latent variables in statistical analysis) and their indicators (the individual items). The other part of SEM is the structural component, or the path model, which shows how the variables of interest are related. CFA is a required first step in running most types of SEM models; performing the CFA verifies the measurement quality of factors that you will use to test experimental manipulations in an SEM.

Conducting CFA on the MUSIC Inventory

Table 1 contains the list of steps necessary to perform a CFA. While this list is presented in order of how one would select options in JASP, these options are more or less the same across statistical software. First and foremost, prepare the data to ensure that agreement scales are converted from text to numeric measurement (scale level of measurement). For example, assign the anchoring agreement labels a numeric value such as strongly disagree = 1, disagree = 2, neither agree nor disagree = 3, agree = 4, and strongly agree = 5. Data considerations for CFA go beyond this paper's scope and are a chapter's topic. Primary concerns, however, include data normality, categorical variables, and sample size adequacy. One source often utilized by beginners in FA is the work by Donna Harrington [16], aimed at providing a non-technical guide to CFA. We rely on Harrington’s work to model the CFA in this paper.

One of the first things to consider when using a data set in any statistical analysis is to examine the data for missing values. A plethora of advice exists for dealing with missing data. Harrington [16] notes that listwise and pairwise deletions are commonly used to handle missing data. Listwise deletion means that all data associated with an individual participant is ignored if there are any missing data points (i.e., one variable missing of 20 total measured). This lowers the total number of cases analyzed in a statistical test. Pairwise deletion attempts to minimize the loss of data that occurs in listwise deletion and includes only the data associated with a particular statistical test. For example, when calculating a correlation, two variables are paired. If one variable is missing, only that pair of variables is excluded from the analysis. All other variables remain included. According to Harrington, the best option is imputation, as no available data is lost. Imputation is a calculated substitute for the missing data point. The most common imputation method is to replace missing values with the mean value of the given variable.

As with most statistical procedures, the assumption of normality of data is also important with CFA. Harrington [16] addresses the normality of data by stating that multivariate normality of data is assumed, although it is difficult to measure. However, as Kline [17] points out, one can check for univariate normality and outliers, which also detect most cases of multivariate non-normality. Harrington [16] also discusses the impact of the variable scale of measurement on CFA model estimations. Focusing on Likert-type items often utilized in educational research, Harrington [16] states that it is acceptable to treat Likert-type data as continuous if there are at least five levels of agreement (for example, strongly disagree, disagree, neither agree nor disagree, agree, strongly agree), normality, and sufficient sample size. This results in data not biased by categorical variables. Finally, an adequate sample size is necessary for CFA, although there is no agreement over what "adequate" means. Factors such as data normality, missing data, the complexity of the model, and others impact sample size needs. Generally, a sample size of n=300 is likely good, with n = 200 being adequate. Anything less than n=100 is likely not adequate. Table 1 lists all the options one should select to set up the CFA.

Next, align the individual scale items (i.e., indicators) with the a priori factors they measure as per scale developers or EFA (See Figure 4).While there are two versions of the MUSIC, the original 26-item and the short form 19-item scale, we opted to use the 26-item version in this example. Each of the five factors has between 4 and 6 items in the larger version. As shown in Figure 4, the 5 factors are named, and then each of the 4-6 items associated with a factor is moved (click and drag, or click and arrow) from the overall list on the left to the appropriate factor on the right.

