MEASURING TEACHER SELF-EFFICACY IN THE MORAL WORK OF TEACHING

Richard D. Osguthorpe, Sarah K. Clark, and Richard R. Sudweeks
Brigham Young University

The Self-Efficacy in the Moral Work of Teaching Scale (SMWTS) is designed to measure the self-efficacy of teachers in relation to the moral work of teaching. We used confirmatory factor analysis (CFA) and exploratory structural equation modeling (ESEM) to examine the factor structure of the SMWTS in a sample of preservice teachers (n = 152) attending a teacher education program in the Western United States. The results of this initial study demonstrate that a bifactor ESEM (B-ESEM) model better accounted for the construct-relevant multidimensionality than alternative models and that the SMWTS can be used to produce scores with good reliability and factor analytic evidence of validity within a sample of preservice teachers.

Keywords: character education, moral education, moral work of teaching, teacher self-efficacy, rating scale

There is a rich and expanding research literature on measuring teacher self-efficacy in specific classroom contexts and content areas. Building on foundational work in teacher self-efficacy (see Bandura, 1997; Henson, 2002, Pajares, 1996), researchers have explored teacher self-efficacy in areas such as science (Riggs & Enochs, 1990), classroom management (Emmer, 1990), special education (Coladarci & Breton, 1997), as well as decision-making, accessing school resources, enlisting parent and community involvement, and considering school climate (Bandura, 1997). There have only been a few studies that have explored the relationship between teacher self-efficacy and teaching moral character to students and engaging in their moral development. For example, Milson and Mehlig (2002) developed the Character Education Efficacy Instrument that combined notions of general and personal teacher self-efficacy. Drawing on that work, Narvaez et. al, (2008) validated the Teacher Efficacy for Moral Education Instrument, which they described as a “more focused measure of personal self-efficacy for fostering moral character in students” (p. 5). These instruments focused primarily on the teacher as character educator and moral educator, respectively.

• **Correspondence concerning this article should be addressed to:** Richard D. Osguthorpe, rich_osguthorpe@byu.edu

Journal of Character Education, Volume 18(2), 2022, pp. 93–111
ISSN 1543-1223

However, there does not appear to be any research or instruments that have explored teacher self-efficacy related, more broadly, to the moral work of teaching. In particular, there is a lack of attention to the teacher's role as moral agent in the classroom, in addition to the teacher's role as moral or character educator. The goal of the current study was to attend to this gap in the research literature. This study describes the validation of an instrument for measuring teacher self-efficacy in the moral work of teaching.

In our exploration of teacher self-efficacy related to the moral work of teaching, we rely on a broad conception of the moral work of teaching, and we make important assumptions. This conception and these assumptions are detailed in previous work (see Sanger & Osguthorpe, 2011). To briefly summarize this conception of the moral work of teaching:

> We conceptualize the [moral work of teaching] using a broad and inclusive view of the moral domain—a complex aspect of human life that involves issues of what is good, right, and virtuous in what and how we think, feel, and act (Darwall, 1998), in who we are, and how we relate to others. In turn, the [moral work of teaching] is taken to be the elements of practice that are within, or meaningfully connected to, the moral domain. These elements prominently include all those things that go into being a morally good person as a teacher, and also in the effects teachers have on the moral functioning and development of students. (p. 569)

Connected to this conceptualization (and following from the same framework), we assume that teaching is an inherently moral endeavor (Fenstermacher, 1990; Hansen, 2001; Jackson et al., 1993; Oser, 1994; Sockett, 1993); and that the moral work of teaching is of significant importance in relation to the aims of education (e.g., Brighouse, 2006; Dewey, 1916; Giroux, 1988; Noddings, 2013; Whitehead, 1959) and its moral mission (e.g. Arthur, 2008; McClellan, 1999; Swartz, 2010; Wanxue & Hanwei, 2004).

Most importantly for the current study, this conceptualization of the moral work of teaching makes a distinction between "being a morally good person" in the role of teacher (teaching morally) and "the effects teachers have on the moral functioning and development of students" (teaching morality; see also Fenstermacher et al., 2009 2009). The latter encompasses a comprehensive view of character education and its scientific underpinnings (see Berkowitz, 2002). And this distinction suggests that a teacher's moral work goes beyond fostering the character development of students to include the ways in which every activity of teaching is informed by the morality of the teacher.

If teaching is an inherently moral endeavor, then it follows that preparation to be a teacher should attend to the moral work of teaching. In a constructivist approach to such preparation, attention to moral matters might begin with an examination of the beliefs that teacher candidates hold in relation to the moral work of teaching, including the strength with which these beliefs are held. In order to construct an instrument for identifying and examining these beliefs, this study reports on the development and validation of a scale for measuring teacher self-efficacy beliefs related to the moral work of teaching. We designed questions specific to one of three factors or traits: (a) teaching morally (to teach in ways that align with what is good, right, virtuous, and caring); (b) teaching morality via modeling (to demonstrate for another what it means to be a good, virtuous, caring person); and (c) teaching morality via instruction (to design and implement learning activities that help others be good, virtuous, and caring).

We define each of these three potential factors in our description of the instrument that follows. It is important to note that the primary purpose of this instrument design was to expand the notion of self-efficacy in the moral work of teaching to not only include teaching for character and moral development (teaching morality), but also the many ways in which every activity of teaching can be described in moral terms (teaching morally). This broader conception of self-efficacy in the moral work of teaching connects directly to the reasons that teachers choose a career in teaching in the first place.

METHOD

In this section, we describe (a) the study participants, (b) the characteristics of the instrument, (c) the procedures used to administer the instrument, and (d) the procedures used to analyze the responses.

Participants

The participants consisted of 152 preservice teachers who were enrolled in coursework connected to a teacher education program at a large comprehensive university in the Intermountain West. Of the 152 participants, 41 were enrolled in a teacher education program in elementary education, 94 were enrolled in a pre-program course that is required for admission to the elementary education program, and the remaining seven had no intention of enrolling in a teacher education program. Participants were selected according to these course enrollments. Most (n = 142) identified as White, while almost all (150) identified as female, and all but four (148) were between ages 18-25.

Instrument

Consistent with our theory that the moral work of teaching includes three different, conceptually-related subdomains, the first version of the instrument, entitled the Self-Efficacy for the Moral Work of Teaching Scale (SMWTS), consisted of 28 items intended to elicit responses that would be indicative of teachers' self-efficacy toward (a) teaching morally, (b) teaching morality via modeling, and (c) teaching morality via instruction. We determined that focusing on virtue and character (as opposed to a deontological ethics or consequentialist ethics) would provide the most accessible moral language for this study. Thus, following Peterson and Seligman's (2004) "six broad categories of virtue [that] emerge consistently from historical surveys," we selected "character strengths" in each virtue category that teachers might use as "distinguishable routes to displaying one or another of the virtues." We limited our selection of character strengths to those we determined to have the most germane application to the practice of teaching.

