One Size Doesn't Fit All: Using Factor Analysis to Gather Validity Evidence When Using Surveys in Your Research

Eva Knekta,[†]* Christopher Runyon,[§‖] and Sarah Eddy[‡]

[†]Department of Science and Mathematics Education, Umeå University, 901 87 Umeå, Sweden; [‡]Department of Biological Sciences, Florida International University, Miami, FL 33199; [§]Department of Educational Psychology, University of Texas at Austin, Austin, TX 78712; [‖]National Board of Medical Examiners, Philadelphia, PA 19104

ABSTRACT

Across all sciences, the quality of measurements is important. Survey measurements are only appropriate for use when researchers have validity evidence within their particular context. Yet, this step is frequently skipped or is not reported in educational research. This article briefly reviews the aspects of validity that researchers should consider when using surveys. It then focuses on factor analysis, a statistical method that can be used to collect an important type of validity evidence. Factor analysis helps researchers explore or confirm the relationships between survey items and identify the total number of dimensions represented on the survey. The essential steps to conduct and interpret a factor analysis are described. This use of factor analysis is illustrated throughout by a validation of Diekman and colleagues' goal endorsement instrument for use with first-year undergraduate science, technology, engineering, and mathematics students. We provide example data, annotated code, and output for analyses in R, an open-source programming language and software environment for statistical computing. For education researchers using surveys, understanding the theoretical and statistical underpinnings of survey validity is fundamental for implementing rigorous education research.

THE USE OF SURVEYS IN BIOLOGY EDUCATION RESEARCH

Surveys and achievement tests are common tools used in biology education research to measure students' attitudes, feelings, and knowledge. In the early days of biology education research, researchers designed their own surveys (also referred to as "measurement instruments"[1]) to obtain information about students. Generally, each question on these instruments asked about something different and did not involve extensive use of measures of validity to ensure that researchers were, in fact, measuring what they intended to measure (Armbruster *et al.*, 2009; Rissing and Cogan, 2009; Eddy and Hogan, 2014). In recent years, researchers have begun adopting existing measurement instruments. This shift may be due to researchers' increased recognition of the amount of work that is necessary to create and validate survey instruments (cf. Andrews *et al.*, 2017; Wachsmuth *et al.*, 2017; Wiggins *et al.*, 2017). While this shift is a methodological advancement, as a community of researchers we still have room to grow. As biology education researchers who use surveys, we need to understand both the theoretical and statistical underpinnings of validity to appropriately employ instruments within our contexts. As a community, biology education researchers need to move beyond simply adopting a "validated" instrument to establishing the validity of the scores produced by the instrument for a researcher's

[1]In this article, we will use the terms "surveys," "measurement instrument," and "instrument" interchangeably. We will, however, put the most emphasis on the term "measurement instrument," because it conveys the importance of considering the quality of the measurement resulting from the instrument's use.

Peggy Brickman, *Monitoring Editor*

Submitted Apr 26, 2018; Revised Sep 20, 2018; Accepted Nov 27, 2018

CBE Life Sci Educ March 1, 2019 18:rm1

DOI:10.1187/cbe.18-04-0064

*Address correspondence to: Eva Knekta (eva.knekta@umu.se).

intended interpretation and use. This will allow education researchers to produce more rigorous and replicable science. In this primer, we walk the reader through important validity aspects to consider and report when using surveys in their specific context.

Measuring Variables That Are Not Directly Observable

Some variables measured in education studies are directly observable. For example, the percent of international students in a class or the amount of time students spend on a specific task can both be directly observed by the researcher. Other variables that researchers may want to measure are not directly observable, such as students' attitudes, feelings, and knowledge. The measurement of unobservable variables is what we focus on in this primer. To study these unobservable variables, researchers collect several related observable variables (responses to survey items) and use them to make inferences about the unobservable variable, termed "latent variable" or "construct"[2] in the measurement literature. For example, when assessing students' knowledge of evolution, it is intuitive that a single item (i.e., a test question) would not be sufficient to make judgments about the entirety of students' evolution knowledge. Instead, students' scores from several items measuring different aspects of evolution are combined into a sum score. The measurement of attitudes and feelings (e.g., students' goals, students' interest in biology) is no different. For example, say a researcher wanted to understand the degree to which students embrace goals focused on improving themselves, *agentic goals*, as will be seen in our illustrating example in this primer. Instead of asking students one question about how important it is for them to improve themselves, an instrument was created to include a number of items that focus on slightly different aspects of improving the self. The observed responses to these survey items can then be combined to represent the construct *agentic goal endorsement*. To combine a number of items to represent one construct, the researcher must provide evidence that these items truly represent the same construct. In this paper, we provide an overview of the evidence necessary to have confidence in using a survey instrument for one's specific purpose and go into depth for one type of statistical evidence for validity: factor analysis.

Aims

The aims of this article are 1) to briefly review the theoretical background for instrument validation and 2) to provide a step-by-step description of how to use factor analysis to gather evidence about the number and nature of constructs in an instrument. We begin with a brief theoretical background about validity and constructs to situate factor analysis in the larger context of instrument validation. Next, we discuss coefficient alpha, a statistic currently used, and often misused, in educational research as evidence for validity. The remainder of the article explores the statistical method of factor analysis. We describe what factor analysis is, when it is appropriate to use it, what we can learn from it, and the essential steps in conducting it. An evaluation of the number and nature of constructs in the Diekman *et al.* (2010) goal-endorsement instrument when used with first-year undergraduate science, technology, engineering, and mathematics (STEM) students is provided to illustrate the steps involved in conducting a factor analysis and how to report it in a paper (see Boxes 1–7). The illustrating example comes from a unique data collection and analysis made by the authors of this article. Data, annotated code, and output from the analyses run in R (an open-source programming language and software environment for statistical computing; R Core Team, 2016) for this example are included in the Supplemental Material.

WHAT IS VALIDITY?

The quality of measurements is important to all sciences. Although different terms are used in different disciplines, the underlining principles and problems are the same across disciplines. For example, in physics, the terms "accuracy" and "precision" are commonly used to describe how confident researchers should be in their measurements. In the discourse about survey quality, validity and reliability are the key concepts for measurement quality. Roughly, validity refers to whether an instrument actually measures what it is designed to measure and reliability is the consistency of the instrument's measurements.

In this section, we will briefly outline what validity is and the many types of validity evidence. Reliability, and its relation to validity, will be discussed in *The Misuse of Coefficient Alpha*. Before getting into the details, we need to emphasize a critical concept about validity that is often overlooked: validity is not a characteristic of an instrument, but rather a characteristic of *the use* of an instrument in a *particular context*. Anytime an instrument is used in a new context, at least some measures of its validity must be established for that specific context.

Validity Is Not a Property of the Instrument

The concept of validity within educational measurements has been acknowledged and discussed for a long time (e.g., Cronbach and Meehl, 1955; Messick, 1995; Cizek, 2016; Kane, 2016; Slaney, 2017). According to the latest *Standards for Educational and Psychological Testing* published by the American Educational Research Association (AERA), American Psychology Association (APA) & National Council on Measurement in Education (NCME) in 2014:

> Validity refers to the degree of which evidence and theory support the interpretations of the test score for the proposed use. (AERA, APA, and NCME, 2014, p. 11)

Thus, validity is not a property of the measurement instrument but rather refers to its proposed interpretation and use. Validity must be considered each time an instrument is used (Kane, 2016). An instrument may be validated for a certain population and purpose, but that does not mean it will work across all populations and for all purposes. For example, a validation of Diekman's goal-endorsement instrument (Diekman *et al.,* 2010) as a reasonable measure of university students' goal endorsement does not automatically validate the use of the

[2]"Latent variables" and "constructs" both refer to phenomena that are not directly observable. Examples could include a student's goals, the strength of his or her interest in biology, or his or her tolerance of failure. The term "latent variable" is commonly used when discussing these phenomena from a measurement point of view, while "construct" is a more general term used when discussing these phenomena from a theoretical perspective. In this article, we will use the term "construct" only when referring to phenomena that are not directly observable.

TABLE 1. Types of validity evidence to consider when validating an instrument according to the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 2014)

Type of validity evidence	Definition	Example considerations[a]
Evidence based on test content	Analyses of the relationship between an instrument's content and the construct it is intended to measure	Does this instrument represent the appropriate aspects of communal goals (construct) as described by the theoretical framework?
Evidence based on response processes	Information on how respondents answer the instrument's items	Is it reasonable to assume that the respondents were motivated and honest when answering the instrument? Did the respondents understand the items as intended by the researcher?
Evidence based on internal structure	Analyses of internal relationships between instrument items and instrument components and how they conform to the intended construct	Does factor analysis support the relationships between items suggested by the theoretical framework?
Evidence based on relations to other variables	Analyses of the relationships of instrument scores to variables external to the instrument and to other instruments that measure the same construct or related constructs	Can the instrument detect differences in the strength of communal goal endorsement between women and men that has been found by other instruments? Does the instrument correlate in expected ways with similar and/or dissimilar measures?
Evidence based on the consequences of testing[b]	The extent to which the consequences of the use of the score are congruent with the proposed uses of the instrument	Will the use of the instrument cause any unintended consequences for the respondent? Is the instrument identifying students who need extra resources as intended?

[a]Many of the example considerations are in reference to the elements in the Diekman *et al.* (2010) instrument; we provide these only as motivating examples and encourage readers to apply the example within their own work.

[b]If and how to include consequences of testing as a measure of validity is highly debated in educational and psychological measurement (see Mehrens, 1997; Lissitz and Samuelsen, 2007; Borsboom *et al.*, 2004; Cizek, 2016; Kane, 2016). We chose to present the view of validity as described in the latest *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 2014).

instrument for measuring 6-year-olds' goal endorsement. Similarly, a test validated for one purpose, such as being a reasonable measure of sixth-grade mathematical achievement, does not automatically validate it for use with other purposes, such as placement and advancement decisions (Kane, 2016). The validation of a survey may also be time sensitive, as cultures continually change. Using a survey from the 1980s about the use of technology would be employing a dated view of what is meant by "technology" today.