Table 1. Steps to CFA analysis

STEPS	Selection of Options	Insight
Prepare Data.	Ensure normality, deal with missing data, and have an adequate sample size.	Practice quality data preparation as you would with any analysis.
Align individual items to intended factors (See Figure 3).	Name the factors you expect based on assessment instructions or EFA. Typically, use a drag-and-drop method to insert the individual scale items under the associated factor.	A minimum of 3 items per factor is considered necessary. In this case, factor names were MUSIC scale names (eMpowerment, Usefulness, Success, Interest, and Caring). Be sure to utilize labels in your data file to select the correct individual items for each factor.
Select Model Parameters	If your data has a lower-level factor that will predict hierarchical or higher-level factors, select the option for second-order factors.	The MUSIC inventory only has first-order factors.
	Under the Model Options tab, the default is to have the actual factor variances serve in the model, which means that factors have a fixed variance of 1.	Select include mean structure and assume factors are uncorrelated.
	Under the Additional Output tab, select additional fit measures, which will provide information regarding how well the data fit the model. Also, select R2, which measures the proportion of variance in the indicators explained by all predictors.	The model is tested using a Chi-square test. However, the Chi-square test is very sensitive to large sample sizes, which are needed in FA. Therefore, always opt for additional fit measures.
	Under Plots, select both misfit and model plots.	These are optional but provide an alternative way of understanding the results for visual learners.
	Leave Advanced options at default unless you have a priori reason to alter the model or analysis.	

