Validation of the Data-Driven Decision-Making Efficacy and Anxiety Inventory (3D-MEA) with U.S. Pre-Service Teachers

Todd D. Reeves
Yasemin Onder
Beheshteh Abdi
Northern Illinois University

Sound measurement of teacher self-efficacy and anxiety surrounding data-driven decision making is crucial in both theory-building and efficacy studies, as well as in practical contexts. The present study (N=457) thusly examined the validity of inferences drawn from the Data-Driven Decision-Making Efficacy and Anxiety Inventory (3D-MEA; Dunn et al., 2013) among U.S. pre-service teachers, a population with whom this instrument has never been fully and rigorously validated. Findings indicated a good fit of the hypothesized five-factor confirmatory factor model and reliable 3D-MEA scores in this population. The reported validity and reliability evidence implies that the 3D-MEA, originally intended for in-service teachers, may too be used meaningfully with U.S. pre-service teachers.

Educators must make countless decisions every day concerning their instructional practice. These include decisions about the scope and sequence of instruction, instructional methods and materials, and the needs of individual students (Hamilton et al., 2009). To aid in these often-complex decisions, educators may deploy what is known as data-driven decision making (DDDM). DDDM involves teachers "systematically collecting and analyzing various types of data, including demographic, administrative, process, perceptual, and achievement data, to guide a range of decisions to help improve the success of students" (Hamilton et al., 2009, p. 46). DDDM is theorized as a process of collecting, analyzing, interpreting, and acting on data, which relies on a diverse body of teacher knowledge termed data literacy (Mandinach & Gummer, 2016).

Despite the popularity of DDDM and rich data in many schools, many teachers lack self-efficacy concerning their capacity to work with data (Dunn et al., 2013b). Grounded in the work of Bandura's (1977) social cognitive theory, Dunn et al. defined DDDM self-efficacy as "teachers' beliefs in their abilities to organize and execute the necessary courses of action to successfully engage in classroom-level DDDM to enhance student performance" (p. 88). Teachers may lack self-efficacy concerning various facets of DDDM, such as data interpretation and the specification of instructional decisions based on data (Pierce et al., 2013; Reeves et al., 2016). At the same time, teacher anxiety—defined by Dunn et al. (2013b) as "the trepidation, tension, and apprehension teachers feel related to their ability to successfully engage in DDDM" (p. 90)—may pose a barrier to DDDM implementation (Jimerson et al., 2015; Onwuegbuzie, 2004). Problematic levels of self-efficacy and anxiety surrounding DDDM have been demonstrated in not only in-service teachers but also pre-service teachers as well (Dunn, 2016; Piro et al., 2014; Volante & Fazio, 2007).

Given theoretical and empirical evidence for their relations with teacher DDDM behavior, psycho-social factors such as self-efficacy and anxiety have become important constructs in research on teacher DDDM (Prenger & Schildkamp, 2018). Defensible measurement of these constructs in relevant populations is thus essential for descriptive, associational, and/or intervention research on DDDM. Along these same lines, sound measurement of such constructs may advance DDDM initiatives in practice. For instance, quality instrumentation may allow practitioners to well estimate the distribution of such factors for the purposes of identifying barriers to DDDM and/or designing targeted solutions to support its enactment.

Accordingly, the present study examines the validity of inferences drawn from the Data-Driven Decision-Making Efficacy and Anxiety Inventory (3D-MEA) (Dunn et al., 2013b) among U.S. pre-service teachers. While the 3D-MEA was developed for use with in-service teachers, and it has been independently validated with this population (Walker et al., 2018), considerably less is known about its psychometric properties among pre-service teachers. Crucially, several 3D-MEA items which reference the respondent's "district" may complicate the item response process for respondents whom have no district affiliation, which would include pre-service teachers. In this study, we offer in particular one form of validity evidence—validity evidence based on internal structure—that may be included in an argument for the interpretation and use of 3D-MEA scores (AERA, APA, & NCME, 2014; Kane, 2016; Lavery et al., 2020). We acknowledge that a fully comprehensive validation effort would consider other sources of validity evidence as well.

Theoretical Framework and Literature Review

The 3D-MEA was designed to measure in-service teacher self-efficacy and anxiety surrounding the practice of DDDM. Dunn et al. (2013b) theorized DDDM self-efficacy as a multi-dimensional construct with four sub-domains: *self-efficacy for data identification and access*; *self-efficacy for data technology use*; *self-efficacy for data analysis and interpretation*; and *self-efficacy for application of data to instruction*. Dunn and her colleagues (2013b) defined *self-efficacy for data identification and access* as "teachers' beliefs in their abilities to identify, access, and gather dependable, high-quality student data" (p. 90); *self-efficacy for data technology use* as "teachers' beliefs in their ability to successfully utilize district and state data technology resources" (pp. 90); *self-efficacy for data analysis and interpretation* as "teachers' beliefs in their abilities to successfully analyze and interpret student data" (pp. 94-95); and *self-efficacy for application of data to instruction* as "teachers' beliefs in their abilities to successfully connect or apply their interpretation of data findings to classroom instruction in order to improve student learning" (pp. 94-95).