Each item was a direct question asking teachers to rate their self-efficacy regarding their ability to successfully perform a specific teaching task related to some specific aspect of one of the three facets of the overall construct. With teacher self-efficacy being a future-oriented belief about one's level of competence within a specific context, the most common item stems we used were "How well can you …" and "To what extent can you …" Each item was accompanied by a common set of five response options representing ordered points on a rating scale ranging from 1 (*not at all well*) to 5 (*extremely well*).

The first page of the questionnaire included a written explanation of the distinction between these three aspects of the moral work of teaching. The 28 questions were grouped by category, and participants were reminded of the distinction between categories prior to each section. We provided more detail about what it means to teach morally by explaining in the preface that:

> When teachers teach morally, their teaching is informed by what is good, right, virtuous, and caring. For example, teachers might respectfully attend to controversial matters in the curriculum, fairly grade student work, compassionately provide extra instruction for complex subject matter, equitably allocate time and effort to disadvantaged students, persistently help a struggling student, and so on.

We also provided additional definition of teaching morality via modeling in the preface, suggesting that:

> When teachers teach morality via modeling, they teach in ways that intentionally demonstrate for students how to be morally good. That is, their

> actions exemplify how to be good, virtuous, and caring. For example, teachers might model kindness and respect by formally greeting individual students as they enter the classroom. Or they might model fairness and compassion when deciding between combatants on the playground. Similarly, teachers might model open-mindedness when discussing a controversial issue in a social studies lesson, and they might model perseverance by participating in remedial efforts during the lunch hour with select students and colleagues.

And, finally, we expanded on the notion of teaching morality via instruction in the preface by stating that:

> When teachers teach morality via instruction, they provide lessons, experiences, and curricula that seek to instill moral goodness. That is, they design and implement learning activities that provide opportunities for students to experience how to be good, virtuous, and caring. For example, teachers might seek to instill honesty in students by defining honesty, by reading a story that highlights the importance of honesty, by engaging students in a discussion of how honesty applies to classroom life, and then by encouraging students to be honest when taking a test. They might also seek to instill a trait such as fairness by engaging students in a role-playing experience or scenario that involves an in-depth analysis and discussion of a moral dilemma related to fairness or other related ethical issues.

This prefatory content was included after we conducted a series of read-aloud sessions on an initial administration of the survey with eight individuals representative of the sample (but who were not included in any additional administration of the survey). In those read-aloud sessions, we determined that participants had difficulty distinguishing the different ways in which they might attend to the moral work of teaching in their classroom. In those early versions, we also limited the type and number of character strengths used in our items, as it was determined that they struggled to see the relevance of some character strengths to classroom practice (such as loyalty and social responsibility). We also eliminated other character strengths that respondents found difficult to quickly grasp and understand (such as prudence). Again, following Peterson and Seligman's (2004) categories of virtues, we ultimately settled on the following character strengths (with their connected category of virtue): fairness (justice), care (humanity), persistence (courage), open-mindedness (wisdom), hopefulness (transcendence), and honesty (courage). Coverage of all categories was not important to the study, but because prudence was eliminated in early versions of the read-aloud, we did not include a character strength for the virtue category of temperance. These character strengths served primarily as a source for illustrating the ways in which teachers might indicate their teacher self-efficacy toward (a) teaching morally, (b) teaching morality via modeling, and (c) teaching morality via instruction. See the Appendix for the final 18 items included on the Self-Efficacy for the Moral Work of Teaching Scale.

Procedures

The instrument was administered using *Qualtrics* software. Participants were given a link to access during class, and they either completed the survey during the immediate class session or soon thereafter. Most students took between 5–10 minutes to respond to the questions.

Analysis

Initial Analysis

After examining the frequency distribution of responses to each item and conducting an initial confirmatory factor analysis which posited that each item would load on one and only one of the three intended factors, we eliminated 10 items which had either a low loading (< .30) on their specified factor or a high cross-loading (> .50) on an unintended factor.

Second Confirmatory Factor Analysis

We then performed a confirmatory factor analysis using the remaining 18 items. The items were hypothesized to measure three factors.

- Items 1–6: Teaching Morally;

- Items 7–12: Teaching Morality via Modeling; and
- Items 13–18: Teaching Morality via Instruction.

The diagram in the upper left corner of Figure 1 depicts this model with the hypothesized three factors. The 18 small rectangles arranged in a horizontal line across the bottom of this diagram each represent one of the items in the final scale. The three ovals in this diagram each represent one of the hypothesized factors. Each straight line with a single arrowhead that connects an item to one of the ovals represents a factor loading which summarizes the effect of the corresponding factor on the students' responses to that item. Each curved line with double-headed arrows connecting two ovals represents the correlation between the corresponding factors.

The results of this CFA indicated that the model fit the data quite well, but the three factors were highly correlated. The correlation between Factors 1 and 2 was .930, while the correlation between Factors 2 and 3 was .966, and the correlation between Factors 1 and 3 was .854. In applied research inter-factor correlations above .850 are indicative of a lack of discriminant validity (Brown, 2015; Marsh et al., 2014). These high correlations caused us to wonder whether the three factors were all manifestations of a single, underlying factor. Therefore, our next step was to test a one-factor model in which all 18 items were hypothesized to load on a single factor. Since this one-factor model did not fit the data as well as the three-factor model, we concluded that the data must be multidimensional rather than unidimensional.

Since 2009, multiple scholars have claimed that the traditional CFA approach which we have just reported has an important limitation because it is based on the questionable assumption that each item in a multifactor instrument should load solely on one factor and that all cross-loadings on other factors should be constrained to zero (Asparouhov & Muthén, 2009; Cooke & Sellbon, 2018; Marsh et al., 2009; Marsh et al., 2014; Morin, Myers, & Lee, 2020).

These scholars typically refer to the traditional approach to CFA which excludes cross-loadings as the Independent Cluster Model of CFA (ICM-CFA). They use the term *independent* not because the resulting factors are uncorrelated, but because the items specified to load on each factor are constrained to be independent of any and all items specified to load on a different factor. These scholars further assert that ICM-CFA models have two deficiencies:

1. Items in an instrument designed to measure multiple constructs are rarely so pure that they measure the target construct exclusively without also measuring other related construct to some degree. Morin et al. (2016) assert that items are especially likely to tap more than one targeted construct when instruments are "designed to assess conceptually-related and partially overlapping domains." They further declare that "Although 'pure' indicators may exist, we surmise that such indicators remain at best a convenient fiction" (p. 279).
2. Even when a CFA model fits the data well, the routine practice of constraining cross-loadings to zero has the undesirable consequence of inflating the correlations among the factors when compared to exploratory factor analysis unless all the non-targeted cross-loadings are close to zero. Morin, Myers and Lee (2020) assert that this practice "undermines the discriminant validity of the factors and may cause multicollinearity in the estimation of relations with covariates" (p. 1046).

Since we were trying to measure constructs that were "conceptually-related" and at least "partially overlapping," when we observed the high inter-factor correlations in the three-factor ICM_CFA model we began to wonder if the high inter-factor correlations which we had observed in the three-factor ICM-CFA model were due to the fact that the cross-loadings had been constrained to zero and if a model that included freely-estimated cross-loadings would produce smaller correlations among the three factors.