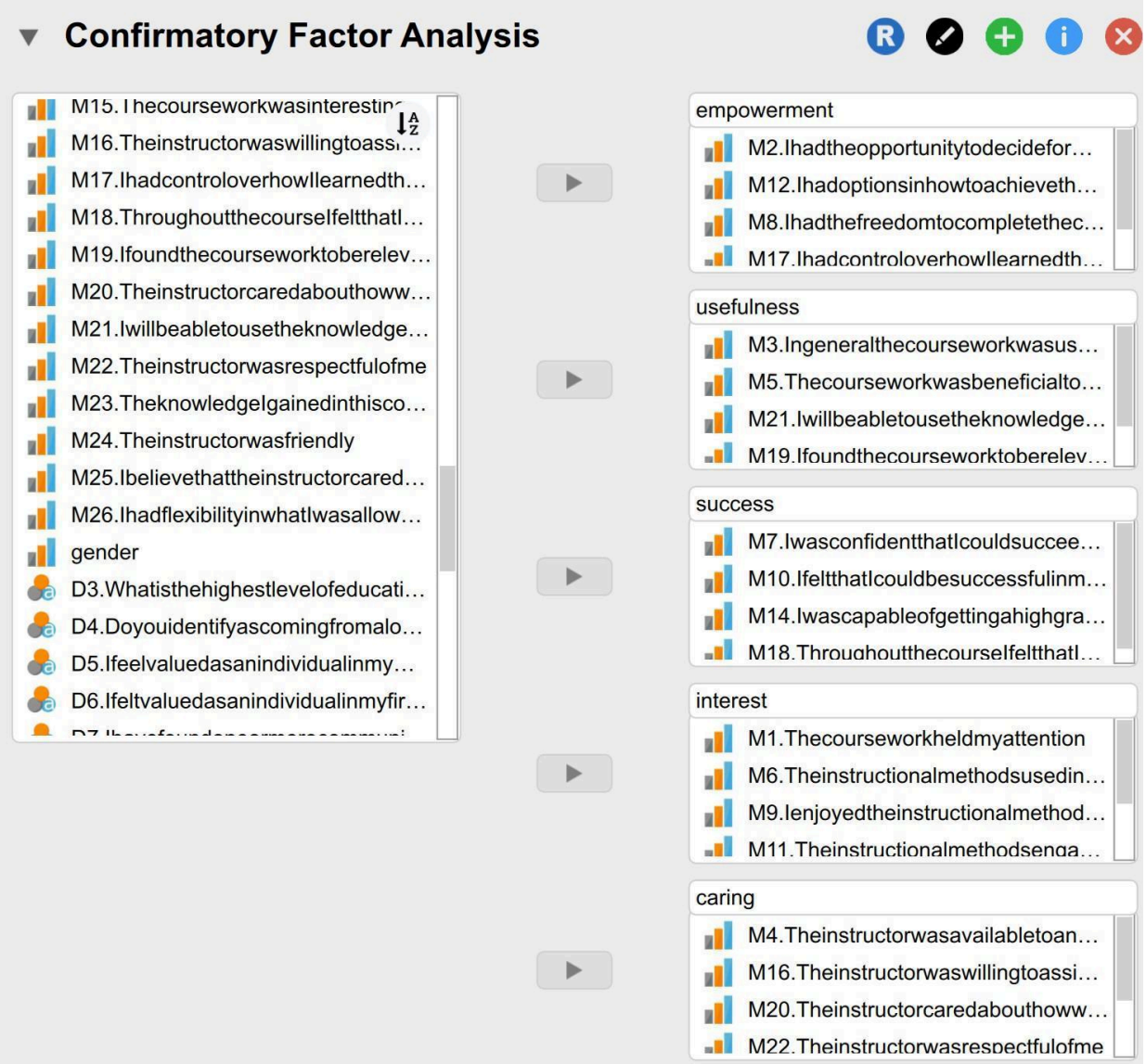

Figure 4. Align Indicators with Factors (JASP Screen Shot)

Interpretation of CFA Results

After selecting model parameters and options, we interpret the results of the CFA. FA models are tested using a global Chi-Square statistic, although the interpretation is the opposite of what one might think. A significant Chi-Square generally means that the “groups differed” or that we are to reject the null hypothesis (which states that no differences existed). As statisticians, this is typically a good finding. In CFA, however, a significant Chi-Square represents a poor model fit to the data. The statistical result from Chi-Square is almost always significant because Chi-Square is negatively affected by a large sample size such as those used in CFA. Therefore, alternative fit indices are utilized to evaluate the model's fit. Figure 5

displays the Chi-Square and alternative fit indices resulting from our MUSIC Inventory data.

Model fit

Chi-square test

Model	X^2	df	p
Baseline model	7567.139	300	
Factor model	1003.213	265	< .001

Note. The estimator is ML.

Additional fit measures

Fit indices

Index	Value
Comparative Fit Index (CFI)	0.898
Tucker-Lewis Index (TLI)	0.885
Bentler-Bonett Non-normed Fit Index (NNFI)	0.885
Bentler-Bonett Normed Fit Index (NFI)	0.867
Parsimony Normed Fit Index (PNFI)	0.766
Bollen's Relative Fit Index (RFI)	0.850
Bollen's Incremental Fit Index (IFI)	0.899
Relative Noncentrality Index (RNI)	0.898

Other fit measures

Metric	Value
Root mean square error of approximation (RMSEA)	0.081
RMSEA 90% CI lower bound	0.076
RMSEA 90% CI upper bound	0.087
RMSEA p-value	3.897×10^{-14}
Standardized root mean square residual (SRMR)	0.060
Hoelter's critical N (α = .05)	129.167
Hoelter's critical N (α = .01)	136.550
Goodness of fit index (GFI)	0.979
McDonald fit index (MFI)	0.418
Expected cross validation index (ECVI)	2.774

Figure 5. Main CFA Results for Interpretation: JASP Screen Shot

A large choice of fit indices exists to quantify the goodness of fit of the sampled data from the perfect model. Interpreting the results of these fit indices has been the subject of a large body of research. Various “rules of thumb” have been developed to interpret how well the data fit the model. While it is beyond the scope of this paper to compare the various rules of thumb, we reference findings related to goodness of fit indices as reviewed by Goretzko and colleagues [18] but present them as done by Hu and Bentler [19]. They suggest using the Comparative Fit Index (CFI), or its non-normed version, the Tucker-Lewis Index (TLI), as an alternative to the Chi-Square. The CFI is a normed fit index adjusted not to be sensitive to sample size. Values > .95 suggest a good model fit. The TLI is more appropriate for small sample sizes, and values should also be > .95. Another non-normed index is the Bentler-Bonet Non-normed Fit Index (NNFI), again useful when the sample size is smaller. As with the previous indices mentioned, a value of > .95 is considered a good model fit. The Relative Fit Index (RFI) is an alternative normed fit index with values ranging from 0 to 1. An RFI > .90 is a good model fit. Depending upon which options are selected when conducting the CFA, another commonly reported index is the RMSEA (root mean square error of approximation). Rather than measuring the goodness of fit, the RMSEA can be considered to measure the badness of fit. Therefore, an RMSEA ≤ .05 is considered a good fit. The SRMR (standardized root mean square) is similar to the standardized CFI; only it measures “bad-ness of fit.” SRMR values ≤ .08 are considered indicative of a good model fit. Generally, reporting the CFI, TLI, and RMSEA (and sometimes the SRMR) is common practice unless concerns with your data warrant reporting one (or more) of the alternative measures.