Types of Validity Evidence

Validation is a continuous and iterative process of collecting many different types of evidence to support that researchers are measuring what they aim to measure. The latest *Standards for Educational and Psychological Testing* describes many types of validity evidence to consider when validating an instrument for a particular purpose (AERA, APA, and NCME, 2014, chap. 1). These types of evidence and illustrative examples are summarized in Table 1. For example, one important aspect to consider is whether the individual items that make up the survey are interpreted by the respondents in the way intended by the researcher. Researchers must also consider whether the individual items constitute a good representation of the construct and whether the items collectively represent all the important aspects of that construct. Looking at our illustrative example (Box 1 and Table 2), we could ask whether items 15–23 (i.e., helping others, serving humanity, serving community, working with people, connection with others, attending to others, caring for others, intimacy, and spiritual rewards) in the goal-endorsement instrument constitute a good representation of the construct "helping others and one's community"? Yet another type of validity evidence involves demonstrating that the scores obtained for a construct on an instrument of interest correlate to other measures of the same or closely related constructs.

The use of existing surveys usually allows the collection of less validity evidence than the creation and use of a new survey. Specifically, if previous studies collected validity evidence for the use of the survey for a similar purpose and with a similar population as the intended research, researchers can then reference that validity evidence and present less of their own. It is important to note that, even if a survey has a long history of established use, this alone does not provide adequate validity evidence for it being an appropriate measurement instrument. It is worth researchers' time to go through the published uses of the survey and identify all the different types of validity evidence that have been collected. They can then identify the additional evidence they want to collect to feel confident applying the instrument for their intended interpretation and use. For a more detailed description of different types of validity evidence and a pedagogical description of the process of instrument validation, see Reeves and Marbach-Ad (2016) and Andrews *et al.* (2017).

In this article, we will focus on the third type of validity evidence listed in Table 1, evidence based on internal structure. Investigating the internal structure of an instrument is crucial in order to be confident that you can combine several related items to represent a specific construct. We will describe an empirical tool to gather information about the internal relationships between items in a measurement instrument: factor analysis. On its own, factor analysis is not sufficient to establish the validity of the use of an instrument in a researcher's context and

BOX 1. How to describe the purpose (abbreviated), instrument, and sample for publication illustrated with the goal-endorsement example

Defining the construct and intended use of the instrument

The aim of this study was to analyze the internal structure of the goal-endorsement instrument described by Diekman *et al.* (2010) for use with incoming first-year university STEM students. The objective is to use the knowledge gained through the survey to design STEM curricula that might leverage the goals students perceive as most important to increase student interest in their STEM classes.

The theoretical framework leading to the development of this survey has a long and well-established history. In 1966, Bakan (1966) originally proposed that two orientations could be used to characterize the human experience: agentic (orientation to the self) and communal (orientation to others). Agentic goals can thus be seen as goals focusing on improving the self or one's own circumstances. Communal goals are goals focusing on helping others and one's community and being part of a community. Gender socialization theory contributed to our understanding of who holds these goals most strongly: women are socialized to desire and assume more communal roles, while males assume more agentic roles (Eagly *et al.*, 2000; Prentice and Carranza, 2002; Su *et al.*, 2009).

This framework and survey were first used in the context of STEM education by Diekman *et al.* (2010). They found these two goal orientations to be predictive of women's attrition from STEM, particularly when they perceive STEM careers to be at odds with the communal goals important to them. Current research in this area has expanded beyond the focus on gender differences and has recognized that all humans value communal goals to some degree and that there is also variation in importance placed on communal goals by racial and ethnic groups (Smith *et al.*, 2014), social class (Stephens *et al.*, 2012), and college generation status (Allen *et al.*, 2015). The majority of this work has been done with the general population of undergraduates. Our proposed use of the survey is to explore the variation in goals among different groups in a STEM-exclusive sample.

The instrument

The goal-endorsement survey described by Diekman *et al.*, (2010) aims to measure how others-focused (communal) versus self-focused (agentic) students are. The instrument asks students to rate "how important each of the following kinds of goals [is] to you personally" on a scale of 1 (not at all important) to 7 (very important). The original measurement instrument has 23 items that have been reported as two factors: agentic (14 items) and communal (nine items) goals (see Table 2 for a listing of the items). The survey has been used many times in different contexts and has been shown to be predictive in ways hypothesized by theory. Diekman *et al.* (2010) briefly report on an EFA supporting the proposed two-factor structure of the instrument with a sample of undergraduates from introductory psychology courses.

Data collection and participants

The questionnaire was distributed in Fall 2015 and 2016 to entering first-year undergraduate students in STEM fields (biology, biochemistry, physics, chemistry, math, and computer science) at a large southern U.S. R1 university. Students took the questionnaire in

TABLE 2. Items included in the Diekman *et al.* (2010) goal-endorsement instrument[a]

	Items	Three-factor solution			Four-factor solution				Five-factor solution				
		1	2	3	1	2	3	4	1	2	3	4	5
1	Power		0.74			0.74				0.76			
2	Recognition		0.69			0.60				0.60			
3	Achievement			0.44				0.69				0.68	
4	Mastery			0.45			0.20	0.39				0.38	
5	Self-promotion		0.56			0.59	0.21			0.56	0.21		
6	Independence			0.65			0.66				0.66		
7	Individualism			0.62			0.65				0.65		
8	Status		0.79			0.75				0.74			
9	Focus on the self			0.50			0.47			0.20	0.47		
10	Success		0.23	0.39				0.65				0.65	
11	Financial rewards		0.59			0.55				0.52			
12	Self-direction			0.64			0.56				0.56		
13	Demonstrating skills or competence			0.48				0.46			0.20	0.43	
14	Competition		0.33	0.25		0.34				0.36			
15	Helping others	0.86			0.86				0.82				
16	Serving humanity	0.72			0.74				0.80				
17	Serving community	0.77			0.76				0.83				
18	Working with people	0.48			0.48								0.65
19	Connection with others	0.49			0.49								0.82
20	Attending to others	0.77			0.78				0.76				0.27
21	Caring for others	0.81			0.80				0.70				0.22
22	Intimacy	0.23	0.24		0.25	0.27							0.30
23	Spiritual rewards	0.46			0.46				0.47				

[a]Items 1–14 originally represented the agentic scale, and items 15–23 represented the communal scale. Standardized pattern coefficients from the initial EFA for the three-, four-, and five-factor solutions are reported in columns 3–14. For clarity, pattern coefficients <0.2 are not shown.

the weeks before their first Fall semester. In total, 796 students (70% women) completed the questionnaire. Fifteen percent of the students were first-generation students, and 24% came from underrepresented minorities.

Sample size
In our study, the total sample size was 796 students. Considering the number of factors (two) and the relatively large number of items per factor (nine and 14), the sample size was deemed more than sufficient to perform factor analysis (Gagne and Hancock, 2006; Wolf *et al.*, 2013).

for their purpose. However, when factor analysis is combined with other validity evidence, it can increase a researcher's confidence that they are invoking the theoretical frameworks used in the development of the instrument: that is, the researcher is correctly interpreting the results as representing the construct the instrument purports to measure.

INSTRUMENT SCOPE: ONE OR SEVERAL CONSTRUCTS?

As described in *Measuring Variables That Are Not Directly Observable*, a construct cannot be directly measured. Instead, different aspects of a construct are represented by different individual items. The foundational assumption in instrument development is that the construct is what drives respondents to answer similarly on all these items. Thus, it is reasonable to distill the responses on all these items into one single score and make inferences about the construct. Measurement instruments can be used to measure a single construct, several distinct constructs, or even make finer distinctions within a construct. The number of intended constructs or aspects of a construct to be measured are referred to as an instrument's *dimensionality*.

Unidimensional Scales

An instrument that aims to measure one underlying construct is a unidimensional scale. To interpret a set of items as if they measure the same construct, one must have both theoretical and empirical evidence that the items function as intended; that they do, indeed represent a single construct. If a researcher takes a single value (such as the mean) to represent a set of responses to a group of items that are unrelated to one another theoretically (e.g., I like biology, I enjoy doing dissection, I know how to write a biology lab report), the resulting value would be difficult to interpret at best, if not meaningless. While all of these items are related to biology, they do not represent a specific, common construct. Obviously, taking the mean response from these three items as a measure of interest in biology would be highly problematic. For example, one could be interested in biology but dislike dissection, and one's laboratory writing skills are likely influenced by aspects other than interest in biology. Even when a set of items theoretically seem to measure the same construct, the researcher must empirically show that students demonstrate a coherent response pattern over the set of items to validate their use to measure the construct. If students do not demonstrate a coherent response, this indicates that the items are not functioning as intended and they may not all measure the same construct. Thus, the single value used to represent the construct from that group of items would contain very little information about the intended construct.

Multidimensional Scales

An instrument that is constructed to measure several related constructs or several different aspects of a construct is called a multidimensional scale. For example, the Diekman *et al.* (2010) goal-endorsement instrument (see items in Box 1 and Table 2) we use in this article is a multidimensional scale: it theoretically aims to measure two different aspects of student goal endorsement. To be able to separate the results into two subscales, one must test that the items measure *distinctly* different constructs. It is important to note that whether a set of items represents different constructs can differ depending on the intended populations, which is why collecting evidence on the researcher's own population is so critical. Wigfield and Eccles (1992) illustrate this concept in a study of children of different ages. Children in early or middle elementary school did not seem to distinguish between their perceptions of interest, importance, and usefulness of mathematics, while older children did appear to differentiate between these constructs. Thus, while it is meaningful to discuss interest, importance, and usefulness as distinct constructs for older children, is it not meaningful to do so with younger children.

In summary, before using a survey, one has to have gathered all the appropriate validity evidence for the proposed interpretations and use. When measuring a construct, one important step in this validation procedure is to explicitly describe and empirically analyze the assumed dimensionality of the survey.

THE MISUSE OF COEFFICIENT ALPHA: UNDERSTANDING THE DIFFERENCE BETWEEN RELIABILITY AND VALIDITY

In many biology educational research papers, researchers only provide coefficient alpha (also called Cronbach's alpha) as evidence of validity. For example, in Eddy *et al.* (2015), the researchers describe the alpha of two scales on a survey and no other evidence of validity or dimensionality. This usage is widely agreed to be a misuse of coefficient alpha (Green and Yang, 2009). To understand why this is the case, we have to understand how validity and reliability differ and what specifically coefficient alpha measures.