To overcome the limitations of ICM-CFA, Asparouhov and Muthén (2009), and Marsh et al. (2009) proposed a hybrid approach that combines the advantages of exploratory factor analysis (EFA) with the advantages of CFA. This hybrid has come to be known as Exploratory Structural Equation Modeling (ESEM). Marsh et al. (2014) declared that ESEM is "an integration of the best features of exploratory and confirmatory factor analysis" (p. 85). One of the distinguishing features of ESEM models is that cross-loadings can be estimated in addition to item loadings on each intended factor.

We utilized the framework proposed by Morin et al. (2016) as a tool to help us determine which multidimensional model best fit the self-efficacy ratings we had collected. Their framework includes six different classes of multidimensional models which differ in terms of two considerations: (a) whether a model permits items to cross-load on to non-targeted factors, and (b) whether the model includes first-order factors only, a second-order factor, or a bifactor structure. This framework proposed by Morin, Arens, & Marsh (2016) recognizes that just because one particular model is found to have acceptable fit to the data does not mean that is the best model. Their framework is based on the assumption that the most defensible way to determine whether the items that make up a scale measure (a) a single unidimensional construct, (b) a first-order model with multiple factors, or (c) a more complex global construct with specificities is to compare and contrast rival models representing each of these alternatives (Morin et al., 2016).

Figure 1 displays the six classes of alternative multidimensional models in the Morin et al. (2016) framework. We have adapted each of the six diagrams to represent the 18 items in our data. The six models are systematically ordered on the basis of the two dimensions suggested by Morin et al. (2016). The distinction between the two columns shown in Figure 1 refers to whether the various models permit cross-loadings. The three models in the column on the left each depict a ICM-CFA model that does not include any cross-loadings. In contrast, each of the three models in the column on the right is an ESEM model which includes cross-loadings symbolized by dashed straight lines with single arrowheads.

The distinction between the models in each row in Figure 1 is also relevant to the present study. The two models in the top row each include curved lines with double-headed arrows connecting a pair of factors. Therefore, each of these two diagrams represents a model that directly estimates the correlation among each pair of factors.

In the two models in the second row, the curved lines in the top row have been replaced with a single, second-order factor which is hypothesized to have a direct effect on each of the three first-order factors, but its effect on the observed variables is hypothesized to be mediated through the first-order factors and is therefore indirect. Consequently, in the models in the second row, the straight lines with single-headed arrows which emanate from the second-order factor ~~but~~ do not connect with any of the individual items.

Each of the two models in the bottom row of Figure 1 is a bifactor model. In the bifactor models, the curved lines with double-headed arrows shown in the models in the first row have been replaced with a general (or global) factor labeled "G" which is hypothesized to account for and explain the association among the factors observed in the models in the first row. Bifactor models have traditionally been characterized by two distinguishing features:

1. Each item is hypothesized to load on a general factor and on one specific factor. The general factor is defined by the shared meaning that all 18 items have in common. In contrast, the meaning of each specific factor is defined by the variance shared by a subset of related items above

and beyond the variance that is common to all 18 items.
2. The specific factors are expected to be uncorrelated with each other and with the general factor.

The task of the researcher using this framework reduces to answering two questions:

1. Does an ESEM model that permits cross-loadings fit better than a model that constrains all cross-loadings to zero? In other words, does a model in the righthand column fit better than the corresponding model in the lefthand column.
2. Which type of model structure best fits the data and best accounts for the relationship among the items and factors? In other words, does a second-order model or a bifactor model fit better than a first-order model?

We used the online syntax generator developed by De Beer and van Zyl (2019) to produce the *M*plus code needed to estimate the three ESEM models. This convenient online tool can be accessed at http://www.surveyhost.co.za/esem/. We followed the guidelines presented in the tutorial published by van Zyl and ten Klooster (2022) to use this tool to generate the syntax for ESEM models.

All analyses were performed using *M*plus version 8.2. Since each item in the scale included only five response options, we avoided using the default Maximum Likelihood estimation procedure in *M*plus which assumes that the responses to the various items are normally distributed. Instead, we used the weighted least squares mean and Variance option to estimate the parameters in each model.

RESULTS

Factor Analytic Studies

We first report the results of our factor analytic studies aimed at finding the best model to use to describe the factor structure of the data being analyzed. Then, we report the results of our reliability analyses.

The Three-Factor Model With No Cross Loadings

The standardized factor loadings for the three-factor ICM-CFA model are relatively high. The loadings of the six items on the Teaching Morally factor ranged from .592 to .732. The loadings of the six items on the Teaching Morality via Modeling factor ranged from .699 to .791, and the loadings of the six items on the Teaching Morality via Instruction factor ranged from .747 to .832. However, the bivariate correlations among the three factors were high as reported in the cells below the main diagonal in Table 1. The lowest of the three inter-factor correlations was .854 and the highest was .966.

Although the fit statistics indicate that this three-factor ICM-CFA model fit the data well (CFI = .974, TLI = .970, RMSEA = .063, SRMR = .055), this finding does not mean that it is the only model that would fit the data. Consequently, we estimated several alternative models in an attempt to find a model which would fit the data better and account for the high correlations among the three first-order factors.

The One-Factor Model

The one-factor model represents the hypothesis that the reason the three factors in the ICM-CFA model are so highly correlated is because they are all manifestations of one common, underlying factor. The fit statistics for this alternative model are reported in Column A in Table 2 (CFI = .969, TLI = .965, RMSEA = .068, SRMR = .059). The values of the relative fit statistics (CFI and TLI) are larger for three-factor CFA model than for the one-factor model. In addition, the values of the absolute fit statistics (RMSEA and SRMR) are smaller for the three-factor CFA model than for the one-factor model. This pattern provides

TABLE 1
Correlations Among Factors in the Three-Factor Models Estimated Using Confirmatory Factor Analysis and Exploratory Structural Equation Modeling Respectively

Confirmatory Factor Analysis	*Exploratory Structural Equation Modeling*		
Variable	*Teaching Morally*	*Teaching Morality Via Modeling*	*Teaching Morality Via Instruction*
Teaching morally	—	.404	.634
Teaching morality via modeling	.930	—	.429
Teaching morality via instruction	.854	.966	—

Note: The three correlation coefficients in the cells below the main diagonal report the degree of bivariate association among pairs of variables estimated by the ICM-CFA model in which cross-loadings were constrained to zero. The coefficients above the main diagonal report the bivariate correlation among pairs of variables estimated by the ESEM model which allowed nonnegligible cross-loadings to be freely estimated.