Looking at the most commonly reported indices (and we have no reason to suggest that one of the other indices would be more appropriate), our CFI = .898 and our TLI = .885. Good model fit is interpreted when the values are > .95. While our values are approaching this level, they do not support an excellent fit index. Some researchers report an “adequate fit” for values such as these. Either way, these indices suggest a better model likely exists. Our RMSEA = .081, which does not fall into the “good” or “adequate” fit categories. RMSEA can be impacted by small degrees of freedom (or sample size). In this analysis, we began with 300 df and maintained 265 df on factor analysis (a smaller number indicates missing values not imputed in this case). This is considered a good sample size for CFA. However, a look at the standardized version of this index, the SRMR, suggests that our sample size may impact the RMSEA. The value of SRMR = .06 is considered an adequate finding. In summary, 3 of our 4 indices support our model's “adequate” fit, and the fourth (RMSEA) does not indicate a good or adequate fit. In this situation, we suggest looking at the loading weights of items on each of the factors.

Factor loadings measure an individual item's predictive power on its factor. Table 6 contains the factor loadings from the JASP output (other output columns are omitted). The column titled “Estimate” lists the factor loading for each item. A general overview suggests that the items in the Success Factor had the lowest weights compared to those from the other factors. This may suggest that a problem exists with the success factor.

To further understand our findings, we look at the factor co-variances, as mentioned earlier regarding our (and other researchers') previous findings that the interest factor was not validated; the high covariances between interest and the other factors would suggest that the interest factor may be redundant in this model. In Figure 7, we see this finding again - the covariances of interest with the other factors are higher than all other covariances. Interestingly, the relatively small covariances between success and other factors suggest independence (a promising finding). A final way to examine any concerns is by looking at the model and misfit plots to see how the individual items contribute to their assigned factor and other factors. It is beyond the scope of this paper to do so. However, enough evidence from our fit indices alone suggests that a better model exists for the data.

At this point, we conducted an EFA on the data (which, unlike CFA, does not force items into their assigned factors). Recall that our overall goal was to validate the MUSIC inventory within a new group of students (engineering majors) and in a required course (vs. the original instrument’s validation from elective courses). We did not have a justified need to conduct an EFA. However, when the CFA suggests a better model exists, returning to conduct an EFA can often elucidate the answer. In doing so, we found a model in which 4 factors exist rather than 5. This model combines the Interest and Usefulness factors into 1 factor. The remaining 3 factors were upheld (Caring, Empowerment, and Success) in the EFA.