Reliability is about consistency when a testing procedure is repeated (AERA, APA, and NCME, 2014). For example, assuming that students do not change their goal endorsement, do repeated measurements of students' goal endorsement using Diekman's goal-endorsement instrument give consistent results? Theoretically, reliability can be defined as the ratio between the true variance in the construct among the participating respondents (the latent, unobserved variance the researcher aims to interpret) and the observed variance as measured by the measurement instrument (Crocker and Algina, 2008). The observed variance for an item is a combination of the true variance and measurement error. Measurement error is the extent that responses are affected by factors other than the construct of interest (Fowler, 2014). For example, ideally, students' responses to Diekman's goal-endorsement instrument

would only be affected by their actual goal endorsement. But students' responses may also be affected by things unrelated to the construct of goal endorsement. For instance, responses on communal goals items may be influenced by social desirability, students' desire to answer in a way that they think others would want them to. Students' responses on items may also depend on external circumstances while they were completing the survey, such as time of the day or the noise level in their environment when they were taking the survey. While it is impossible to avoid measurement error completely, minimizing measurement error increases the ratio between the true and the observed variance, which increases the likelihood that the instrument will yield similar results over repeated use.

Unfortunately, a construct cannot, by definition, be directly measured; the true score variance is unknown. Thus, reliability itself cannot be directly measured and must be estimated. One way to estimate reliability is to distribute the instrument to the same group of students multiple times and analyze how similar the responses of the same students are over time. Often it is not desirable or practically feasible to distribute the same instrument multiple times. Coefficient alpha provides a means to estimate reliability for an instrument based on a single distribution.[3] Simply put, coefficient alpha is the correlation of an instrument to itself (Tavakol and Dennick, 2011). Calculation of coefficient alpha is based on the assumption that all items in a scale measure the same construct. If the average correlation among items on a scale is high, then the scale is said to be reliable.

The use and misuse of coefficient alpha as an estimate of reliability has been extensively discussed by researchers (e.g., Green and Yang, 2009; Sijtsma, 2009; Raykov and Marcoulides, 2017; McNeish, 2018). It is outside the scope of this article to fully explain and take a stand among these arguments. Although coefficient alpha may be a good estimator of reliability under certain circumstances, it has limitations. We will further elaborate on two limitations that are most pertinent within the context of instrument validation.

Limitation 1: Coefficient Alpha Is about Reliability, Not Validity

A high coefficient alpha does not prove that researchers are measuring what they intended to measure, only that they measured the same thing consistently. In other words, coefficient alpha is an estimation of reliability. Reliability and validity complement each other: for valid interpretations to be made using an instrument, the reliability of that instrument must be high. However, if the test is invalid, then reliability does not matter. Thus, high reliability is necessary, but not sufficient, to make valid interpretations from scores resulting from instrument administration. Consider this analogy using observable phenomena: a calibrated scale might produce consistent values for the weight of a student and thus the measure is reliable, but using this score to make interpretations about the students' height would be completely invalid. Similarly, a survey's coefficient alpha could be high, but the survey instrument could still not be measuring what the researcher intended it to measure.

[3]In addition to coefficient alpha, there are a number of other reliability estimates available. We refer interested readers to Bandalos (2018), Sijtsma (2009), and Crocker and Algina (2008).

Limitation 2: Coefficient Alpha Does Not Provide Evidence of Dimensionality of the Scale

Coefficient alpha does not provide evidence for whether the instrument measures one or several underlying constructs (Schmitt, 1996; Sijtsma, 2009; Yang and Green, 2011). Schmitt (1996) provides two illustrative examples of why a high coefficient alpha should not be taken as a proof of a unidimensional instrument. He shows that a six-item instrument, in which all items have equal correlations to one another (unidimensional instrument), could yield the same coefficient alpha as a six-item instrument with item correlations clearly showing a two-dimensional pattern (i.e., an instrument with item correlation of 0.5 across all items has the same coefficient alpha as an instrument with item correlations of 0.8 between some items and items correlations of 0.3 between other items). Thus, as Yang and Green (2011) conclude, "A scale can be unidimensional and have a low or a high coefficient alpha; a scale can be multidimensional and have a low or a high coefficient alpha" (p. 380).

In conclusion, reporting only coefficient alpha is not sufficient evidence 1) to make valid interpretations of the scores from an instrument or 2) to prove that a set of items measure only one underlying construct (unidimensionality). We encourage readers interested in learning more about reliability to read chapters 7–9 in Bandalos (2018). In the following section, we describe another statistical tool, factor analysis, which actually tests the dimensionality among a set of items.

FACTOR ANALYSIS: EVIDENCE OF DIMENSIONALITY AMONG A SET OF ITEMS

Factor analysis is a statistical technique that analyzes the relationships between a set of survey items to determine whether the participant's responses on different subsets of items relate more closely to one another than to other subsets, that is, it is an analysis of the dimensionality among the items (Raykov and Marcoulides, 2008; Leandre *et al.*, 2012; Tabachnick and Fidell, 2013; Kline, 2016; Bandalos, 2018). This technique was explicitly developed to better elucidate the dimensionality underpinning sets of achievement test items (Mulaik, 1987). Speaking in terms of constructs, factor analysis can be used to analyze whether it is likely that a certain set of items together measure a predefined construct (collecting validity evidence relating to internal structure; Table 1). Factor analysis can broadly be divided into exploratory factor analysis (EFA) and confirmatory factor analysis (CFA).

Exploratory Factor Analysis

EFA can be used to *explore* patterns underlying a data set. As such, EFA can elucidate how different items and constructs relate to one another and help develop new theories. EFA is suitable during early stages of instrument development. By using EFA, the researcher can identify items that do not empirically belong to the intended construct and that should be removed from the survey. Further, EFA can be used to explore the dimensionality of the instrument. Sometimes EFA is conflated with principal component analysis (PCA; Leandre *et al.*, 2012). PCA and EFA differ from each other in several fundamental ways. EFA is a statistical technique that should be used to identify plausible underlying constructs for a set of items. In EFA, the variance the items share is assumed to represent the construct and the nonshared variance is assumed to represent

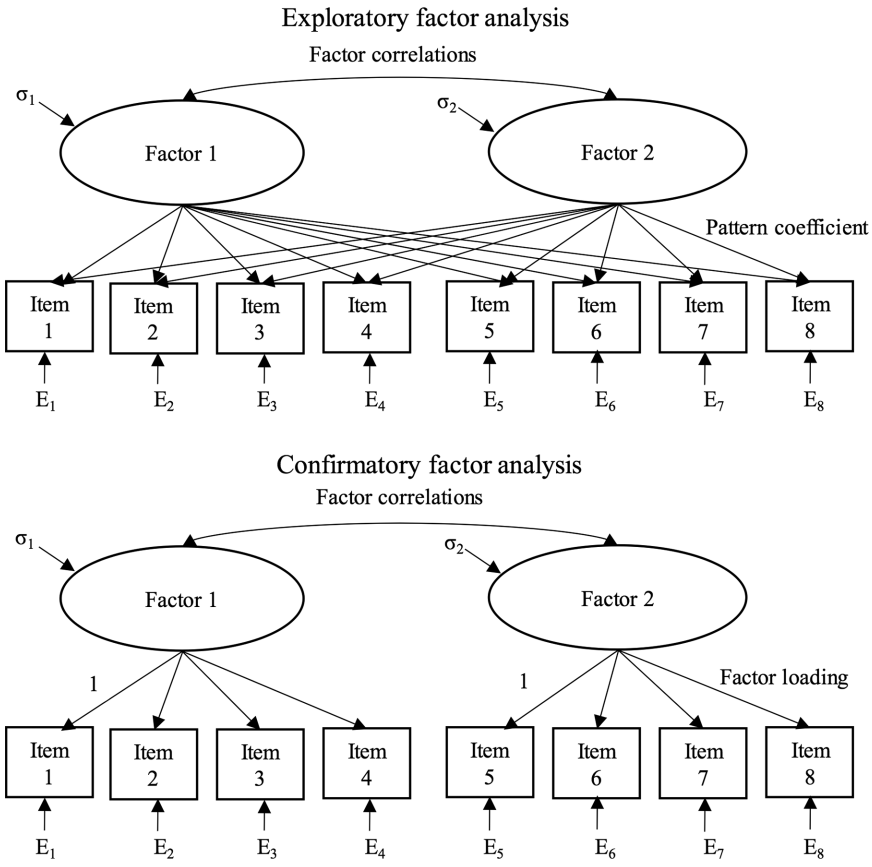

FIGURE 1. Conceptual illustration of EFA and CFA. Observed variables (items 1–8) by squares, and constructs (factors F1 and F2) are represented by ovals. Factor loading/pattern coefficients representing the effect of the factor on the item (i.e., the unique correlation between the factor and the item) are represented by arrows. σ_j, variance for factor *j*; E_i, unique error variance for item *i*. The factor loading for one item on each factor is set to 1 to give the factors an interpretable scale.

measurement errors. PCA is a data reduction technique that does not assume an underlying construct. PCA reduces a number of observed variables to a smaller number of components that account for the most variance in the observed variables. In PCA, all variance is considered, that is, it assumes no measurement errors. Within educational research, PCA may be useful when measuring multiple observable variables, for example, when creating an index from a checklist of different behaviors. For readers interested in reading more about the distinction between EFA and PCA and why EFA is the most suitable for exploring constructs, see Leandre *et al.* (2012) or Raykov and Marcoulides (2008).

Confirmatory Factor Analysis

CFA is used to *confirm* a previously stated theoretical model. Essentially, when using CFA, the researcher is testing whether the data collected supports a hypothesized model. CFA is suitable when the theoretical constructs are well understood and clearly articulated and the validity evidence on the internal structure of the scale (the relationship between the items) has already been obtained in similar contexts. The researcher can then specify the relationship between the item and the construct and use CFA to *confirm* the hypothesized number of constructs, the relationship between the constructs, and the relationship between the constructs and the items. CFA may be appropriate when a researcher is using a preexisting survey that has an established structure with a similar population of respondents.