The 3D-MEA is intended primarily to measure each of the above four sub-domains of DDDM self-efficacy. In addition, the 3D-MEA is designed to measure in-service teacher *anxiety* related to DDDM. Dunn et al. (2013b) defined DDDM anxiety as "the trepidation, tension, and apprehension teachers feel related to their ability to successfully engage in DDDM" (p. 90). Per self-efficacy theory, the DDDM self-efficacy dimensions should be correlated with one another, but negatively correlated with DDDM anxiety (Bandura, 1988), a pattern that has borne out in prior research.

In a study with in-service teachers from the Pacific Northwest, the developers of the 3D-MEA reported good fit of a confirmatory factor model comprising the four DDDM self-efficacy dimensions and one DDDM anxiety dimension (Dunn et al., 2013b). The DDDM self-efficacy dimensions were positively correlated with one another, and each DDDM self-efficacy dimension was negatively correlated with DDDM anxiety. Its developers also reported 3D-MEA score reliability estimates ranging from .81 to .92 in this in-service teacher population. A subsequent, independent study with in-service teachers in Illinois found high score reliability and confirmed both a good fit of the 5-factor model, as well as the earlier-reported pattern of inter-factor correlations (Walker et al., 2018).

Four other published studies have used the 3D-MEA with *in-service* teachers and offer relevant validity evidence. In two studies, Dunn et al. (2013a, 2013b) found that the 3D-MEA measures were expectedly related to teacher concerns about DDDM in large samples of in-service teachers. In a correlational study of in-service teachers, Reeves et al. (2016) linked self-efficacy with in-service teacher-reported implementation of data use practices. Finally, in a small study with 15 in-service teachers, Green et al. (2016) did not explicitly address reliability or validity issues, but did find that 3D-MEA measures changed during a professional development intervention. Overall, this body of research well supports the reliability and validity of the information provided by the 3D-MEA when used with in-service teachers.

After a review of the literature, we identified only three studies that employed the 3D-MEA with *pre-service* teachers, all of which were conducted by our research group. All three were small-scale intervention studies which used the 3D-MEA as an outcome measure (Reeves & Chiang, 2017, 2018, 2019), and reported limited reliability and validity evidence. As the present study with 457 pre-service teachers includes data from the 207 pre-service teachers represented in those earlier studies, we do not take time here to discuss in detail the reliability and dimensionality evidence available from these earlier studies. However, we note summarily that the aforementioned studies contributed unique validity evidence that 3D-MEA measures change in response to intervention, and are related to relevant criterion measures, among pre-service teachers.

Substantive analysis of the 3D-MEA items indicates possible challenges associated with the use of the measure with pre-service teachers. In particular, all three *self-efficacy for data technology use* items and one of five *anxiety* items reference the respondent's (school) "district." For example, one of the *self-efficacy for data technology use* items reads, "I am confident I can use the tools provided by my district's data technology system to retrieve charts, tables or graphs for analysis". Such references to a district may complicate item response processes and introduce measurement error for pre-service teachers who lack a district affiliation. While the evidence available to date is promising, there remains yet to be a rigorous psychometric evaluation of the 3D-MEA in a large, heterogeneous sample of pre-service teachers.

Method

Participants

Our study relied on data collected during several published and unpublished studies of pre-service teacher data use in the U.S, as well as data collected specifically for this validation study. In the existing studies, 3D-MEA scores were used either as outcome measures (for intervention studies) or explanatory variables (for other studies). The new data were collected from pre-service elementary teachers at one institution who received extra credit in an assessment course for participation in 3D-MEA validation research over multiple semesters. All data collection relied on online survey instruments administered via Qualtrics survey software. For those data gathered during pretest-posttest intervention studies, only pretest 3D-MEA responses were retained when both pretest and posttest data were available. In total, 3D-MEA and available basic demographic data from seven data files were merged for this study.

The analytic sample comprised U.S. pre-service teachers (N=457) from at least 38 states and 132 pre-service institutions who consented to participate in the study. The majority of the participants were from Illinois, New York, Texas, Florida, Pennsylvania, New Jersey, and Michigan. Based on available data, the sample was 91.46% female, 87.27% white, and 9.57% Hispanic/Latino. With a mean age of 24.89 (SD=6.78), the pre-service participants intended to work in a variety of PK-12 grades and subjects (though a large majority of the sample, 86.12%, intended to be K-5 teachers).