TABLE 2
Fit Statistics for Five Rival Models

	CFA Models[a]		*ESEM Models*[b]		
Statistic	*One Single Factor (A)*	*Three First-Order CFA Factors (B)*	*Three First-Order ESEM Factors (C)*	*Second-Order ESEM Model (D)*	*Bifactor ESEM Model (E)*
Number of parameters freely estimated	70	73	104	104	119
Chi square measure of misfit	230.458	212.398	159.487	159.487	126.237
Degrees of freedom	135	132	102	102	87
Probability	.0000	.0000	.0002	.0002	.0038
Comparative Fit Index	.969	.974	.982	.982	.988
Tucker-Lewis Index	.965	.970	.973	.973	.979
Root mean square error of approximation	.068	.063	.061	.061	.054
90% confidence interval	.053—.083	.047—.078	.042—.078	.042—.078	.032—.074
Probability RMSEA < .0	.028	.088	.164	.164	.352
Standardized root mean square residual	.059	.055	.042	.042	.036
Adjusted chi square test of difference in nested models	—	15.569	63.281		35.307
Degrees of freedom	—	3	30		15
Statistical significance	—	.001	.000		.0022

Notes: [a]All cross-loadings are constrained to zero in the CFA models. [b]In the ESEM models, target rotation was used to estimate nonnegligible cross-loadings.

support for the three-factor model. Since the one-factor model is nested within the three-factor model, we used the adjusted chi square DIFFTEST option in *M*plus to empirically test whether the difference in the fit of these two models was statistically significant. The results indicated that the three-factor model significantly reduced the misfit (adjusted chi square = 15.569, $df = 3$, $p < .001$) as reported in the last three rows of Table 2 in the column for the three-factor CFA model. Therefore, we rejected the hypothesis that the 18 items were all manifestations of a single latent variable.

The Three-Factor ESEM Model

Having discounted the unidimensional model, we were left with the task of ascertaining which multidimensional model best fit the data and accounted for the high correlations among the three factors in the ICM-CFA model. Therefore, our next step involved comparing the three-factor CFA model (Panel A in Figure 1) with the three-factor ESEM model (Panel B in Figure 1). These are the two models depicted in the top row of Figure 1. This comparison focuses on the question of whether including cross-loadings in the model leads to better fit and a substantial reduction in the inter-factor correlations.

The cells above the main diagonal of the correlation matrix in Table 1 display the correlation coefficients describing the relationship between each pair of factors estimated by the ESEM model. The extent to which each coefficient below the diagonal is larger than its counterpart above the main diagonal indicates the degree to which constraining cross-loadings to zero inflated estimates of the inter-factor correlations obtained from the ICM-CFA model. Including cross-loadings in the model decreased the correlation between Factors 1 and 2 from .930 to .404, while the correlation between Factors 1 and 3 decreased from .854 to .634 and the correlation between Factors 2 and 3 decreased from .966 to .429.

The fit statistics for the two rival three-factor models are displayed in columns B and C respectively of Table 2. The CFI and TLI indices are higher for the ESEM model than for the ICM-CFA model, and the RMSEA and SRMR fit statistics are smaller for the ESEM model than for the ICM-CFA model. This pattern supports the conclusion that the ESEM model should be preferred. This conclusion is further evidenced by the fact that the chi square DIFFTEST indicated that the ESEM model fit the data significantly better than the ICM-CFA model (chi square = 63.281, $df = 30$, $p < .0001$) as reported in the last three rows of Column C in Table 2.

The Hierarchical ESEM Model

From a statistical point of view, a second-order model with only three first-order factors is "just identified" (Brown, 2015, p. 292). Consequently, the fit statistics for the second-order ESEM model are equivalent to the corresponding fit statistics of the three-factor ESEM model reported in Column C of Table 2. Therefore, the question of whether the second-order ESEM model fits better than the first-order ESEM model is moot in this instance.

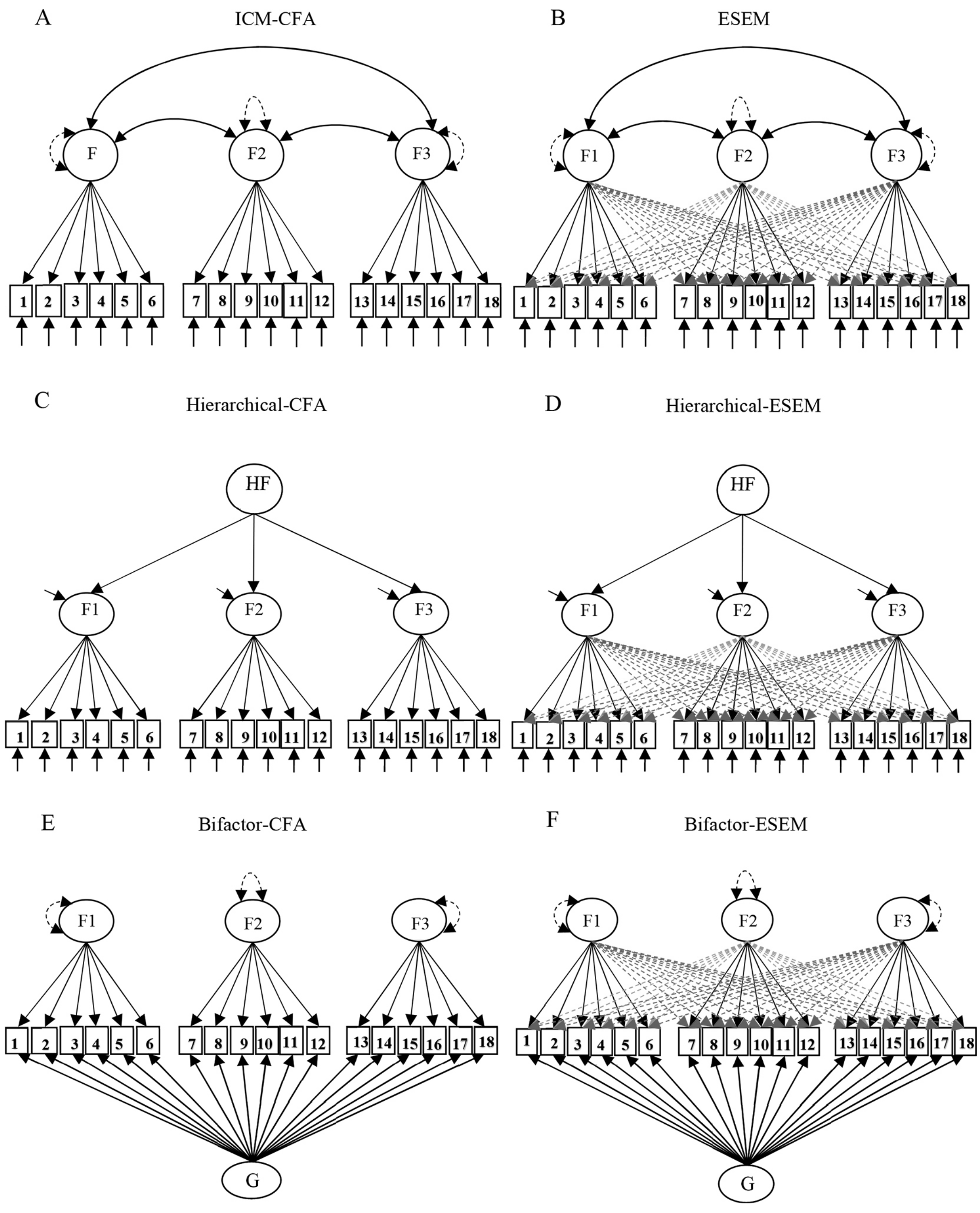
The Bifactor ESEM Model

The final model we estimated was the bifactor ESEM (B-ESEM) model depicted in Panel F of Figure 1**.** In order for the B-ESEM to be preferred, the following conditions should exist in addition to a significant improvement in the model fit statistics (Morin, et al., 2016). First, and most importantly, the global factor in the B-ESEM should be well defined. Second, ideally at least one of the specific factors should be well defined. Third, the absolute value of the cross-loadings estimated by the B-ESEM model should be smaller than in the three-factor ESEM model.