A Brief Technical Description of Factor Analysis

Mathematically, factor analysis involves the analysis of the variances and covariances among the items. The shared variance among items is assumed to represent the construct. In factor analysis, the constructs (the shared variances) are commonly referred to as factors. Nonshared variance is considered error variance. During an EFA, the covariances among all items are analyzed together, and items sharing a substantial amount of variance are collapsed into a factor. During a CFA the shared variance among items that are *prespecified* to measure the same underlying construct is extracted. Figure 1 illustrates EFA and CFA on an instrument consisting of eight observable variables (items) aiming to measure two constructs (factors): F1 and F2. In EFA, no a priori assumption of which items represent which factors is necessary: the EFA determines these relationships. In CFA, the shared variance of items 1–4 are specified by the researcher to represent F1, and the shared variance of items 5–8 are specified to represent F2. Even further, part of what CFA tests is that items 1–4 *do not* represent F2, and items 5–8 *do not* represent F1. For both EFA and CFA, nonshared variance is considered error variance.

Figures illustrating the relationships between items and factors (such as Figure 1) are interpreted as follows. The double-headed arrow between the factors represents the correlation between the two factors (factor correlations). Each one-directional arrow between the factors and the items represents the unique correlation between the factor and the item (called "pattern coefficient" in EFA and "factor loading" in CFA). The pattern coefficients and factor loadings are similar to regression coefficients in a multiple regression. For example, consider the self-promotion item on Diekman's goal-endorsement instrument. The factor loading/pattern coefficient for this item tells the researcher how much of the average respondent's answer on this item is due to his or her general interest in agentic goals versus something unique about that item (error variance). For readers interested in more mathematical details about factor analysis, we recommend Kline (2016), Tabachnick and Fidell (2013), or Yong and Pearce (2013).

Should EFA and CFA Be Applied on the Same Sample?

If a researcher decides that EFA is the best approach for analyzing the data, the results from the EFA should ideally be confirmed

with a CFA before using the measurement instrument for research. This confirmation should never be conducted on the same sample as the initial EFA. Doing so does not provide generalizable information, as the CFA will be (essentially) repeating many of the relationships that were established through the EFA. Additionally, there could be something nuanced about the way the particular sample responds to items that might not be found in a second sample. For these reasons (among others), it is best practice to perform an EFA and CFA on independent samples. If a researcher has a large enough sample size, this can be done by randomly dividing the initial sample into two independent groups. It is also not uncommon for a researcher using an existing survey to decide that a CFA is suitable to start with but then discover that the data do not fit to the theoretical model specified. In this case, it is completely justified and recommended to conduct a second round of analyses starting with an EFA on half of the initial sample followed by a CFA on the other half of the sample (Bandalos and Finney, 2010).

FACTOR ANALYSIS STEP BY STEP

In this section, we 1) describe the important considerations when preparing to perform a factor analysis, 2) introduce the essential analytical decisions made during an analysis, and 3) discuss how to interpret the outputs from factor analyses. We illustrate each step with real data using factor analysis to analyze the dimensionality of a goal-endorsement instrument (Diekman *et al.*, 2010). Further, annotated code and output for running and analyzing EFA and CFA in R are provided as Supplemental Material (R syntax and Section 2) along with sample data.

Before delving into the technical details, we would like to be clear that conducting a factor analysis involves many decisions. There are no golden rules to follow to make these decisions. Instead, the researcher must make holistic judgments based on his or her specific context and available theoretical and empirical information. Factor analysis requires collecting evidence to build an argument to support a suggested instrument structure. The more time a researcher spends with the data investigating the effect of different possible decisions, the more confident they will be in finalizing the survey structure. As always, it is critical that a researcher's decisions are guided by previously collected evidence and empirical information and not by a priori assumptions that the researcher wishes to support.

Defining the Construct and Intended Use of the Instrument

An essential prerequisite when selecting (or developing) and analyzing an instrument is to explicitly define the intended purpose and use of the instrument. Further, the theoretical construct or constructs that one aims to measure should be clearly defined, and the current general understanding of the construct should be described. The next step is to confirm a good alignment between the construct of interest and the instrument selected to measure it, that is, that the items on the instrument actually represent what one aims to measure (evidence based on content; Table 1). For a researcher to be able to use CFA for validation, an instrument must include at least four items in total. A multidimensional scale should have at least three but preferably five or more items for each theorized subscale. In very special cases, two items can be acceptable for a subscale (Yong and Pearce, 2013; Kline, 2016).[4] For an abbreviated example of how to write up this type of validity for a manuscript using a survey instrument, see Box 1.

Sample Size

The appropriate sample size needed for factor analysis is a multifaceted question. Larger sample sizes are generally better, as they will enhance the accuracy of all estimates and increase statistical power (Gagne and Hancock, 2006). Early guidelines on sample sizes for factor analysis were general in their nature, such as a minimum sample size of 100 or 200 (e.g., see Boomsma, 1982; Gorsuch, 1983; Comrey and Lee, 1992). Although it is very tempting to adopt such general guidelines, caution must be taken, as they might lead to underestimating or overestimating the sample size needed (Worthington and Whittaker, 2006; Tabachnick and Fidell, 2013; Wolf *et al.*, 2013).

The sample size needed depends on many elements, including number of factors, number of items per factor, size of factor loadings or pattern coefficients, correlations between factors, missing values in the data, reliability of the measurements, and the expected effect size of the parameters of interest (Gagne and Hancock, 2006; Worthington and Whittaker, 2006; Wolf *et al.*, 2013). Wolf *et al.* (2013) showed that a sufficient sample size for a one-factor CFA with eight items and factor loadings of 0.8 could be as low as 30 respondents. For a two-factor CFA with three or four items per scale and factor loadings of 0.5, a sample size of ~450 respondents is needed. For EFA, Leandre *et al.* (2012) recommend that "under moderately" good conditions (communalities[5] of 0.40–0.70 and at least three items for each factor), a sample of at least 200 should be sufficient, while under poor conditions (communalities lower than 0.40 and some factors with only two items for each factor), a sample size of at least 400 is needed. Thus, when deciding on an appropriate sample size, one should consider the unique properties of the actual survey. The articles written by Wolf *et al.* (2013) and Gagne and Hancock (2006) provide a good starting point for such considerations. See Box 1 for an example of how to discuss sample size decisions in a manuscript.

In some cases, it may be implausible to have the large sample sizes necessary to obtain stable estimates from an EFA or a CFA. Often researchers must work with data that have already been collected or are using a study design that simply does not include a large number of respondents. In these circumstances, it is strongly recommended that one use a measurement instrument that has already been validated for use in a similar population for a similar purpose. In addition to considering and analyzing other relevant types of validity evidence (see Table 1), the researchers should report on validity evidence based on internal structure from other studies and describe the context of those studies relative to their own context. The researchers should also acknowledge in the methods and limitation sections that they could not run dimensionality checks on their sample. Further, researchers can also analyze a correlation matrix[6] of the responses to the survey items from their own

[4]This is partly due to identification issues (see *Specifying the Model*).

[5]In EFA, communalities describe how much of the variance in an item is explained by the factor. For more information about communalities, see *Interpreting Output from EFA*.

[6]For a description of a correlation matrix, see the Supplemental Material, Sections 1 and 2.

data collection to get a sense of how the items may relate to one another in their context. This correlation matrix may be reported to help provide preliminary validity evidence based on internal structure.

Properties of the Data

As with any statistical analysis, before performing a factor analysis the researcher must investigate whether the data meet the assumptions for the proposed analysis. Section 1 of the Supplemental Material provides a summary of what a researcher should check for in the data for the purposes of meeting the assumptions of a factor analysis and an illustration applied to the example data. These include analyses of missing values, outliers, factorability, normality, linearity, and multicollinearity. Box 3 provides an example of how to report these analyses in a manuscript.

Analytic Considerations for CFA

Once the data are screened to determine their properties, several analytical decisions must be made. Because there are some differences in analytical decisions and outputs for EFA and CFA, we will discuss EFA and CFA in separate sections. We will start with CFA, as most researchers adopting an existing instrument will use this method first and may not ever need to perform an EFA. See Box 2 for how to report analytical considerations for a CFA in a manuscript.

Selecting an Estimator. When performing a CFA, a researcher must choose a statistical method for extracting the variance from the data. There are several different methods available, including unweighted least squares, generalized least squares, maximum likelihood, robust maximum likelihood, principal axis factoring, alpha factoring, and image factoring. Each of these methods has its strengths and weaknesses. Kline (2016) and Tabachnick and Fidell (2013) provide a useful discussion of several of these methods and when best to apply each one. In general, because data from surveys are often on an ordinal level (e.g., data from Likert scales) and sometimes slightly nonnormally distributed, estimators robust against nonnormality, such as maximum-likelihood estimation with robust standard errors (MLR) or weighted least-squares estimation (WLS), are often suitable for performing CFA. Whether or not MLR or WLS is most suitable depends partly on the number of response options for the survey items. MLR work best when data can be considered continuous. In most cases, scales with seven response options work well for this purpose, whereas scales with five response options are questionably continuous. MLR is still often used in estimation for five response options, but with four or fewer response options, WLS is better (Finney and DiStefano, 2006). The decision regarding the number of response options to include in a survey should not be driven by these considerations. Rather, the number of response options and properties of the data should drive the selection of the CFA estimator. Although more response options for an item allow researchers to model it as continuous, respondents may not be able to meaningfully differentiate between the different response options. Fewer response options usually offer less ambiguity, but usually result in less variation in the response. For example, if students are provided with 10 options to indicate their level agreement with a given item, it is possible that not all of the response options may be used. In such a case, fewer response options may better capture the latent distribution of possible responses to an item.

Specifying the Model. The purpose of a CFA is to test whether the data collected with an instrument support the hypothesized model. Using theory and previous validations of the instrument, the researcher specifies how the different items and factors relate to one another (see Figure 1 for an example model). For a CFA, the number of parameters that the researcher aims to

BOX 2. What to report in the methods of a publication for a CFA using the goal-endorsement example

We chose to start with a CFA to confirm a two-factor solution, because 1) the theoretical framework underlying the instrument is well understood and articulated and 2) Diekman *et al.* (2010) performed an EFA on a similar population to ours that supported the two-factor solution. If the assumed factor model was confirmed, then we could confidently combine the items into two sum scores and interpret the data as representing both an agentic and a communal factor. CFA was run using the R package lavaan (Rosseel, 2012).