Instrumentation

As discussed, the 3D-MEA instrument's 20 items are intended to measure four DDDM self-efficacy dimensions and one DDDM anxiety dimension. Three items are intended to measure *self-efficacy for data identification and access*, another three items *self-efficacy for data technology use*, three items *self-efficacy for data analysis and interpretation*, six items *self-efficacy for the application of data to instruction*, and five items *anxiety*. All items are shown in Table 1. The response format for these 20 items was a 5-point rating scale: 1=*Strongly disagree,* 2=*Disagree,* 3=*Neither disagree nor disagree,* 4=*Agree,* and 5=*Strongly agree.*

Analytic Approach

To examine the 3D-MEA's internal score structure as evidence of validity with pre-service teachers, we used confirmatory factor analysis (CFA) via the lavaan package in R (Rosseel, 2012). To avoid overfitting of the model to these data (e.g., capitalization on chance), we also performed cross-validation to provide evidence of the stability of our findings. In particular, we sub-divided our full sample of 457 persons into two random subsets, and formally tested the invariance of model parameters across these to sub-samples using multi-group CFA (Meredith, 1993). This involved comparing the fit of successive multi-group CFA analyses in which parameters were increasingly constrained equal across groups. We relied on both likelihood-ratio tests of goodness-of-fit and changes in the comparative fit index in evaluating invariance. Modification indices were also examined for the development dataset to determine if fit could be

improved through model re-specification. However, no such changes were ultimately made owing to the apparent strong evidence of model fit in the development sample.

In all CFA analyses, the indicator variables were treated as continuous (Rhemtulla et al., 2012); a supplemental analysis with the items treated as ordered categories had very similar findings.[1] Given evidence of non-normality, we relied on robust maximum likelihood estimation with the Satorra-Bentler scaled chi-square test statistic ("MLM" estimation in the lavaan package). Non-normality can inflate the chi-square statistic and deflate standard errors under traditional maximum likelihood estimation (Rhemtulla et al., 2012; Yuan & Bentler, 2000). The sample size was adequate by conventional standards (Gagné & Hancock, 2006), and there were no missing data.[2]

We relied on the chi-square goodness-of-fit test as well as relative and absolute fit indices (Hu & Bentler, 1999), namely: the comparative fit index (CFI) (Bentler, 1990); the standardized root mean square residual (SRMR); and the root mean square error of approximation (RMSEA) (Steiger & Lind, 1980). Given the large sample size, the fit indices were emphasized relative to the results of chi-square goodness-of-fit significance tests. Fit indices were evaluated in relation to literature-defined thresholds (Hu & Bentler, 1999; Marsh et al., 2004).

Prior to confirmatory factor analysis, we also examined both univariate and multivariate normality issues, particularly skewness and kurtosis. This included formally testing univariate normality using the Shapiro-Wilke test, and multivariate skewness and kurtosis using Mardia's (1970) tests. Furthermore, we used Mahalanobis distance (*D*) measures and their statistical significance to indicate possible multivariate outliers. We also compared results with and without these outliers included. Finally, to examine the reliability of 3D-MEA's scores, we computed coefficient omegas (ω; McDonald, 2013) for all five factors. We also computed a measure of maximal reliability, coefficient *H* (Hancock & Mueller, 2001).

Results

Table 1 contains descriptive statistics for the 20 3D-MEA items. There was adequate variation in responses to all items, as well as some variation among the item means themselves. All five response categories were used for all items except for the item, "I am confident I can use

[1] The results of this supplemental analysis are available from the authors upon request.

[2] In compiling our analytic dataset, we retained from each of seven source datasets only those cases with complete 3D-MEA data. However, this was all or nearly all cases from each of the datasets, which is likely attributable to a number of factors. First, the survey software used to collect all data reminded participants at the end of each page to complete any unanswered questions. In the three published intervention studies, for example, missingness for the 3D-MEA items within a given time point was either nominal or non-existent. Across these three studies, only one individual in one study did not respond to a 3D-MEA item at one time point. Missingness tended to occur between rather than within survey sections (with individuals ceasing participation after fully completing a given block of questions, such as the 3D-MEA items). Second, for existing studies, we began with the only available version of the datasets, the anonymized, archival versions of the analytic datasets which featured no case-wise missing data. Third, some datasets included 3D-MEA data collected at two time points, and when pretest 3D-MEA data were missing we were able to use posttest 3D-MEA data instead. Fourth, and as related to the new data collection only, we speculate that respondents may have been highly motivated by the erroneous belief that they must respond to all items in order to receive the incentive (even though the consent form stated otherwise).

assessment data to identify gaps in my instructional curriculum" (*self-efficacy for the application of data to instruction*). For this one item, no individual had responded *strongly disagree*.