Parameter estimates			
Factor loadings			
Factor	**Indicator**	**Estimate**	**Std. Error**
Empowerment	M2.Ihadtheopportunitytodecideformyselfhowtomeetthecoursegoals	0.728	0.043
	M8.Ihadthefreedomtocompletethecourseworkmyownway	0.824	0.044
	M12.Ihadoptionsinhowtoachievethegoalsofthecourse	0.824	0.04
	M17.Ihadcontroloverhowllearnedthecoursecontent	0.76	0.042
	M26.Ihadflexibilityinwhatlwasallowedtodointhiscourse	0.777	0.043
Usefulness	M3.Ingeneralthecourseworkwasusefultome	0.908	0.043
	M5.Thecourseworkwasbeneficialtome	0.842	0.04
	M19.Ifoundthecourseworktoberelevanttomyfuture	0.893	0.043
	M21.Iwillbeabletousetheknowledgelgainedinthiscourse	0.754	0.037
	M23.Theknowledgelgainedinthiscourseisimportantformyfuture	0.849	0.04
Success	M7.IwasconfidentthatIcouldsucceedinthecoursework	0.636	0.036
	M10.IfeltthatIcouldbesuccessfulinmeetingtheacademicchallengesint	0.609	0.034
	M14.Iwascapableofgettingahighgradeinthiscourse	0.434	0.04
	M18.ThroughoutthecourselfeltthatIcouldbesuccessfulonthecoursewor	0.58	0.036
Interest	M1.Thecourseworkheldmyattention	0.852	0.049
	M6.Theinstructionalmethodsusedinthiscourseheldmyattention	1.047	0.049
	M9.Ienjoyedtheinstructionalmethodsusedinthiscourse	1.02	0.05
	M11.Theinstructionalmethodsengagedmeinthecourse	1.018	0.044
	M13.Ienjoyedcompletingthecoursework	0.877	0.052
	M15.Thecourseworkwasinterestingtome	0.729	0.045
Caring	M4.Theinstructorwasavailabletoanswermyquestionsaboutthecoursewor	0.854	0.046
	M16.Theinstructorwaswillingtoassistmeiflneededhelpinthecourse	0.779	0.039
	M20.Theinstructorcaredabouthowwellldidinthiscourse	0.806	0.041
	M22.Theinstructorwasrespectfulofme	0.646	0.034
	M24.Theinstructorwasfriendly	0.683	0.034
	M25.Ibelievethattheinstructorcaredaboutmyfeelings	0.857	0.042

Figure 6: CFA MUSIC Factor Loadings: JASP Partial Screen Shot

Factor Covariances

			Estimate	Std. Error	z-value	p	95% Confidence Interval Lower	95% Confidence Interval Upper
Empowerment	↔	Usefulness	0.606	0.036	16.774	< .001	0.535	0.677
Empowerment	↔	Success	0.540	0.043	12.597	< .001	0.456	0.624
Empowerment	↔	Interest	0.700	0.030	23.087	< .001	0.640	0.759
Empowerment	↔	Caring	0.583	0.038	15.468	< .001	0.509	0.657
Usefulness	↔	Success	0.474	0.045	10.530	< .001	0.386	0.562
Usefulness	↔	Interest	0.757	0.026	29.188	< .001	0.707	0.808
Usefulness	↔	Caring	0.561	0.038	14.875	< .001	0.487	0.635
Success	↔	Interest	0.473	0.045	10.464	< .001	0.384	0.562
Success	↔	Caring	0.464	0.046	10.185	< .001	0.375	0.554
Interest	↔	Caring	0.672	0.031	21.532	< .001	0.611	0.734

Figure 7: CFA MUSIC Factor Covariances: JASP Screen Shot

Discussion

The goal of this paper was to provide a guide for conducting CFA for DBER work when the primary researchers are using measurement instruments that were validated in another field of study. For several reasons, one would wish to validate the instrument for use in STEM or engineering. However, most researchers from these fields need to become more familiar with CFA or have limited experience conducting and interpreting the results of CFA. We utilized a tool we have previously worked with, the MUSIC Inventory (a measure of climate motivation for use in academics), to demonstrate CFA and result interpretation. The original MUSIC Model was validated across numerous studies, almost all tested on students who spanned elective courses, most of which were not engineering students, and included freshmen through seniors (and even graduate students). Given our CFA results, our interpretation is that there are better models than the current MUSIC Model of motivation, as assessed through the MUSIC Inventory, to understand the academic motivational climate of engineering students in a first-year engineering course. Three of our 4 fit indices suggest that our model is adequate rather than good. The fourth index does not support the resulting model with our data. Other results, such as the individual item loadings on each factor and factor co-variances, often assist in interpreting the model. In our situation, the item loadings on their factors show that the Success items had the lowest factor loadings as compared to all other item loadings on the other factors. Factor covariances demonstrate that the interest factor has high covariances with all other factors.