Selecting an estimator

In consideration of the ordinal and nonnormal nature of the data, the robust maximum-likelihood estimation (MLR) was used to extract the variances from the data. Full-information maximum likelihood in the estimation procedure was used to handle the missing data.

Specifying a two-factor CFA

To confirm the factor structure proposed by Diekman *et al.* (2010), we specified a two-factor CFA, with items 1–14 representing the agentic scale and items 15–23 representing the communal factor (Table 2). Correlation between the two factors was allowed. For identification purposes, the factor loading for one item on each factor was set to 1. The number of variances and covariances in the data was 276 (23(23 + 1)/2), which was larger than the number of parameter estimates (one factor correlation, 23 error terms, 21 factor loadings, and variances for each factor). Thus, the model was overidentified.

Selecting model fit indices and setting cutoff values

Multiple fit indices (chi-square value from robust MLR [MLR χ^2]; comparative fit index [CFI]; the root-mean-square error of approximation [RMSEA]; and the standardized root-mean-square residual [SRMR]) were consulted to evaluate model fit. The fit indices were chosen to represent an absolute, a parsimony-adjusted, and an incremental fit index. Consistent with the recommendations by Hu and Bentler (1999), the following criteria were used to evaluate the adequacy of the models: CFI > 0.95, SRMR < 0.08, and RMSEA < 0.06. Coefficient alpha was computed based on the model results and used to assess reliability. Values > 0.70 were considered acceptable.

estimate (e.g., error terms, variances, correlations and factor loadings) must be *less than or equal to* the number of possible variances and covariances among the items (Kline, 2016). For a CFA, a simple equation tells you the number of possible variances and covariances: $p(p + 1)/2$, where p = number of items. If the number of parameters to estimate is *more than* the number of possible variances and covariances among the items, the CFA is called "underidentified" and will not provide interpretable results. When the number of parameters to be estimated *equals* the number of covariances and variances among the items, the model is deemed "just identified" and will result in perfect fit of the data to the model, regardless of the true relationship between the items. To test whether the data fit the theoretical model, the number of parameters that are being estimated needs to be *less* than the number of variances and covariances observed in the data. In this case, the model is "overidentified." For the example CFA in Figure 1, the number of possible variances and covariances is $8(8 + 1)/2 = 36$, and the number of parameters to estimate is 17 (one factor correlation, eight error terms, six factor loadings, and variances for each of the two factors[7]), thus the model is overidentified.

Choosing Appropriate Model Fit Indices. The true splendor of CFA is that so-called model fit indices have been developed to help researchers understand whether the data support the hypothesized theoretical model.[8] The closest statistic to an omnibus test of model fit is the model chi-square test. The null hypothesis for the chi-square test is that there is no difference between the hypothesized model and the observed relationships within the data. Several researchers argue that this is an unrealistic hypothesis (Hu and Bentler, 1999; Tabachnick and Fidell, 2013). A close approximation of the data to the model is more realistic than a perfect model fit. Further, the model chi-square test is very sensitive to sample size (the chi-square statistic tends to increase with an increase in sample size, all other considerations constant; Kline, 2016). Thus, while large sample sizes provide good statistical power, the null hypothesis that the factor model and the data do not differ from each other may be rejected although the difference is actually quite small. Given these concerns, it is important to consider the result of the chi-square test in conjunction with multiple other model fit indices.

Many model fit indices have been developed that quantify the *degree* of fit between the model and the data. That is, the values provided by these indices are not intended to make binary (fit vs. no fit) judgments about model fit. These model fit indices can be divided into absolute, parsimony-adjusted, and incremental fit indices (Bandalos and Finney, 2010). Because each type of index has its strengths and weaknesses (e.g., sensitivity to sample size, model complexity, or misspecified factor correlations), using at least two different types of fit indices is recommended (Hu and Bentler, 1999; Tabachnick and Fidell, 2013). The researcher should decide a priori which model fit indices to use and the cutoff values that will be considered a good enough indicator of model fit to the data. Hu and Bentler (1999) recommend using one of the relative fix indices such as comparative fit index (CFI) with a cutoff of >0.95 in combination with standardized root-mean-square residual (SRMR; absolute fit indices, good model < 0.08) or root-mean-square error of approximation (RMSEA; parsimony-adjusted fit indices, good model < 0.06) as indicators for good fit. Some researchers, including Hu and Bentler (1999), caution against using these cutoff values as golden rules because it might lead to incorrect rejection of acceptable models (Marsh *et al.*, 2004; Perry *et al.*, 2015).

Interpreting the Outputs from CFA

After making all the suggested analytical decisions, a researcher is now ready to apply a CFA to the data. Model fit indices that the researcher a priori decided to use are the first element of the output that should be interpreted from a CFA. If these indices suggest that the data do not fit the specified model, then the researcher does not have empirical support for using the hypothesized survey structure. This is exactly what happened when we initially ran a CFA on Diekman's goal-endorsement instrument example (see Box 3). In this case, focus should shift to understanding the source of the model misfit. For example, one should ask whether there are any items that do not seem to correlate with their specified latent factor, whether any correlations seem to be missing, or whether some items on a factor group together more strongly than other items on that same factor. These questions can be answered by analyzing factor loadings, correlation residuals, and modification indices. In the following sections, we describe these in more detail. See Boxes 3, 6, and 7 for examples of how to discuss and present output from a CFA in a paper.

Factor Loadings. As mentioned in *Brief Technical Description of Factor Analysis*, factor loadings represent how much of the respondent's response to an item is due to the factor. When a construct is measured using a set of items, the assumption is that each item measures a slightly different aspect of the construct and that the common variance among them is the best possible representation of the construct. High, but not too high, factor loadings for these items are preferred. If several items have high standardized factor loadings[9] (e.g., above 0.9), this suggests that they share a lot of variance, which indicates that these items may be too similar and thus do not contribute unique information (Clark and Watson, 1995). On the other hand, if an item has a low factor loading on its focal factor, it means that item shares no or little variance with the other items that theoretically belong to the same focal factor and thus its contribution to the factor is low. Including items with low factor loadings when combining the scores from several items into a single score

[7]It is necessary to set the metric to interpret factor loadings and variances in a CFA model. This is commonly done by either 1) choosing one of the factor loadings and fixing it to 1 (this is done for each factor in the model) or 2) by fixing the variance of the latent factors to 1. We have chosen the former approach for this example.

[8]For some software and estimation methods, model fit indices are also provided for EFA. In a similar way as for CFA, these model fit indices can be used to evaluate the fit of the data to the model.

[9]When using CFA, the default setting in most software is to provide factor loadings in the original metric of the items, such that the results are covariances between the items and the factor. Because these values are unstandardized, it is sometimes hard to interpret these relationships. For this reason, it is common to standardize factor loadings and other model relationships (e.g., correlations between latent factors), which puts them in the more familiar correlation format that is bounded by –1 and +1.

BOX 3. How to interpret and report CFA output for publication using the goal-endorsement example, initial CFA

Descriptive statistics

No items were missing more than 1.3% of their values, and this missingness was random (Little's MCAR test: chi-square = 677.719, *df* = 625, $p = 0.075$ implemented with the BaylorEdPsych package; Beaujean, 2012). Mean values for the items ranged from 4.1 to 6.3. Most items had a skewness and kurtosis below |1.0|, and all items had a skewness below |2.0| and kurtosis below |4.0|. Mardia's multivariate normality test (implemented with the psych package; Revelle 2017) showed significant multivariate skewness and kurtosis values. Intra-subscale correlations ranged from 0.02 to 0.73, and the lowest tolerance value was 0.36.

Interpreting output from the initial two-factor CFA

Results from the initial two-factor CFA indicated that, in our population, the data did not support the model specified. The chi-square test of model fit was significant ($\chi^2 = 1549$, $df = 229$, $p < 0.00$), but this test is known to be sensitive to minor model misspecification with large sample sizes ($n = 796$). However, additional model fit indices also indicated that the data did not support the model specified. SRMR was 0.079, suggesting good fit, but CFI was 0.818, and RMSEA was 0.084. Thus, the hypothesized model was not empirically supported by the data.

To better understand this model misspecification, we explored the factor loadings, correlational residuals, original interitem correlation matrix, and modification indices. Several factor loadings were well below 0.7, indicating that the factors did not explain these items well. Analysis of correlational residuals did not point out any special item-pair correlation as especially problematic; rather, several correlational residuals were residuals greater than |0.10|. Consequently, the poor model fit did not seem to be primarily caused by a few ill-fitting items. A reinvestigation of the interitem correlation matrix made when analyzing the factorability of the data (see the Supplemental Material, Section 1) suggested the presence of more than two factors. This was most pronounced for the agentic scale, for which some items had a relatively high correlation to one another and lower correlations to other items in that scale. Inspection of the modification indices suggested adding correlations between, for example, the items achievement and mastery. Together, these patterns indicate that the data might be better represented by more than two factors.

(sum, average, or common variance) will introduce bias into the results.[10] There is, however, no clear rule for when an item has a factor loading that is too low to be included. Bandalos and Finney (2010) argue that, because the items are specifically chosen to indicate a factor, one would hope that the variability explained in the item by the factor would be high (at least 50%). Squaring the standardized factor loadings provides the amount of variability explained in the item by the factor (R^2), indicating that it is desirable to have standardized factor loadings of at least 0.7 ($R^2 = 0.7^2 = \sim 50\%$). However, the acceptable strength of the factor loading depends on the theoretically assumed relationship between the item and the factor. Some items might be more theoretically distant from the factor and therefore have lower factor loadings, but still comprise an essential part of the factor. This reinforces the idea that there are no hard and fast rules in factor analysis. Even if an item does not reach the suggested level for factor loading, if a researcher can argue from a theoretical basis for its inclusion, then it could be included.

Correlation Residuals. As mentioned before, CFA is used to *confirm* a previously stated theoretical model. In CFA, the collected data are used to evaluate the accuracy of the proposed model by comparing the discrepancy between what the theoretical model implies (e.g., a two-factor model in the Diekman *et al.* [2010] example) and what is observed in the actual data. Correlation residuals represent the differences between the observed correlations in the data and the correlations implied by the CFA (Bandalos and Finney, 2010). Local areas of misfit can be identified by inspecting correlational residuals. Correlation residuals greater than |0.10| are indicative of a specific item-pair relationship that is poorly reproduced by the model (Kline, 2016). This guideline may be too low when working with small sample sizes and too large when working with large samples sizes and, as with all other fit indices, should only be used as one element among many to understand model fit.