While all but one item was negatively skewed and all items were leptokurtotic, skewness and kurtosis values larger than 2 and 7 were not evidence (respectively). However, Shapiro-Wilk tests of univariate normality were statistically significant (p<.001) for all items. At the same time, Mardia's (1970) multivariate skewness ($b_1 = 4103.79, p < .001$) and kurtosis ($b_2 = 50.78, p < .001$) tests were both statistically significant. Maximum likelihood estimation's assumption of multivariate normality was consequently not tenable, and for this reason, we used robust maximum likelihood estimation as noted earlier (Olsson, Foss, Troye, & Howell, 2000).

Twenty-six cases (5.69%) had statistically significant (p<.001) Mahalanobis distance (D) values indicating that they were potential multivariate outliers. Unreported analyses showed that substantive results were the same for both sample sub-sets one and two with these cases removed. We therefore report results based on all cases here.

Table 1
Descriptive Statistics (N=457)

Item	Mean (M)	Std. Deviation (SD)	Skewness (S_3)	Kurtosis (K_4)
I am confident in my ability to access state assessment results for my students	3.21	1.03	-0.25	2.16
I am confident that I know what types of data or reports I need to assess group performance	3.18	0.98	-0.21	2.08
I am confident that I know what types of data or reports I need to assess student performance	3.37	0.94	-0.43	2.41
I am confident I can use the tools provided by my district's data technology system to retrieve charts, tables or graphs for analysis	3.37	0.97	-0.37	2.52
I am confident I can use the tools provided by my district's data technology system to filter students into different groups for analysis	3.32	0.97	-0.34	3.39
I am confident that I can use my district's data analysis technology to access standard reports	3.26	0.97	-0.22	2.33
I am confident in my ability to understand assessment reports	3.58	0.89	-0.62	2.86
I am confident in my ability to interpret student performance from a scaled score	3.64	0.84	-0.65	2.89
I am confident in my ability to interpret subtest or strand scores to determine student strengths and weaknesses in a content area	3.55	0.85	-0.61	2.83

I am confident that I can use data to identify students with special learning needs	3.50	0.92	-0.56	2.75
I am confident that I can use data to identify gaps in student understanding of curricular concepts	3.65	0.82	-0.77	3.30
I am confident that I can use assessment data to provide targeted feedback to students about their performance or progress	3.75	0.76	-0.78	3.80
I am confident I can use assessment data to identify gaps in my instructional curriculum	3.65	0.80	-0.63	2.93
I am confident that I can use data to group students with similar learning needs for instruction	3.83	0.75	-0.85	3.97
I am confident in my ability to use data to guide my selection of targeted interventions for gaps in student understanding	3.53	0.83	-0.52	2.63
I am intimidated by statistics	2.91	1.09	0.10	2.16
I am intimidated by the task of interpreting students' state level standardized assessments	3.05	1.01	-0.11	2.23
I am concerned that I will feel or look "dumb" when it comes to data driven decision-making	3.02	1.11	-0.11	2.11
I am intimidated by my district's data retrieval technology	2.94	0.89	-0.09	2.99
I am intimidated by the process of connecting data analysis to my instructional practice	3.03	0.96	-0.13	2.21

Note. Response format was: 1=*Strongly disagree,* 2=*Disagree,* 3=*Neither disagree nor disagree,* 4=*Agree,* and 5=*Strongly agree.*

Table 2 summarizes the fit of all CFA models estimated during the present study. This includes single-group CFA analyses fit with each of two random sub-samples, five multi-group CFA models used to evaluate model invariance across sub-samples, and, a final model including all members of the sample. As evidenced by the fit indices in Table 2, all models fit quite well in an absolute sense. As fit was good in the first random sub-sample, we did not opt to make model modifications to then be cross-validated in the second sub-sample.

However, we still deployed multi-group CFA (Meredith, 1993) to test the invariance of parameter estimates across two separate samples to bolster evidence for the stability of the solution. To compare the fit of successive nested models we relied on the likelihood-ratio test of model fit based on the adjusted chi-square statistic. However, given that chi-square tests are sensitive with large samples, we also considered changes in model fit indices (e.g., Cheung & Rensvold, 2002). In particular, we relied on changes in the comparative fit index (ΔCFI), with reductions of less than .01 considered indicative of invariance.

Invariance testing begins by fitting a multi-group CFA model in which parameters are unconstrained across groups (model 0), followed by a series of multi-group CFA models with increasing group-equality constraints. These subsequent models feature equal factor loadings across groups (model 1); equal factor loadings and item intercepts across groups (model 2); equal factor loadings, item intercepts, and residual item variances/covariances across groups (model 3); and, finally, equal item factor loadings, item intercepts, residual item variances/covariances, and factor variances/covariances across groups (model 4; Dimitrov, 2010). While models 0 through 3 are used to test measurement invariance per se, model 4 is used to evaluate structural equivalence.