Table 3 reports the standardized loadings for the general factor and each of the specific factors estimated by the B-ESEM. Inspection of these loadings indicates that the general factor is well defined by strong and significant standardized loadings ranging from .547 to .812 with a mean equal to .701. Seventeen (94%) of the 18 items have a loading greater than .610. Item 3 has the weakest loading on the general factor, but its loading (.547) is still greater than .500. Therefore, we concluded that the global factor is well defined.

The absolute values of the standardized loadings on the three specific factors range

FIGURE 1
Path Diagram for Six Rival Models



from .002 to .455 on the Teaching Morally factor, .017 to .325 on the Teaching Morality vis Modeling factor, and .009 to .430 on the Teaching Morality via Instruction factor. None of the 18 items has a higher standardized loading on a specific factor than on the general fac-

TABLE 3
Factor Loadings for the Bifactor ESEM Model

Item	*General Factor*	*Specific Factors*			*Residuals*
		Teaching Morally	*Teaching Morality Via Modeling*	*Teaching Morality Via Instruction*	
1	.653	.002	.182	–.152	.517
2	.728	.041	.289	–.240	.327
3	.547	.393	–.174	–.096	.507
4	.732	.065	–.302	–.274	.294
5	.613	.455	.209	.136	.355
6	.623	.212	-.036	–.009	.566
7	.751	–.046	.080	–.070	.423
8	.726	.006	.325	.021	.367
9	.688	.133	–.241	.260	.383
10	.740	–.108	–.253	–.066	.372
11	.742	.178	.069	.257	.347
12	.649	.179	.031	.256	.480
13	.812	–.130	.054	.170	.292
14	.774	–.223	.068	.213	.301
15	.668	.034	–.170	.430	.339
16	.754	–.098	–.238	.051	.363
17	.680	.070	.132	.371	.378
18	.738	.072	–.017	.396	.293

tor. Three of the five loadings on the Teaching Morality via Instruction specific factor exceed .36 which provides evidence that the second condition specified above is satisfied.

The absolute value of the cross-loadings for the B-ESEM model ($|\lambda_i$ = .006 — .289$|$, *M* = .009) were smaller than the cross-loadings estimated by the three-factor ESEM model ($|l_i$ =.016 — .678$|$, *M* = .211) which provides further evidence in support of the B-ESEM model.

The fit statistics reported in Column E of Table 2 indicate that the B-ESEM model fit the data better than three-factor ESEM model. The CFI and TLI indices for the B-ESEM model are higher than for any of the other models, and the RMSEA and SRMR fit indices are smaller than for any of the other models. Since the three-factor ESEM is nested within the B-ESEM model, we invoked the DIFFTEST option in *M*plus to test whether the fit of the bifactor model represented a significant improvement in model fit compared to the three-factor ESEM model. The results show that the improvement in fit was statistically significant (adjusted chi square = 35.307, *df* = 15, *p* = .0022).

Model-Based Reliability Statistics

McDonald's coefficient omega (ω) and a variant of omega which has come to be known as omega hierarchical (ω_H) are appropriate statistics for estimating reliability when the bifactor CFA model is used (Flora, 2020; Reise et al., 2013; Rodriguez, Reise, & Haviland, 2016a & 2016b, Watkins, 2017). ω and ω_H are both model-based reliability coefficients which can be computed directly from the factor loadings and residuals obtained from a bifactor CFA model. However, estimating reliability coefficients for a bifactor ESEM model is somewhat more complex than computing reliability estimates for a bifactor CFA model because the researcher must resolve two questions about whether and how the cross-loadings should be included in the calculations:

1. Should cross loadings be included or ignored in the process of estimating reliability coefficients for a bifactor ESEM model?
2. Should the observed value of negative cross-loadings be summed together with the positive loadings when comping w and w_H, or should each negative cross-loading be replaced by its absolute value prior to being included in the summation process?

To the best of our knowledge, at present the published literature on bifactor ESEM models does not provide a clear, consensus answer to either of these questions. van Zyl and ten Klooster (2022) state that "there is currently no accepted means to account for the cross-loadings in ESEM models when estimating the scales' reliability" (p. 21). Morin, et al. (2016) do not discuss issues related to estimating reliability. Consequently, they do not address these two questions.

Morin, et al. (2020) acknowledge that they are unaware of any recommendations in the published literature about how omega reliability coefficients should be calculated in the presence of cross-loadings in bifactor ESEM models. They conclude their discussion of this issue by stating:

> In practice, we currently ignore these cross-loadings in the calculation of these coefficients (ω, ω_H, and other variants of omega], but reinforce that this is simply a matter of preference and common sense, not a statistically driven guideline. Clearly, future research is needed on this topic. (p. 1068)

However, in another paper in which Morin is listed as the second author, Gillet et al. (2020) report the results of a bifactor ESEM model that includes negative cross-loadings. Their report includes estimates of ω for the general factor and omega subscale (ω_S) for each of several subscales, but they do not explain how they dealt with cross-loadings that had negative values.

In an attempt to obtain an approximation of the reliability of the 18-item scale developed in this study, we chose to report two estimates of ω and two estimates of ω_H based on different ways of considering the cross-loadings produced by the bifactor ESEM model. In one approach, we simply summed the observed values of the loadings and cross-loadings including those with positive values and those with negative values. The results are reported in the top row of Table 4. In the second analysis, we included the absolute value of each loading and cross-loading when calculating w and ω_H. The results of this approach are reported in the second row of Table 4. Then, in a third analysis, we calculated ω and ω_H based only on the targeted loadings without including any cross-loadings. In other words, we simply replaced each negative cross-loading with a zero. The results of this approach are presented in the bottom row of Table 4.

Inspection of the reliability estimates reported in Table 4 shows that all six estimates are greater than .82, and five of the six exceed .90. Comparison of the estimates in the left column of Table 4 with the corresponding estimates in the column on the right shows that the ω estimates are all higher than the ω_H estimates as should be expected. Note also that the largest and smallest estimates of ω differ by only .010 in contrast to a difference of .106 between the largest and smallest estimates of ω_H which is more than 10 times larger.

Coefficient omega (ω) represents the proportion of variance in total scores that can be attributed to all modeled sources of common variance in the model including the general factor and all three specific factors. The three estimates of w reported in Table 4 indicate that

TABLE 4
Omega and Omega Hierarchical Reliability Coefficients Computed Three Different Ways

How Cross-Loadings Were Treated in Computing the Reliability Coefficients	*Model-Based Reliability Coefficients*	
	Omega Hierarchical (ω_H)	*Omega (ω)*
• The observed value (whether positive or negative) of each loading and each modeled cross-loading were summed together.	.934	.959
• The absolute value of each loading and each modeled cross-loading were summed together.	.828	.964
• All cross-loadings were ignored. Only the loadings on targeted factors were summed.	.925	.954

at least 95% of the common variance in the teachers' responses can be attributed to the general factor and the three specific factors combined.