Modification Indices. Most statistical software used for CFA provides modification indices that can easily be viewed by the user. Modification indices propose a series of possible additions to the model and estimate the amount the model's chi-square value would decrease if the suggested parameter were added (recall that a lower chi-square value indicates better model fit). For example, if an item strongly correlates with two factors but is constrained to only correlate with one, the modification index associated with adding a relationship to the second factor would indicate how much the chi-square model fit is expected to improve with the addition of this factor loading. In short, modification indices can be used to better understand which items or relationships might be driving the poor model fit.

If (and only if) theoretically justified, a suggested relationship can be added or problematic items can be removed during a CFA. However, caution should be taken before adding or removing any parameters (Bandalos and Finney, 2010). As Bandalos and Finney (2010) state, "Researchers must keep in mind that the purpose of conducting a CFA study is to gain a better understanding of the underlying structure of the variables, not to force models to fit" (p. 112). If post hoc changes to the model are made, the analysis becomes more explorative in nature, and thus tenuous. The modified model should ideally be confirmed with a new data set to avoid championing a model that has an artificially good model fit.

Best practice if the model does not fit (as noted in *Factor Analysis*) is to split the data and conduct a second round of

[10]When distilling the responses of several items into a single score, one is implicitly assuming that all of the items measure the underlying construct equally well (usually without measurement error) and are of equal theoretical importance. Fully discussing the nuances of how to create a single score from a set of items is beyond the scope of this paper, but we would be remiss if we did not at least mention it and encourage the reader to seek more information, such as DiStefano *et al.* (2009).

BOX 4. What to report in the methods of a publication for an EFA using the goal-endorsement example

Because the results from the initial CFA indicated that the data did not support a two-factor solution, we proceeded with an EFA to explore the factor structure of the data. The original sample was randomly divided into equal-sized parts, and EFA was performed on half of the sample ($n = 398$) to determine the dimensionality of the goal-endorsement scale and detect possible problematic items. This was followed by a CFA ($n = 398$) to confirm the result gained from the EFA. EFA and CFA were run using the R package lavaan (Rosseel, 2012).

Selecting an estimator for the EFA
Considering the ordinal and nonnormal nature of the data, a principal axis factor estimator was used to extract the variances from the data. Only cases with complete items were used in the EFA.

Factor rotation
Due to the fact that theory and the preceding CFA indicated that the different subscales are correlated, quartimin rotation (an oblique rotation) was chosen for the EFA.

Determining the number of factors
Visual inspection of the scree plot, parallel analysis (PA) based on eigenvalues from the principal components and factor analysis in combination with theoretical considerations were used to decide on the appropriate number of factors to retain. PA was implemented with the psych package (Revelle, 2017).

analyses starting with an EFA using half of the sample and then conducting a CFA with the other half (Bandalos and Finney, 2010). To see an example of how to write up this secondary CFA analysis, see Boxes 6 and 7 of the goal-endorsement example.

When the Model Fit Is Good. When model fit indices indicate that the hypothesized model is a plausible explanation of the relationships between the items in the data, factor loadings and the correlation between the latent variables in the model (so-called factor correlations) can be interpreted and a better understanding of the construct can be gained. It is also now appropriate to calculate and report the coefficient alpha, omega, or any other index of reliability for each of the subscales. The researcher can more confidently use the results from the instrument to make conclusions about the intended constructs based on combined scale scores (given that other relevant validity evidence presented in Table 1 also supports the intended interpretations).

If a researcher has used CFA to examine the dimensionality of the items and finds that the scale functions as intended, this information should be noted in the methods section of the research manuscript when describing the measurement instruments used in the study. At the very least, the researcher should report the estimator and fit indices that were used and accompanying values for the fit indices. If the scale has been adapted in some way, or if it is being empirically examined for the first time, all of the factor loadings and factor correlations should also be reported so future researchers can compare their values with these original estimates. These could be reported as a standalone instrument validation paper or in the methods section of a study using that instrument.

Analytical Considerations for EFA

If a researcher's data do not fit the model proposed in the CFA, then using the items as indicators of the hypothesized construct is not sensible. If the researcher wants to continue to use the existing items, it is prudent to investigate this misfit to better understand the relationships between the items. This calls for the use of an EFA, where the relationships between variables and factors are not predetermined (i.e., a model is not specified a priori) but are instead allowed to emerge from the data. As mentioned before, EFA could also be the first choice for a researcher if the instrument is in an early stage of development. We outline the steps for conducting an EFA in the following sections. See Box 4 for a description of how to describe analytical considerations for an EFA in the methods section.

Selecting an Estimator. Just as with CFA, the first step in an EFA is selecting a statistical method to use to extract the variances from the data. The considerations for the selection of this estimator are similar to those for CFA (see *Selecting an Estimator*). One of the most commonly used methods for extracting variance when conducting an EFA on ordinal data with slight nonnormality is principal axis factoring (Leandre *et al.*, 2012). If the items in one's instrument have fewer than five response options, WLS can be considered.

Factor Rotation. Factor rotation is a technical step to make the final output from the model easier to interpret (see Bandalos, 2018, pp. 327–334, for more details). The main decision for the researcher to make here is whether the rotation should be orthogonal or oblique (Raykov and Marcoulides, 2008; Leandre *et al.*, 2012; Bandalos, 2018). Orthogonal means that the factors are uncorrelated to one another in the model. Oblique allows the factors to correlate to one another. In educational studies, factors are likely to correlate to one another; thus oblique rotation should be chosen unless a strong hypothesis for uncorrelated factors exists (Leandre *et al.*, 2012). Orthogonal and oblique are actually families of rotations, so once the larger choice of family is made, a specific rotation method must be chosen. The specific rotation method within the oblique category that is chosen does not generally have a strong effect on the results (Bandalos and Finney, 2010). However, the researcher should always provide information about which rotation method was used (Bandalos and Finney, 2010).

Determining the Number of Factors. After selecting the methods for estimation and rotation, researchers must determine how many factors to extract for EFA. This step is recognized as the greatest challenge of an EFA, and the issue has generated a large amount of debate (e.g., Cattell, 1966; Crawford *et al.*,

BOX 5. How to interpret and report EFA output for publication using the goal-endorsement example

Initial EFAs

Parallel analysis based on eigenvalues from the principal components and factor analysis indicated three components and five factors. The scree plot indicated an initial leveling out at four factors and a second leveling out at six factors.

We started by running a three-factor model and then increased the number of factors by one until we had run all models ranging from three to six factors. The pattern matrices were then examined in detail with a special focus on whether the factors made theoretical sense (see Table 2 for pattern matrices for the three-, four-, and five-factor models). The three-factor solution consisted of one factor with high factor loadings for the items representing communal goals (explaining 17% of the variance in the data). The items originally representing agentic goals were split into two factors. One factor included items that theoretically could be described as prestige (explaining 12% of the variance in the data) and the other items related to autonomy and competency (explaining 11% of the variance in the data). The total variance explained by the three-factor model was 41%. In the four-factor solution, the autonomy and competency items were split into two different factors. In the five-factor solution, three items from the original communal goals scale (working with people, connection to others, and intimacy) contributed most to the additional factor. In total, 48% of the variance was explained by the five-factor model. For a six-factor solution, the sixth factor included only one item with pattern loadings greater than 0.40, and thus a six-factor solution was deemed to be inappropriate.

In conclusion, the communal scale might represent one underlying construct as suggested by previous research or it might be split into two subscales represented by items related to 1) serving others and 2) connection. Our data did not support a single agentic factor. Instead, these items seemed to fit on two or three subscales: prestige, autonomy, and possibly competency. Because all the suggested solutions (three-, four-, and five-factor solutions) included a number of poorly fitting items, we decided to remove items and run a second set of EFAs before proceeding to the CFA.

Second round of EFAs

On the basis of the results from the initial EFAs, we first continued with a three-factor solution, removing items with low pattern coefficients (<0.40; 10: success, 14: competition, and 22: intimacy, to begin with; Table 2). When these variables were removed in a stepwise manner, additional items now showed low pattern coefficients (<0.40) and/or low communalities in the new EFA solutions. The new items showing low pattern coefficients were items belonging to their own factors in the five-factor EFA (i.e., items representing competency and connection). Not until all items from these two scales were removed was a stable three-factor solution achieved with pattern coefficients >0.40. Thus, to achieve a three-factor solution, including only items with pattern coefficients >0.40, we had to drop 30% of the items and, consequently, extensively narrow the content validity of the scale.

TABLE 3. Standardized pattern coefficients for the Diekman *et al.* (2010) goal-endorsement instrument from the second EFA for the five-factor solutions[a]

		1	2	3	4	5
1	Power		0.75			
2	Recognition		0.60			
3	Achievement					0.81
5	Self-promotion		0.56			
6	Independence			0.65		
7	Individualism			0.69		
8	Status		0.76			
9	Focus on the self			0.50		
10	Success					0.55
11	Financial rewards		0.55			
12	Self-direction			0.55		
13	Demonstrating skills or competence					0.40
15	Helping others	0.84				
16	Serving humanity	0.80				
17	Serving community	0.80				
18	Working with people				0.94	
19	Connection with others				0.53	
20	Attending to others	0.75				
21	Caring for others	0.74				
23	Spiritual rewards	0.50	0.20			

[a]For clarity, pattern coefficients <0.2 are not shown.

To further explore a five-factor solution, we decided, on the basis of the empirical results and the theoretical meaning of the items, to stepwise remove items 4 (mastery), 14 (competition), and 22 (intimacy). We used an inclusive pattern coefficient cutoff (<0.40) for this initial round of validation, because we wanted to keep as many items as possible from the original scale. If some items continue to show pattern coefficients below 0.5 over repeated data collections, researchers should reconsider whether these items should be kept in the scale. The new 20-item five-factor solution resulted in theoretically the same factors as for the first five-factor EFA, but now all pattern coefficients but one were above 0.50 on the primary factor and below 0.20 on the other factors (Table 3). In total, 52% of the variance in the data was explained.