The multi-group CFA analyses indicated strict measurement invariance across sub-samples (and furthermore structural invariance). While the likelihood ratio test was statistically significant when comparing model 2 and model 3, possibly indicating a violation of strict measurement invariance, the corresponding ΔCFI was ignorably small. Given that the chi-square test is sensitive to large sample sizes, we accordingly emphasized the negligible decrease in the CFI in interpreting our comparison of models 2 and 3. These cross-validation findings provide evidence for the stability of our finding of the fit of the hypothesized five-factor CFA model. In other words, we are confident that the observed model-data fit is not due to the overfitting of a model to our overall (*N*=457) sample.

In light of these invariance findings, we report our substantive findings based on a CFA conducted with our full sample. Figure 1 provides a graphical representation of the five-factor model estimated for all data (*N*=457). The chi-square test of fit was statistically significant, $\chi^2_{SB}(160,\ N = 457) = 316.81, p < .001$. However, the chi-square method of fit testing is known to be sensitive to large sample sizes, so we relied on other measures of model fit for goodness-of-fit evaluation. The robust comparative fit index values—.96 for the Satorra-Bentler-adjusted robust CFI (2001) and .96 for the Brosseau-Liard et al.-adjusted robust CFI (2012)—imply good model fit (Bentler & Bonett, 1980; Hu & Bentler, 1999; Schumacker & Lomax, 1996).

At the same time, the robust root mean square error of approximation (RMSEA) values implied good fit of our full model with all respondents, with point estimates for these indices less than .06 (Hu & Bentler, 1999; Marsch et al., 2004): .05 for the Satorra-Bentler-adjusted robust RMSEA (90% confidence interval [0.04, 0.05]); and .05 for the Brosseau-Liard and Savalei-adjusted robust RMSEA (90% confidence interval [0.05, 0.06]). Finally, the standardized root mean square residual (SRMR=.05) value less than .08 indicated good fit of our final, full model (Hu & Bentler, 1999).

Table 2
Summary of Fit for All Confirmatory Factor Models

CFA model	χ^2[a]	*df* χ^2	$\Delta\chi^2$[a]	*df* $\Delta\chi^2$	RMSEA [90% CI][b]	RMSEA [90% CI][c]	CFI[b]	ΔCFI[b]	CFI[c]	ΔCFI[c]	SRMR
Single-group	-	-	-	-	-	-	-	-	-	-	-
Sample 1 (*n*=228)	246.85***	160	-	-	.049 [.038, .059]	.056 [.042, .069]	.950	-	.952	-	.052
Sample 2 (*n*=229)	232.98***	160	-	-	.045 [.034, .055]	0.53 [.037, .067]	.964		.968	-	.053
Multi-group (*N*=457)	-	-	-	-	-	-	-	-	-	-	-
Model 0	479.43***	320	-	-	.047 [.039, .054]	.054 [.044, .064]	.960	-	.959	-	.050
Model 1	495.87***	335	15.30	15	.046 [.038, .053]	.053 [.043, .063]	.960	.000	.959	.000	.055
Model 2	512.37***	350	13.88	15	.045 [.038, .052]	.052 [.042, .061]	.959	-.001	.959	.000	.056
Model 3	546.73***	370	33.37*	20	.046 [.039, .053]	.053 [.043, .062]	.956	-.003	.955	-.004	.055
Model 4	559.97***	385	12.07	15	.045 [.037, .051]	.052 [.042, .061]	.956	-.000	.955	-.000	.068
Final model *(N*=457)	316.81***	160	-	-	.046 [.040, .053]	.054 [.045, .063]	.960	-	.959	-	.046

Note. CFA=confirmatory factor analysis. Model 0 = parameters unconstrained across groups. Model 1 = factor loadings constrained equal across groups. Model 2 = factor loadings and item intercepts constrained equal across groups. Model 3 = factor loadings, item intercepts, and residual item variances/covariances constrained equal across groups. Model 4 = item factor loadings, item intercepts, residual item variances/covariances, and factor variances/covariances constrained equal across groups. Individual model chi-square and chi-square difference statistics are robust, using the Satorra-Bentler (2001) adjustment.

[a]Chi-square statistic adjusted for non-normality using Satorra-Bentler (2001) method.

[b]Calculated using the Satorra-Bentler-adjusted (2001) chi-square statistic.

[c]Calculated using the methodology outlined in Brosseau-Liard et al. (2012) and/or Brosseau-Liard and Savalei (2014).