Omega hierarchical (ω_H) is a special case of coefficient omega that is particularly relevant to a bifactor model because the size of w_H indicates the proportion of variance in the total scores that can be explained solely by the general factor. The conservative estimate of ω_H that results when the absolute values of the cross-loadings are included indicates that 82% of the variance in the total scores can be attributed to individual differences on the general factor.

Rodriguez et al. (2016b) claim that it is "critical" (p. 141) to compare omega$_H$ to omega by computing ratio of ω_H divided by ω. In this case, a conservative estimate of that ratio is .828/.964 = .859. This means that at least 85% of the reliable variance in the total scores can be attributed to the general factor.

DISCUSSION

In summary, the three-factor CFA model fit the data reasonably well, but the three factors were highly correlated indicating the need to search for another model that might help us understand whether and to what extent these observed correlations were inflated due to all cross-loadings being constrained to zero. If separate subscores were reported for each respondent for each of these three highly correlated factors, those scores would lack discriminant validity and users would experience difficulty in their attempts to conceptually distinguish between the meaning of the three scores.

The one-factor model did not fit the data as well the three-factor CFA model, so the notion that the 18 items all measured a single, latent variable did not provide an adequate account of the construct-relevant variability in the data or the high correlations among the three factors. The three-factor ESEM model fit significantly better than the three-factor ICM-CFA model indicating the presence of construct-relevant multidimensionality (Morin et al., 2016) due to cross-loadings. The results indicated that including freely estimated cross-loadings reduced the correlations among each of the three factor pairs.

The bifactor ESEM model fit the data better than all the other models. The general factor was well defined by the loadings on this factor, and the relative size of the various omega statistics indicated that the general factor was the dominant source of reliable variance. Coefficient omega-hierarchical indicated that this global factor accounted for 96% percent of the reliable variance. The bifactor ESEM model provides evidence that the data includes construct-relevant multidimensionality due to both (a) cross-loadings on unintended factors and (b) the existence of a global factor that is common to all of the items (Morin, et al.,

2016) Using an ESEM approach allowed cross-loadings to be estimated among items designed to assess conceptually-adjacent traits and had the desirable effect of substantially reducing the inflated inter-factor correlations obtained from the three-factor ICM-CFA model. The B-ESEM model resulted in the same corrected estimates of the inter-factor correlations but had the added advantage of accounting for construct-relevant variance due to the presence of specific factors above and beyond what is accounted for by the global factor that accounts for the shared variance among all 18 items.

These initial findings provide evidence the 18-item scale can be used to obtain reliable measures of preservice teachers' self-efficacy toward the moral work of teaching. Using bifactor ESEM to model the underlying structure of the data not only shows the presence of a dominant global factor but also accounts for true variance among the items that is above and beyond that which is explained by the global factor. Importantly, this 18-item scale also calls attention to the interconnectedness of the moral work of teaching. The one-factor model did not fit the data as well as the bifactor ESEM model, suggesting that measuring teacher self-efficacy in the moral work of teaching is more nuanced than separately measuring teacher self-efficacy for teaching morally, teaching morality via instruction, or teaching morality via modeling. Put another way, the possible subscales in this study (teaching morally, teaching morality via instruction, and teaching morality via modeling) do not have discriminant validity, pointing up the possibility that teacher self-efficacy related to any one factor is enhanced and better explained by self-efficacy in other factors related to this broader notion of the moral work of teaching. Thus validated scales that explore a single factor, such as character development or moral education (see Milson & Mehlig, 2002; Narvaez et al., 2008) are not a substitute for understanding teacher self-efficacy in the moral work of teaching (nor were they intended to do so). In this way, the current study highlights the complexity of understanding the teachers' moral role in the classroom.

Limitations

One important limitation of this study is the small sample size and the composition of the sample—primarily young, White, women with varying levels of coursework. Ideally, we would have randomly divided the sample into two subsamples and then performed the various analyses in one of the two subsamples prior to checking to see if similar results would be obtained in the other subsample. Because of the small sample size ($n = 152$) we elected not to perform this cross-validation check. Consequently, the results of this study should be considered tentative and subject to further research. A second limitation is the fact that the respondents were all preservice teachers which prevents generalization to in-service teachers. Moreover, the questions are connected to classroom teaching and not preservice experiences (which is not uncommon for self-efficacy instruments). Hence, the study needs to be replicated by administering the scale to a larger and more diverse sample that includes preservice teachers plus in-service teachers with varying levels of teaching experience. Although the results of the factor analysis reported here provides initial evidence of structural validity, further studies should investigate questions related to criterion-related validity. Moreover, evidence of content validity might be enhanced by considering additional factors related to teacher self-efficacy that might not be fully captured by the conceptions of teaching morally and teaching morality included in this study.

This conclusion suggests the need for further exploration of the initial conception of the moral work of teaching. That is, there may be additional factors or traits or virtues that contribute to self-efficacy in this domain. For example, teachers' self-efficacy in their ability to conduct and interpret assessments of students' progress and achievement in a moral way is a needed area of study. Although assessment tasks might be assumed in some of the questions related to teaching morality, the current instantiation of the scale might not fully capture teacher self-efficacy to assess the moral functioning and development of students. Similarly, the current instrument might not address teacher self-efficacy related to improving and developing the ability to teach morally and to teach morality. Moreover, teacher self-efficacy toward additional classroom practices that signal additional virtues and that were not explored in the current instrument (e.g., loyalty and social responsibility), should be further explored.

Finally, we are indebted to an anonymous reviewer for pointing out that the reliability coefficient proposed by Green and Yang (2009) and subsequently labeled "categorical omega" by Kelley and Pornprasertmanit (2016) may have been more appropriate than McDonald's omega for estimating reliability in the context of this study. Categorical omega is a variation of McDonald's omega which allows for a nonlinear rather than a linear relationship between the factors and item scores. It also allows for thresholds to differ across items.

Yang and Xia (2019) demonstrated how to compute categorical omega using Bayesian estimation methods with Mplus and SAS software. When we tailored their syntax to estimate categorical omega for a bifactor ICM-CFA model that fit our data, we obtained .943 as the resulting reliability estimate. However, this model does not account for cross-loadings of items on nontargeted factors. We have not been able to find any examples in the published literature about how to compute categorical omega for a bifactor ESEM model which is the model that best fit our data.

REFERENCES

Armor, D., Conroy-Oseguera, P., Cox, M., King, N., McDonnell, L., Pascal, A., Pauly, E., & Zellman, G. (1976). Analysis of the school preferred reading programs in selected Los Angeles minority schools, Report No. R-2007LAUSD. Rand Corporation.