In conclusion, the initial CFA, as well as the EFA analysis, indicated that the two-dimensional scale previously suggested was not supported in our sample. The EFA analysis mainly indicated a three- or a five-factor solution. To achieve a good three-factor solution, we had to exclude 30% of the original items. The final three factors were labeled "prestige," "autonomy," and "service." Both the empirical data and theoretical consideration suggested two additional factors: a competency factor and a connection factor. We continued with this five-factor solution, as it allowed us to retain more of the original items and made theoretical sense, as the five factors were just a further parsing of the original agentic and communal scales.

2010; Leandre *et al.*, 2012). Commonly used methods are to retain all factors with an eigenvalue >1 or to use a scree plot. Eigenvalues are roughly a measure of the amount of information contained in a factor, so factors with higher eigenvalues are the most useful for understanding the data. A scree plot is a plot of eigenvalues versus number of factors. Scree plots allow researchers to visually estimate the number of factors that are informative by considering the shape of the plot (see the annotated output

BOX 6. How to interpret and report CFA output for publication using the goal-endorsement example, second CFA

Based on the results from the EFAs, a second CFA was specified using the five-factor model with 20 items (excluding 4: mastery, 10: competition, and 22: intimacy). The specified five-factor CFA demonstrated appropriate model fit ($\chi^2 = 266$, $df = 160$, $p < 0.00$, CFI = 0.959, RMSEA = 0.046, and SRMR = 0.050). Factor loadings were close to or above 0.70 for all but three items (Figure 2), meaning that, for most items, around 50% of the variance in the items was explained ($R^2 \approx 0.5$) by the theorized factor. This means that the factors explained most of the items well. Factor correlations were highest between the service and connection factors (0.76) and the autonomy and competency (0.67) factors. The lowest factor correlation found was between the prestige and service factors (0.21). Coefficient alpha values for the subscales were 0.81, 0.77, 0.66, 0.87, and 0.77 for prestige, autonomy, competency, service, and connection, respectively.

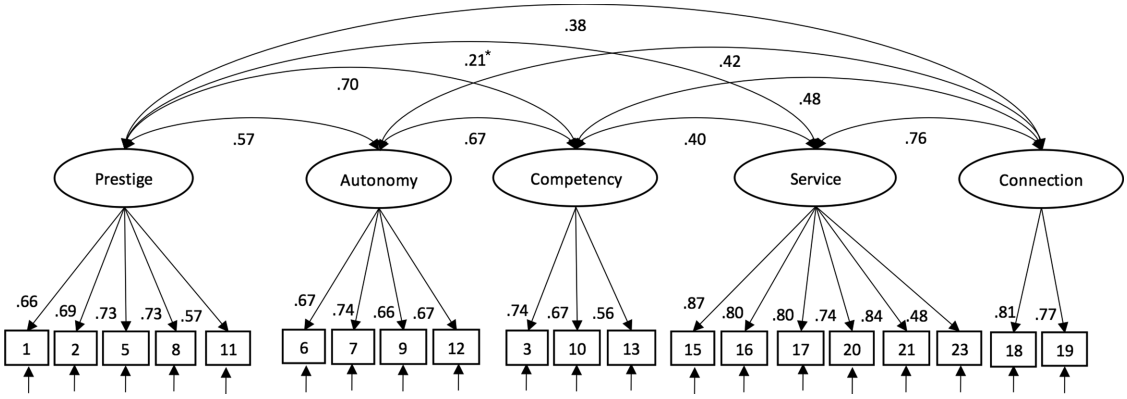

FIGURE 2. Results from the final five-factor CFA model. Survey items (for items descriptions see Table 3) are represented by squares and factors are represented by ovals. The numbers below the double-headed arrows represent correlations between the factors; the numbers by the one-directional arrows between the factors and the items represent standardized factor loadings. Small arrows indicate error terms. *, $p < 0.01$; $p < 0.001$ for all other estimates.

in the Supplemental Material, Section 2, for an example of a scree plot). These two methods are considered heuristic, and many researchers recommend also using parallel analysis (PA) or the minimum average partial correlation test to determine the appropriate number of factors (Ledesma and Valero-Mora, 2007; Leandre *et al.*, 2012; Tabachnick and Fidell, 2013). In addition, several statistics that mathematically analyze the shape of the scree plot have been developed in an effort to provide a nonvisual method of determining the number of factors (Ruscio and Roche, 2012; Raiche *et al.*, 2013).

We recommend using a number of these indices, as well as theoretical considerations, to determine the number of factors to retain. The results of all of the various methods discussed provide plausible solutions that can all be explored to evaluate the best solution. When these indices are in agreement, this provides more evidence of a clear factor structure in the data. To make each factor interpretable, it is of outmost importance that the number and nature of factors retained make theoretical sense (see Box 5 for a discussion on how many factors to retain). Further, the intended *use* for the survey should also be considered. For example, say a researcher is interested in studying two distinct populations of students. If the empirical and theoretical evidence supports both a two-factor and a three-factor solution, but the three-factor solution provides a clearer distinction between two populations of interest, then the researcher might choose the three-factor solution (see Box 7).

Interpreting Output from EFA

The aim of EFA is to gain a better understanding of underlying patterns in the data, investigate dimensionality, and identify potentially problematic items. In addition to the results from parallel analysis or other methods used to estimate the number of factors, other informative measures include pattern coefficients and communalities. These outputs from an EFA will be discussed in this section. See Box 5 for an example of how to write up the output from an EFA.

Pattern Coefficients and Communalities. Pattern coefficients and communalities are parameters describing the relationship between the items and the factors. They help researchers understand the meaning of the factors and identify items that do not empirically appear to belong to their theorized factor.

Pattern coefficients closely correspond to factor loadings in CFA, and they are commonly the focal output from an EFA (Leandre *et al.*, 2012). Pattern coefficients represent the impact each factor has on an item after controlling for the impact of all the other factors on that item. A high pattern coefficient suggests that the item is well explained by a particular factor. However, as with CFA, there is no clear rule as to when an item has a pattern coefficient too low to be considered part of a particular factor. Guidelines for minimum pattern coefficient values range from 0.40 to 0.70. In other words, all items with pattern coefficients equal to or higher than the chosen cutoff value can

BOX 7. Writing conclusions from factor analysis for publication using the goal-endorsement example

Conclusions

The results from the factor analysis did not confirm the proposed two-factor goal-endorsement scale for use with college STEM majors. Instead, our results indicated five subscales: prestige, autonomy, competency, service, and connection (Table 4). The five-factor solution aligned with Diekman *et al.*'s (2010) original two-factor scale, because communal items did not mix with agentic items. Our sample did, however, allows us to further refine the solution for the original two scales. Finer parsing of the agentic and communal scales may help identify important differences between students and allow researchers to better understand factors contributing to retention in STEM majors. In addition, with items related to autonomy and competency moved to their own scales, the refined prestige scale focusing on factors like power, recognition, and status may be a more direct contrast to the service scale. Additional evidence in support of this refinement include that the five-factor solution better distinguishes the service scale and the prestige scale (factor correlation = 0.21) than the two-factor solution (factor correlation between agentic and communal factors = 0.35). Further, retention may be significantly correlated to prestige but not to autonomy. Alternatively, differences between genders may exist for the service scale but not the connection scale.

On the basis of the result of this factor analysis, we recommend using the five-factor solution for interpreting the results of the current data set, but interpret the connection and competency scales with some caution, for reasons summarized in the next section.

Limitations and future studies

The proposed five-factor solution needs additional work. In particular, both the competency and connection scales need further development. Only two items represented connection, and this is not adequate to represent the full aspect of this construct, especially to make it clearly distinct from the construct of service. The competency scale included only three items, coefficient alpha was 0.66, and factor loadings for the scale were low (<0.40) for demonstrating skills or competency.

Another limitation of this study is that the sample consisted of 70% women, an overrepresentation of women for a typical undergraduate STEM population. Further studies should confirm whether the suggested dimensionality holds in a more representative sample. Future studies should also test whether the instrument has the same structure with STEM students from different backgrounds (i.e., measurement invariance should be investigated). The work presented here only establishes the dimensionality of the survey. We recommend the collection of other types of validity evidence, such as evidence based on content or relationships to other variables, to further strengthen our confidence that the scores from this survey represent STEM students' goal orientation.

TABLE 4. Proposed five-factor solution. Items within each factor are ordered by highest to lowest factor loadings

Service	Prestige	Autonomy	Connection	Competency
Helping others	Status	Individualism	Working with people	Achievement
Serving humanity	Power	Independence	Connection with others	Success
Serving community	Recognition	Self-direction		Competence
Attending to others	Self-promotion	Focus on the self		
Caring for others	Financial rewards			
Spiritual rewards				

be considered "good" items and should be kept in the survey (Matsunaga, 2010).

It is also important to consider the magnitude of any cross-loadings. Cross-loading describes the situation in which an item seems to be influenced by more than one factor in the model. Cross-loading is indicated when an item has high pattern coefficients for multiple factors. Using that item is problematic when creating a summed/mean score for a factor, as responses to that item are not uniquely driven by its hypothesized factor, but instead by additional measured factors. Cross-loadings higher than 0.20 or 0.30 are usually considered to be problematic (Matsunaga, 2010), especially if the item does not have a particularly strong loading on a focal factor.

Communality represents the percentage of the variance in responses on an item accounted for by all factors in the proposed model. Communalities are similar to R^2 in CFA (see *Factor Loadings*). However, in CFA, the variance in an item is only explained by one factor, while in EFA, the variance in one item can be explained by several factors. Low communality for an item means that the variance in the item is not well explained by any part of the model, and thus that item could be a subject for elimination.

We emphasize that, even if pattern coefficients or communalities indicate that an item might be subject for elimination, it is important to consider the alignment between the item and hypothesized construct before actually eliminating the item. The items in a scale are presumably chosen for some theoretical reason, and eliminating any items can cause a decrease in content validity (Bandalos and Finney, 2010). If any item is removed, the EFA should be rerun to ensure that the original factor structure persists. This can be done on the same data set, as EFA is exploratory in nature.