*p<.05, **p<.01, ***p<.001

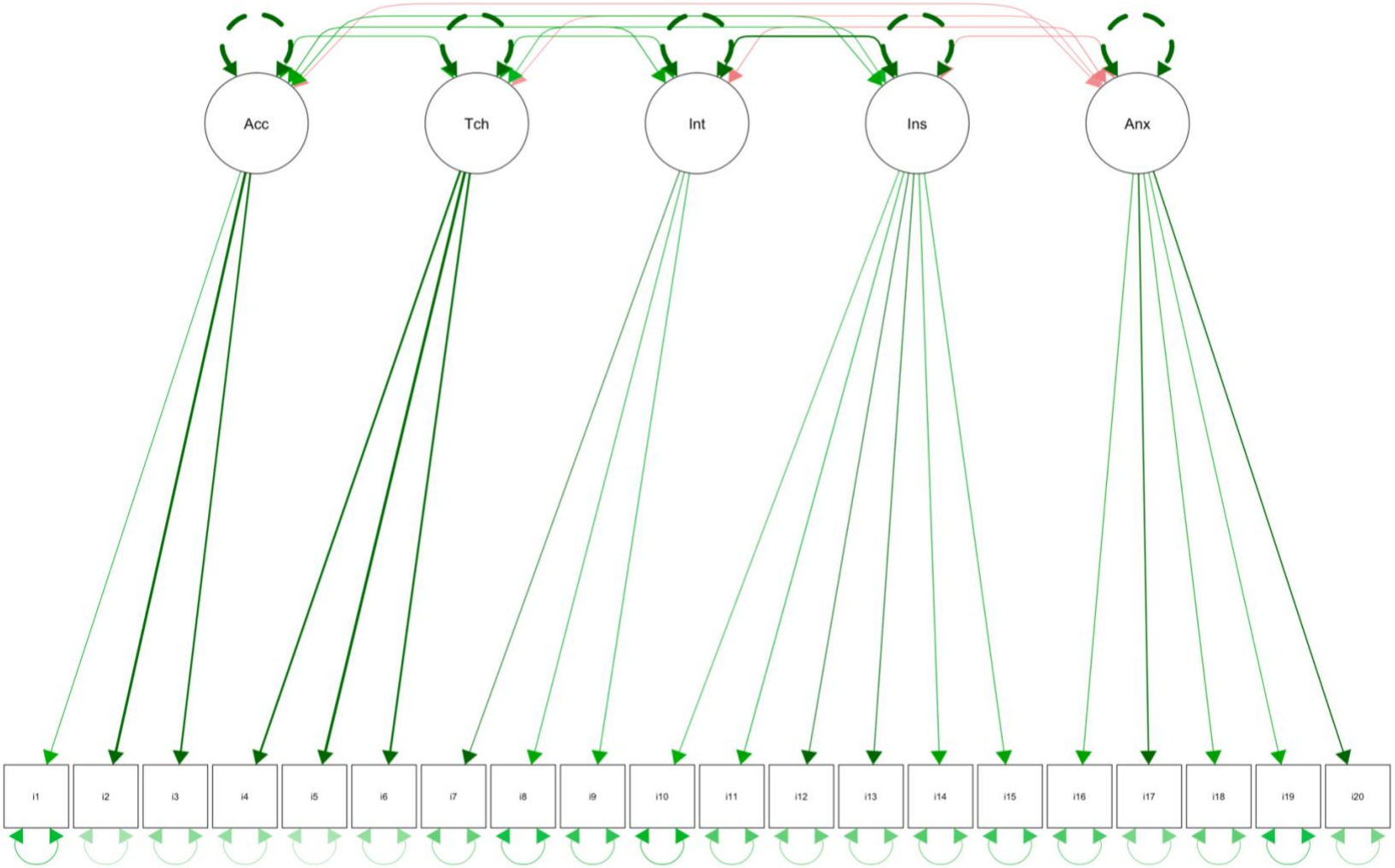

Figure 1. Graphical representation of the five-factor model estimated for all data (*N*=457). The confirmatory factor model featured five inter-correlated latent variables, each inferred by between three to six indicator variables. Standardized factor loadings, factor correlations, and communalities (and their complements, error variances) for the model are tabulated or reported elsewhere in the manuscript. Color indicates the nature of the relationship (green indicates a positive relationship and red indicates a negative relationship). Thickness of line indicates the magnitude of the estimate (thicker or darker lines indicate larger estimates, and thinner or lighter lines indicate smaller estimates).