Arthur, J. (2008). Traditional approaches to character education in Britain and America. In L. Nucci & D. Narvaez (Eds.), *Handbook of moral and character education* (pp. 88–98). Routledge.

Ashton, P. T., Buhr, D., & Crocker, L. (1984). Teachers' sense of efficacy: A self- or norm-referenced construct? *Florida Journal of Educational Research, 26*(1), 29–41.

Ashton, P. T., Olejnik, S., Crocker, L., & McAuliffe, M. (1982). *Measurement problems in the study of teachers' sense of efficacy* (Presentation). American Educational Research Association, New York.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397–438. https://doi.org/10.1080/10705510903008204

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Worth.

Bergem, T. (1992). Effective and responsible teaching: The new synthesis. In F. K. Oser, A. Dick & J.-L. Patry (Eds.), *Teaching the art of living: Lessons learned from a study of teacher education* (pp. 349–364). Jossey-Bass.

Berkowitz, M. W. (2002). *The science of character education*. In W. Damon (Ed.), *Bringing in a new era in character education* (pp. 43–63). Hoover Press.

Brighouse, H. (2006). *On education*. Routledge.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford.

Campbell, E. (1997). Administrators' decisions and teachers' ethical dilemmas: implications for moral agency. *Leading and Managing*, *3*(4), 245–257.

Coladarci, T., & Breton, W. (1997). Teacher efficacy, supervision, and the special education resource-room teacher. *Journal of Educational Research, 90*, 230–239. https://doi.org/10.1080/00220671.1997.10544577

Cooke, D. J., & Sellbom, M. (2019). An examination of Psychopathy Checklist-Revised latent factor structure via exploratory structural equation modeling. *Psychological Assessment, 31*(5), 581–591.

Darwall, S. (1998). *Philosophical ethics*. Westview Press.

de Beer, L. T., & van Zyl, L. E. (2019). ESEM code Generator for Mplus. http://www.survey-host.co.za/esem/

Dewey, J. (1916). Democracy and education by John Dewey. *Project Gutenberg*.

Duffin, L. C., French, B. F., & Patrick, H. (2012). The Teachers' Sense of Efficacy Scale: Confirming the factor structure with beginning pre-service teachers. *Teaching and Teacher Education*, *28*(6), 827–834. https://doi.org/10.1016/j.tate.2012.03.004

Emmer, E. (1990). *A scale for measuring teacher efficacy in classroom management and discipline* (Presentation). American Educational Research Association, Boston, MA.

Fenstermacher, G. (1990). Some moral considerations on teaching as a profession. In J. Goodlad, R. Soder, & K. A. Sirotnik (Eds.), *The moral dimensions of teaching* (pp. 130-151). Jossey-Bass.

Fenstermacher, G. D., Osguthorpe, R. D., & Sanger, M. N. (2009). Teaching morally and teaching morality. *Teacher Education Quarterly*, 36(3), 7–19.

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science, 3*(4), 484-501. https://doi.org/10.1177%2F2515245920951747

Gibson, S., & Dembo, M. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology, 76*, 569–582. https://psycnet.apa.org/doi/10.1037/0022-0663.76.4.569

Giroux, H. A. (1988). *Schooling and the struggle for public life: Critical pedagogy in the modern age.* University of Minnesota Press.

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*(1), 155–167.

Guskey, T., & Passaro, P. (1994). Teacher efficacy: A study of construct dimensions. *American Educational Research Journal, 31*, 627–643. https://doi.org/10.3102%2F00028312031003627

Hansen, D. (2001). *Teaching as a moral activity*. In V. Richardson (Ed.), *Handbook of research on teaching* (pp. 826–857). American Educational Research Association.

Hansen, D. T. (2002). The moral environment in an inner-city boys' high school. *Teaching and Teacher Education*, *18*(2), 183–204. https://doi.org/10.1016/S0742-051X(01)00063-4

Henson, R. K. (2002). From adolescent angst to adulthood: Substantive implications and measurement dilemmas in the development of teacher efficacy research. *Educational Psychologist*, *37*, 137–150. https://doi.org/10.1207/S15326985EP3703_1

Jackson, P. W., Boostrom, R. E., & Hansen, D. T. (1993). *The moral life of schools*. Jossey-Bass.

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods, 21*(1), 69–92.

Mahony, P. (2009). Should ought be taught? *Teaching & Teacher Education, 25*, 983–989. https://doi.org/10.1016/j.tate.2009.04.006

Marsh, H.W., Morin, A.J.S., Parker, P.D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*, 85-110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Marsh, H.W., Muthén, B., Asparouhov, T., Lüdtke, O, Robitzsch, A., Morin, A.J.S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16*, 439-476. https://doi.org/10.1080/10705510903008220

Maxwell, B., & Schwimmer, M. (2016). Professional ethics education for future teachers: A narrative review of the scholarly writings. *Journal of Moral Education*, *45*(3), 354–371.https://doi.org/10.1080/03057240.2016.1204271

McClellan, B. E. (1999). Moral education in America: Schools and the shaping of character from colonial times to the present. *Teachers College Press*.

Milson, A. J., & Mehlig, L.M. (2002). Elementary school teachers' sense of efficacy for character education. *The Journal of Educational Research, 96*(1), 47–53. https://doi.org/10.1080/00220670209598790

Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling, 23*, 116–119. https://doi.org/10.1080/10705511.2014.961800

Morin, A. J. S., Myers, N. D., & Lee, S. (2020). Modern factor analytic techniques: Bifactor models, exploratory structural equation modeling (ESEM), and Bifactor-ESEM. In G. Tenenbaum, R. C. Eklund, & N. Boiangin (Eds.), *Handbook of sport psychology: Exercise, methodologies, & special topics* (pp. 1044–1073). John Wiley & Sons.

Moseley, C., Reinke, K., & Bookout, V. (2003). The effect of teaching outdoor environmental education on elementary preservice teachers' self-efficacy. *Journal of Elementary Science Education*, *15*(1), 1-14. https://doi.org/10.1080/00958960209603476

Narvaez, D., & Lapsley, D. K. (2008). Teaching moral character: Two alternatives for teacher education. *The Teacher Educator*, *43*(2), 156–172. https://doi.org/10.1080/08878730701838983

Narvaez, D., Vaydich, J. L., Turner, J. C., & Khmelkov, V. (2008). Measuring teacher self-efficacy for moral education. *Journal of Research on Character Education*, (*6*)2, 3–15.

Noddings, N. (2013). *Caring: A relational approach to ethics and moral education.* University of California Press.

Oser, F. K. (1994). Chapter 2: Moral perspectives on teaching. *Review of Research in Education*, *20*(1), 57-127.