When a model has a lower than "good fit" with one's data, it suggests that the a priori hypotheses about what items and factors contribute to a construct may be incorrect. In our case, our model is not a good fit. Other than the reasons mentioned above (i.e., sample size, factor covariances, item loadings), some other common reasons for a poorly fitting model may be that some items measure multiple factors. Looking at the CFA Misfit plot, which is too large to include as a figure in this paper, one's attention is quickly drawn to one of the Interest items (*The coursework was interesting to me*) as it is significantly related to 4 Usefulness items. This confluence between the Interest and Usefulness factors has shown up in some of our previous work with the MUSIC Inventory on engineering students. It seems to be a plausible explanation for our lack of model fit. Another possible reason for a poor model fit is that some items within a factor are more related than others. In other words, if 5 items are in a factor, and 1 is highly correlated with 2 or 3 items designed to measure that factor, this could result in a poor model fit.

Once we determined that the MUSIC Model did not fit our data well, we opted to conduct an EFA. As our CFA suggested, the Interest factor needs to be revised. The EFA resulted in 4, not 5, factors. The interest items and usefulness items collapsed to measure one factor. Possible theoretical reasons for the collapse of the interest and usefulness factors will be presented below.

Conclusion

Conducting FA in discipline-based education research is a large task for many. However, demonstrating that one's assessment tool is a valid indicator of what it is intended to measure is critical to the overall validity of the work. We suggest working in cross-disciplinary teams if that allows CFA to be a routine ingredient for disciplines in which it is not now the norm. Other options may be utilizing a stats help group organized by your mathematics or statistics department. The value added to your future work when you can confirm a good model fit is critical. Imagine a scenario in which one determines that, using the MUSIC Inventory, first-year engineering students have very low motivation compared to non-engineering students. Does the research team proceed to figure out why? Or should they be certain that the assessment tool, MUSIC Inventory, is measuring what it claims to measure? In other words, confirming a fit of the model to one's data is critical in trusting the results. Validation of the instrument in new populations, new historical periods, or under different circumstances than when the model was originally validated strengthens the entire research process, especially the conclusions and future research directions.

In the work presented above, we did not find that the MUSIC model of motivation was a good fit for engineering students enrolled in a required first-year course. Of note - after our data was collected, we learned that the author had validated a short form of the instrument - a 19-item scale - partially created because of a few other reports of problems with the Interest Factor in validation studies. Thus, the abbreviated version included a narrower view of "interest" in addition to simply creating a shorter inventory. With this narrower view, we have examined validity within our student group and still maintain some questions about the Interest factor. We will next seek to adapt the tool in a manner that can better model engineering student perceptions of motivation.

From a theoretical perspective, our findings may suggest that our first-year engineering students do not know enough about their future coursework or careers to know if the content in the first-year course will be useful. Alternatively, our findings could also suggest that the personal meaning of "being interested in something" is too closely connected to the utility of said content. In this sense, engineering freshmen don't distinguish between usefulness and interest in academic topics like non-engineering students. Most engineering programs have a very narrow set of required courses that engineering students are enrolled in during their first year. An engineering student rarely has the pleasure (or pain?) of taking an elective course such as history, psychology, or art. Perhaps it is after taking some of these non-engineering courses that engineering students would develop more of a distinction between interesting and useful material.

Despite the lack of fit of the MUSIC model, we do believe that it still has value for use. By design, it allows instructors to understand how changes in the course may affect student motivation - which we believe to be a very valuable asset of this instrument. We envision creating a STEM version of the MUSIC model or perhaps a first-year engineering student model of motivation more accurately representing our students' perceptions of academic motivation. In the longer term, we also hope to strengthen our understanding of the link between student engagement and academic motivation. Anyone interested in utilizing the MUSIC instrument should conduct a CFA on their data. If the model is not a good fit, contact us for updates on future versions of the MUSIC inventory as a valid assessment of academic motivation in first-year engineering students. We would be happy to team with others, or share our future findings as we seek to re-think the MUSIC model.