As shown in Table 3, the standardized factor loadings ranged from 0.64 to 0.88. All of the factor loadings were statistically different from zero (*p*s<.001) and exceeded the item salience threshold of .5 (Kline, 2014). In addition, 90% of the factor loadings were ≥ .70 in absolute value (Hair et al., 1998). The lowest standardized factor loadings were for the items: "I am confident in my ability to access state assessment results for my students" (.66) from the *self-efficacy for data identification and access* scale and "I am confident that I can use data to identify students with special learning needs" (.64) from the *self-efficacy for the application of data to instruction* scale. Communality indices (h^2) for the manifest variables ranged from 0.41 to 0.78, and 80% were larger than .50, indicating that the latent variables well account for variation in all items.

Table 3
Standardized Factor Loadings

Self-efficacy for data identification and access	-
I am confident in my ability to access state assessment results for my students	0.66
I am confident that I know what types of data or reports I need to assess group performance	0.88
I am confident that I know what types of data or reports I need to assess student performance	0.84
Self-efficacy for data technology use	-
I am confident I can use the tools provided by my district's data technology system to retrieve charts, tables or graphs for analysis	0.85
I am confident I can use the tools provided by my district's data technology system to filter students into different groups for analysis	0.88
I am confident that I can use my district's data analysis technology to access standard reports	0.85
Self-efficacy for data analysis and interpretation	-
I am confident in my ability to understand assessment reports	0.78
I am confident in my ability to interpret student performance from a scaled score	0.70
I am confident in my ability to interpret subtest or strand scores to determine student strengths and weaknesses in a content area	0.72
Self-efficacy for the application of data to instruction	-
I am confident that I can use data to identify students with special learning needs	0.64
I am confident that I can use data to identify gaps in student understanding of curricular concepts	0.76
I am confident that I can use assessment data to provide targeted feedback to students about their performance or progress	0.79
I am confident I can use assessment data to identify gaps in my instructional curriculum	0.79
I am confident that I can use data to group students with similar learning needs for instruction	0.76
I am confident in my ability to use data to guide my selection of targeted interventions for gaps in student understanding	0.73
Anxiety	-
I am intimidated by statistics	0.76

I am intimidated by the task of interpreting students' state level standardized assessments	0.81
I am concerned that I will feel or look "dumb" when it comes to data driven decision-making	0.78
I am intimidated by my district's data retrieval technology	0.70
I am intimidated by the process of connecting data analysis to my instructional practice	0.81

Note. All loadings were statistically significant at $p<.001$.

All five latent variables were statistically related to one another (all *p*s<.001), with correlations ranging between .31 (*self-efficacy for data technology use* and *anxiety*) and .83 (*self-efficacy for data analysis and interpretation* and *self-efficacy for the application of data to instruction*) in absolute magnitude. All factor correlations are shown in Table 4. The four self-efficacy dimensions were positively correlated with one another, and each was negatively correlated with the anxiety dimension as consistent with theory and prior research. A very strong correlation between *self-efficacy for data analysis and interpretation* and *self-efficacy for the application of data to instruction* has also been observed in prior research, though theory concerning the DDDM process implies these factors are indeed distinct (Mandinach & Gummer, 2016).

Table 4
Factor Correlations

Factor	Access	Technology	Interpretation	Instruction	Anxiety
Access	1.00	.68	.75	.70	-.33
Technology	-	1.00	.70	.60	-.31
Interpretation	-	-	1.00	.83	-.38
Instruction	-	-	-	1.00	-.37
Anxiety	-	-	-	-	1.00

Note. Access=self-efficacy for data identification and access. Technology=self-efficacy for data technology use. Interpretation=self-efficacy for data analysis and interpretation. Instruction=self-efficacy for the application of data to instruction.

While the fit was good, we also examined modification indices (MIs) for the final model (N=457). The largest MIs were for error covariances for two almost-identically worded and adjacent items (items 2-3) for the *self-efficacy for data identification and access* factor, and two almost-identically worded and adjacent items (items 8-9) for the *self-efficacy for data analysis and interpretation* factor. The MIs also indicated cross-loadings for three items and *the self-efficacy for data technology use* factor, but content overlap among these three items was less obvious. These may indicate local dependence or multidimensionality issues or may be false positives. The decision to make model modifications involves important trade-offs between goodness-of-fit on the one hand, and stability, parsimony, theoretical soundness on the other. As fit was good and a strong theoretically-coherent dimensional structure was evident, we did not opt to make any model modifications (as with the development sample).

Reliability estimates (ω) for the factors were as follows: self-efficacy for data identification and access (.83), self-efficacy for data technology use (.89), self-efficacy for data analysis and interpretation (.78), self-efficacy for the application of data to instruction (.88), and anxiety (.88). The upper-bound estimate of reliability, Hancock's Coefficient *H*, was .93, which was certainly also acceptable (Hancock & Mueller, 2001).