O'Sullivan, S. (2005). The soul of teaching: Educating teachers of character. *Action in Teacher Education, 26*(4), 3–9. https://doi.org/10.1080/01626620.2005.10463338

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*, 533–578. https://doi.org/10.3102%2F00346543066004543

Pendergast, D., Garvis, S., & Keogh, J. (2011). Pre-service student-teacher self-efficacy beliefs: An insight into the making of teachers. *Australian Journal of Teacher Education*, *36*(12), 4. https://doi.org/10.14221/ajte.2011v36n12.6

Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A classification and handbook*. American Psychological Association/Oxford University Press.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*(2), 129-140. https://doi.org/10.1080/00223891.2012.725437

Riggs, I., & Enochs, L. (1990). Toward the development of an elementary teacher's science teaching efficacy belief instrument. *Science Education, 74*, 625–638.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137–150. https://doi.apa.org/doi/10.1037/met0000045

Ryan, K. (1988). Teacher education and moral education. *Journal of Teacher Education*, *39*(5), 18–23. https://doi.org/10.1177%2F002248718803900505

Sanderse, W., & Cooke, S. (2021). Being prepared to become a moral teacher: UK teachers' experiences of initial teacher education. *Scandinavian Journal of Educational Research*, *65*(2), 226–239. https://doi.org/10.1080/00313831.2019.1664628

Sanger, M. N., & Osguthorpe, R. D. (2011). Teacher education, preservice teacher beliefs, and the moral work of teaching. *Teaching and Teacher Education, 27*(3), 569–578.

Sockett, H. (1993). *The moral base for teacher professionalism*. Teachers College Press.

Sockett, H., & LePage, P. (2002). The missing language of the classroom. *Teaching and Teacher Education, 18*(2), 159–171. https://doi.org/10.1016/S0742-051X(01)00061-0

Swartz, S. (2010). The pain and the promise of moral education in sub-Saharan Africa. *Journal of Moral Education*, *39*(3), 267–272. https://doi.org/10.1080/03057240.2010.497608

Thornberg, R. (2008). The lack of professional knowledge in values education. *Teaching and Teacher Education, 24*(7), 1791–1798. https://doi.org/10.1016/j.tate.2008.04.004

Toom, A., Husu, J., & Tirri, K. (2015). Cultivating student teachers' moral competencies in teaching during teacher education. In C. J. Craig & L. Orland-Barak (Eds.), *International Teacher Education: Promising Pedagogies (Part C) Advances in Research on Teaching* (Vol. 22C, pp. 13–31). Emerald Group.

Tschannen-Moran, M., Hoy, A., & Hoy, W. (1998). Teacher efficacy: its meaning and measure. *Review of Educational Research*, *68*(2), 202-248. https://doi.org/10.3102%2F00346543068002202

Tschannen-Moran, M., & Woolfolk Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*(7), 783–805. https://doi.org/10.1016/S0742-051X(01)00036-1

Tschannen-Moran, M., & Hoy, A. (2007). The differential antecedents of self-efficacy beliefs of novice and experienced teachers. *Teaching and Teacher Education, 23*, 944–956. https://doi.org/10.1016/j.tate.2006.05.003

Usher, E. L., & Pajares, F. (2008). Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research*, *78*(4), 751–796. https://doi.org/10.3102%2F0034654308321456

van Zyl, L. E., & Ten Klooster, P. M. (2022). Exploratory structural equation modelling: Practical 530 guidelines and tutorial with a convenient online tool for Mplus. *Frontiers in Psychiatry*, *531*(13), 795672 https://doi.org:0.3389/fpsyt.2021.795672

Wanxue, Q., & Hanwei, T. (2004). The social and cultural background of contemporary moral education in China. *Journal of Moral Education*, *33*(4), 465–480. https://doi.org/10.1080/0305724042000315590

Warnick, B., & Silverman, S. (2011). A framework for professional ethics courses in teacher education. *Journal of Teacher Education, 62*(3), 273–285. https://doi.org/10.1177%2F0022487110398002

Watkins, M. W. (2017). The reliability of multidimensional neuropsychological measures: From alpha to omega. *The Clinical Neuropsychologist*, *31*(6-7), 1113–1126. https://doi.org/10.1080/13854046.2017.1317364

Whitehead, A. N. (1959). The aims of education. *Daedalus*, *88*(1), 192–205. https://www.jstor.org/stable/20026487

Willemse, M., Lunenberg, M., & Korthagen, F. (2005). Values in education: A challenge for teacher educators. *Teaching and Teacher Education, 21*(2), 205–217. https://doi.org/10.1016/j.tate.2004.12.009

Woolfolk Hoy, A. W., & Burke Spero, R. B. (2005). Changes in teacher efficacy during the early years of teaching: A comparison of four measures. *Teaching and Teacher Education*, *21*(4), 343–356. https://doi.org/10.1016/j.tate.2005.01.007

Yang, Y.. & Xia, Y. (2019). Categorical omega with small sample sizes via Bayesian estimation: An alternative to frequentist estimators. *Educational and Psychologist Measurement, 79*(1), 19–39.

APPENDIX: SELF-EFFICACY FOR THE MORAL WORK OF TEACHING SCALE

Part A. This first section of the survey relates to a teacher's self-efficacy to teach morally. Rate yourself on a scale of 1–5 (1 = *not at all well*, 2 = *not very well*, 3 = *moderately well,* 4 = *quite well*, and 5 = *extremely well*) on how well you can teach in ways that might be described as morally good.

1. How well can you provide equitable learning opportunities for all students?
2. How well can you develop appropriate caring relationships with your students?
3. To what extent are you able to persist in helping struggling students learn something that is challenging for them?
4. How well can you teach in open minded ways that respect different points of view?
5. To what extent are you able to each in ways that are hopeful in the face of discouraging events?
6. How well can you balance honesty with compassion in your assessment of students?

Part B. This next section relates to a teacher's self-efficacy to teach morality, via modeling. Rate yourself on a scale of 1–5 (1 = *not at all well*, 2 = *not very well*, 3 = *moderately well*, 4 = *quite well*, and 5 = *extremely well*) on how well you can teach in ways that successfully demonstrate for students how to be morally good.

7. How well can you demonstrate fairness to students in a way that successfully helps them to be fair?
8. How well can you demonstrate compassion for your students in a way that suc-

cessfully helps them to be compassionate?

9. How well can you demonstrate persistence to struggling students in a way that helps them become more persistent?
10. To what extent can you demonstrate respect for different points of view in a way that successfully helps your students to be more open-minded toward points of view different from their own?
11. To what extent can you demonstrate hopefulness in a way that successfully helps your students become more hopeful?
12. How well do you typically demonstrate honesty in your teaching in ways that successfully helps your students refrain from cheating?

Part C. This final section relates to a teacher's self-efficacy to teach morality via instruction. Please rate yourself on a scale of 1–5 (1 = *not at all well*, 2 = *not very well*, 3 = *moderately well*, 4 = *quite well*, and 5 = *extremely well*) on how well you can teach lessons to students that successfully instill moral goodness.

13. To what extent can you instruct students in a way that helps them effectively learn to be fair in their interactions with others?
14. How well can you instruct students in a way that effectively helps them to care for other classmates?
15. How well can you instruct students in a way that successfully helps them become more persistent?
16. How well can you instruct students in a way that successfully helps them to thoughtfully consider and evaluate other people's points of view before accepting or rejecting them?
17. How well can you instruct students to be more hopeful in a way that results in students becoming more hopeful?
18. How well can you instruct students in a way that helps them successfully learn to be honest?