Interpreting the Final Solution. Once the factors and the items make empirical and theoretical sense, the factor solution can be interpreted, and suitable names for the factors should be chosen (see Box 5 for a discussion of the output from an EFA). Important sources of information for this include: the amount variance explained by the whole solution and the factors, factor correlations, pattern coefficients, communality values, and the

underlying theory. Because the names of the factors will be used to communicate the results, it is crucial that the names reflect the meaning of the underlying items. Because the item responses are manifestations of the constructs, different sets of items representing a construct will, accordingly, lead to slightly different nuanced interpretations of that construct. Once a plausible solution has been identified by an EFA, it is important to note that stronger support for the solution can be obtained by testing the hypothesized model using a CFA on a new sample.

CONCLUDING REMARKS

In this article, we have discussed the need for understanding the validity evidence available for an existing survey before its use in discipline-based educational research. We emphasized that validity is not a property of the measurement instrument itself but is instead a property of the instrument's use. Thus, each time a researcher decides to use an instrument, they have to consider to what degree evidence and theory support the intended interpretations and use of the instrument. A researcher should always review the different kinds of validity evidence described by AERA, APA, and NCME (2014; Table 1) before using an instrument and should identify the evidence they need to feel confident when employing the instrument for an intended use. When using several related items to measure an underlying construct, one important validity aspect to consider is whether a set of items can confidently be combined to represent that construct. In this paper, we have shown how factor analysis (both exploratory and confirmatory) can be used to investigate that.

We recognize that the information presented herein may seem daunting and a potential barrier to carrying out important, substantive, educational research. We appreciate this sentiment and have experienced those fears ourselves, but we feel that properly understanding procedures for vetting instruments before their use is essential for robust and replicable research. To reiterate, at issue here is the confidence and trust one can have in one's own research, both after its initial completion and in future studies that will rely on the replicability of results. Again, we can use an analogy for the measurement of unobservable phenomena: one would not expect an uncalibrated and calibrated scale to produce the same values for the weight of a rock. This does not mean that the uncalibrated scale will necessarily produce invalid measurements, only that one's confidence in its ability to do so should be tempered by the knowledge that it has not yet been calibrated. Research conducted using uncalibrated or biased instruments, regardless of discipline, is at risk of inferring conclusions that are incorrect. The researcher may make the appropriate inferences given the values provided by the instrument, but if the instrument itself is invalid for the proposed use, then the inferences drawn are also invalid. Our aim in presenting these methods is to strengthen the research conducted in biology education and continue to improve the quality of biology education in higher education.

ACKNOWLEDGMENTS

We are indebted to Ashely Rowland, Melissa McCartney, Matthew Kararo, Julie Charbonnier, and Marie Janelle Tacloban for their comments on earlier versions of this article. The research reported in this paper was supported by awards from the National Science Foundation (NSF DUE 1534195 and 1711082). This research was conducted under approved IRB 2015-06-0055, University of Texas at Austin.

REFERENCES

Allen, J. M., Muragishi, G. A., Smith, J. L., Thoman, D. B., & Brown, E. R. (2015). To grab and to hold: Cultivating communal goals to overcome cultural and structural barriers in first-generation college students' science interest. *Translational Issues in Psychological Science*, *1*(4), 331.

American Educational Research Association, American Psychological Association, and National Council for Measurement in Education (AERA, APA, and NCME). (2014). *Standards for educational and psychological testing*. Washington, DC.

Andrews, S. E., Runyon, C., & Aikens, M. L. (2017). The math–biology values instrument: Development of a tool to measure life science majors' task values of using math in the context of biology. *CBE—Life Sciences Education*, *16*(3), ar45.

Armbruster, P., Patel, M., Johnson, E., & Weiss, M. (2009). Active learning and student-centered pedagogy improve student attitudes and performance in introductory biology. *CBE—Life Sciences Education*, *8*(3), 203–213.

Bakan, D. (1966). *The duality of human existence: An essay on psychology and religion*. Oxford, UK: Rand McNally.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. New York: Guilford.

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis. Exploratory and confirmatory. In Hancock, G. R., & Mueller, R. O. (Eds.), *The reviewer's guide to quantitative methods in the social science* (pp. 93–114). New York: Routledge.

Beaujean, A. A. (2012). *BaylorEdPsych: R package for Baylor University educational psychology quantitative courses*. Retrieved from https://CRAN.R-project.org/package=BaylorEdPsych

Boomsma, A. (1982). Robustness of LISREL against small sample sizes in factor analysis models. In Joreskog, K. G., & Wold, H. (Eds.), *Systems under indirect observation: Causality, structure, prediction* (Part 1, pp. 149–173). Amsterdam, Netherlands: North Holland.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 212–225. doi: 10.1080/0969594X.2015.1063479

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Crawford, A.V., Green, S.B., Levy, R., Lo W-J., Scott, L., Svetina, D., & Thompson, M.S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement*, *70*(6), 885–901.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Diekman, A. B., Brown, E. R., Johnston, A. M., & Clark, E. K. (2010). Seeking congruity between goals and roles: A new look at why women opt out of science, technology, engineering, and mathematics careers. *Psychological Science*, *21*(8), 1051–1057.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1–11.

Eagly, A. H., Wood, W., & Diekman, A. (2000). Social role theory of sex differences and similarities: A current appraisal. In Eckes, T., & Trautner, H. M. (Eds.), *The developmental social psychology of gender* (pp. 123–174). Mahwah, NJ: Erlbaum.

Eddy, S. L., Brownell, S. E., Thummaphan, P., Lan, M. C., & Wenderoth, M. P. (2015). Caution, student experience may vary: Social identities impact a student's experience in peer discussions. *CBE—Life Sciences Education*, *14*(4), ar45.

Eddy, S. L., & Hogan, K. A. (2014). Getting under the hood: How and for whom does increasing course structure work?. *CBE—Life Sciences Education*, *13*(3), 453–468.

Finney, S. J., & DiStefano, C. (2006). Nonnormal and categorical data in structural equation modeling. In Hancock, G. R., & Mueller, R. O. (Eds.), *A second course in structural equation modeling* (pp. 269–314). Greenwich, CT: Information Age.

Fowler, F. J. (2014). *Survey research methods*. Los Angeles: Sage.

Gagne, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, *41*(1), 65–83.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*(1), 121–135.

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 198–211. doi: 10.1080/0969594X.2015.1060192

Kline, R. B. (2016). *Principles and practise of structural equation modeling* (4th ed.). New York: Guilford.

Leandre, R., Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford, UK: Oxford University Press.

Ledesma, R.D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation*, *12*(2)

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437–448.

Marsh, H.W., Hau, K-T, & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341, doi: 10.1207/s15328007sem1103_2

Matsunaga, M. (2010). How to factor analyze your data right: Do's, don't and how-to's. *International Journal of Psychological Research*, *3*. Retrieved February 24, 2019, from www.redalyc.org/html/2990/299023509007/

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://dx.doi.org/10.1037/met0000144

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practise*, *16*(2), 16–18.

Messick, S. (1995). Validity of psychological-assessment—Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.

Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Journal of Multivariate Behavioral Research*, *22*(3), 267–305. doi: 10.1207/s15327906mbr2203_3

Perry, J. L., Nicholls, A. R, Clough, P. J., & Crust, L. (2015). Assessing model fit: Caveats and recommendations for confirmatory factor analysis and exploratory structural equation modeling. *Measurement in Physical Education and Exercise Science*, *19*(1), 12–21.

Prentice, D. A., & Carranza, E. (2002). What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly*, *26*(4), 269–281.

Raîche, G., Walls, T., Magis, D., Riopel, M., & Blais, J.-G., (2013). Non-graphical solutions for Cattell's scree test. *Methodology*, *9*, 23–29. doi: 10.1027/1614-2241/a000051

Raykov, T., & Marcoulides, G. A. (2008). *An introduction to applied multivariate analysis*. New York: Routledge.

Raykov, T., & Marcoulides, G. A. (2017). Thanks coefficient alpha, We still need you! *Educational and Psychological Measurement*. doi: 10.1177/0013164417725127

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved February 24, 2019, from www.R-project.org

Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: A primer for discipline-based education researchers. *CBE—Life Sciences Education*, *15*(1), rm1.

Revelle, W. (2017). *psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA*. Retrieved February 24, 2019, from https://CRAN.R-project.org/package=psychVersion=1.7.8

Rissing, S. W., & Cogan, J. G. (2009). Can an inquiry approach improve college student learning in a teaching laboratory? *CBE—Life Sciences Education*, *8*(1), 55–61.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment*, *24*(2), 282.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107.

Slaney, K. (2017). Construct validity: Developments and debates. In *Validating psychological constructs:Historical, Philosophical, and Practical Dimensions* (pp. 83–109) (Palgrave studies in the theory and history of psychology). London: Palgrave Macmillan.

Smith, J. L., Cech, E., Metz, A., Huntoon, M., & Moyer, C. (2014). Giving back or giving up: Native American student experiences in science and engineering. *Cultural Diversity and Ethnic Minority Psychology*, *20*(3), 413.

Stephens, N. M., Fryberg, S. A., Markus, H. R., Johnson, C. S., & Covarrubias, R. (2012). Unseen disadvantage: How American universities' focus on independence undermines the academic performance of first-generation college students. *Journal of Personality and Social Psychology*, *102*(6), 1178–1197. Retrieved from http://doi.org/10.1037/a0027143

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin*, *135*(6), 859.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed). Boston: Pearson.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. Retrieved February 24, 2019, from http://doi.org/10.5116/ijme.4dfb.8dfd

Wachsmuth, L. P., Runyon, C. R., Drake, J. M., & Dolan, E. L. (2017). Do biology students really hate math? Empirical insights into undergraduate life science majors' emotions about mathematics. *CBE—Life Sciences Education*, *16*(3), ar49.

Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review*, *12*(3), 265–310. doi: 10.1016/0273-2297(92)90011-p

Wiggins, B. L., Eddy, S. L., Wener-Fligner, L., Freisem, K., Grunspan, D. Z., Theobald, E. J., & Crowe, A. J. (2017). ASPECT: A survey to assess student perspective of engagement in an active-learning classroom. *CBE—Life Sciences Education*, *16*(2), ar32.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*(6), 913–934.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806–838.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, *29*(4), 377–392.

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79–94.