References

[1] D. J. Bartholomew, "Spearman and the origin and development of factor analysis," *British Journal of Mathematical and Statistical Psychology,* vol. 48, pp. 211-220, 1995, doi https:// doi.org/10.1111/j.2044-8317.1995.tb01060.

[2] T. R. Hinkin, "A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires," *Organizational Research Methods,* vol. 1, no. 1, pp. 104-121, 1998, doi: https://doi.org/10.1177/109442819800100106.

[3] T. R. Hinkin, "Scale Development Measures.," in *Research in Organizations*, R. A. Swanson and E. F. H. III Eds. San Francisco, California: Berrett-Koehler Publishers, Inc, 2005, ch. 10.

[4] A. Costello and J. Osborne, "Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis," *Practical Assessment, Research, and Evaluation,* vol. 10, Article 7, 2019, doi: https://doi.org/10.7275/jyj1-4868.

[5] K. Popper, *The Logic of Scientific Discovery*. London, UK: Routledge, 2005.

[6] National Academies of Sciences and Medicine, "Reproducibility and replicability in science," 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK547521/.

[7] National Academies of Sciences, Medicine, "Reproducibility and replicability in science," 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK547521/.

[8] B. D. Jones, "User guide for assessing the components of the MUSIC® Model of Motivation," *Dilayari January,* vol. 18, p. 2018, 2017. [Online]. Available: http://www.theMUSIC- model.com.

[9] B. D. . Jones, Y. Miyazaki, M. Li, and S. Biscotte, "Motivational climate predicts student evaluations of teaching: Relationships between students' course perceptions, ease of course, and evaluations of teaching," *AERA Open,* vol. 8, p. 23328584211073167, 2022, doi https:// doi.org/10.1177/2332858421107316.

[10] B. D. Jones, *Motivating students by design: Practical strategies for professors*. CreateSpace Independent Publishing Platform, 2018.

[11] Wong, Z. Y., Liem, G. A. D., Chan, M., & Datu, J. A. D. (2024). Student engagement and its association with academic achievement and subjective well-being: A systematic review and meta-analysis. Journal of Educational Psychology, 116(1), 48–75. https://doi.org/10.1037/edu0000833

[12] S. L. Amato-Henderson and J. Sticklen, "Work in Progress: Utilizing the MUSIC In-instrument to Gauge Progress in First-Year Engineering Students," in *2022 IEEE Frontiers in Education Conference (FIE)*, 2022: IEEE, pp. 1-6.

[13] S. L. Amato-Henderson and J. Sticklen, "Towards the Use of the MUSIC Inventory for Measuring Engineering Student Empowerment," in *FYEE (ASEE Division, First-Year Experience)*, East Lansing, MI, 2022: ASEE, 2022.

[14] B. D. Jones. "The MUSIC Model of Motivation." https://www.themusicmodel.com (accessed Feb. 1, 2024).

[15] JASP Team. JASP. version: 0.17.2, version: 0.17.2.1. Date-released: 2023-05-11. Retrieved from https://jasp-stats.org.

[16] Harrington, Donna, 'Requirements for Conducting Confirmatory Factor Analysis: Data Considerations,' Confirmatory Factor Analysis (New York, 2008; online edn, Oxford Academic, 1 Jan. 2009), https://doi.org/10.1093/acprof:oso/9780195339888.003.0003, accessed 28 Mar. 2024.

[17] R. B.. Kline and D. A. Santor, "Principles & practice of structural equation modeling," *Canadian Psychology,* vol. 40, no. 4, p. 381, 1999.

[18] D. Goretzko, K. Siemund, and P. Sterner, "Evaluating model fit of measurement models in confirmatory factor analysis," *Educational and Psychological Measurement,* vol. 84, no. 1, pp. 123-144, 2024.

[19] L. t. Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," *Structural equation modeling: a multidisciplinary journal,* vol. 6, no. 1, pp. 1-55, 1999.