Discussion

Years on, research continues to examine data-driven decision making (DDDM) as a mechanism by which to optimize teaching and learning processes, and in turn improve student outcomes. These efforts require high-quality instrumentation to measure important antecedents of DDDM, such as teacher self-efficacy and anxiety, as well as the effects of teacher DDDM interventions. Proper measurement of such constructs in key populations is also essential for practical purposes (e.g., identifying barriers to DDDM, needs assessment). Information concerning the score reliability and validity of relevant instruments when used with critical populations, including the 3D-MEA, is therefore of considerable value to the field.

The structural validity evidence reported here is favorable indeed. We find strong evidence for the fit of a five-factor structure of the 3D-MEA in U.S. pre-service teachers. However, while validity evidence based on internal structure is necessary for test validation processes, it is not sufficient. Evidence related to an instrument's internal score structure is only one source of validity evidence. Contemporary measurement theory views test validation as a process of formulating an argument for the interpretation and use of an instrument's scores. That argument ought rely on other forms of validity evidence as well, including evidence based on test content, evidence based on response processes, evidence based on relations to other variables, and evidence based on consequences of testing (AERA, APA, & NCME, 2014; Kane, 2016; Lavery et al., 2020).

Any given individual study is unlikely to provide sufficient evidence to fully validate an instrument's scores. As such, it would be prudent for researchers to gather additional evidence in order to build a more comprehensive argument for the intended interpretations and uses of this instrument's scores. Along these lines, the reader will recall that one motivation for this study was the question of whether pre-service teachers may appropriately respond to items referencing their "district." There may be similar issues associated with other items when used with pre-service teachers as well. For instance, do pre-service teachers have an appropriate conception of a "data technology system" in order to respond to related items? While beyond the scope of this study, others may choose in the future to collect validity evidence based on response processes, such as think-alouds as pre-service teachers respond to such items. As another form of validity evidence, it may also be beneficial to test correlations among particular DDDM self-efficacy dimensions and objective measures of related teacher data literacy skills (e.g., accessing data, interpreting data).

Our findings imply from a psychometric perspective that the 3D-MEA measure may be used meaningfully with the U.S. pre-service teacher population. That said, our findings were derived from a non-probabilistic sample that is clearly non-representative of U.S. pre-service teachers nationally. These findings should, of course, be replicated using data collected from scientific

and nationally-representative samples. We must also acknowledge the limited variation in one item—the item for which *strongly disagree* was never endorsed—as a potential threat to the validity of our findings. While the expected effect of restriction of range would be attenuation of observed covariation, we note that the factor loading for this item was still quite large (.79).

These findings also bode well for use of 3D-MEA with other populations without a "district" affiliation such as in-service teachers working in private schools. Regardless, it would be interesting to see if reliability and validity improve by generalizing the content of those few items referencing "my district" or other items for which item content may pose interpretational challenges to respondents. Another valuable future direction in this realm is the formal investigation of the invariance of the 3D-MEA's factor structure between in-service and pre-service teachers. The latter would afford evidence for the comparability of findings derived from the 3D-MEA in each of these two populations, as well as evidence measurement-error free comparisons between them.

Author Notes

Todd Reeves is an Associate Professor of Educational Research and Evaluation in the Department of Educational Technology, Research and Assessment at Northern Illinois University. His research addresses problems related to teacher education, assessment, and technology.

Beheshteh Abdi is a doctoral candidate in Instructional Technology in the Department of Educational Technology, Research and Assessment at Northern Illinois university under the supervision of Dr. Todd Reeves. Her research is centered on student engagement and international large-scale data analysis in education.

Yasemin Onder is a doctoral student in Instructional Technology in the Department of Educational Technology, Research and Assessment at Northern Illinois University under the supervision of Dr. Reeves. Her research interests include statistics, research methodology, and learning analytics in education.

Correspondence concerning this article should be addressed to Todd D. Reeves, 204A Gabel Hall, Northern Illinois University, DeKalb, Illinois, 60115, United States, 1-815-753-9427, treeves@niu.edu

References

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

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215.

Bandura, A. (1988). Self-efficacy conception of anxiety. *Anxiety Research*, *1*(2), 77-98.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238-246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588-606.

Brosseau-Liard, P. E., & Savalei, V. (2014). Adjusting incremental fit indices for nonnormality. *Multivariate Behavioral Research*, *49*(5), 460-470.

Brosseau-Liard, P. E., Savalei, V., & Li, L. (2012). An investigation of the sample performance of two nonnormality corrections for RMSEA. *Multivariate Behavioral Research*, *47*(6), 904-930.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*(2), 121-149.

Dunn, K. E., (2016). Educational psychology's instructional challenge: Pre-service teacher concerns regarding classroom-level data-driven decision-making. *Psychology Learning & Teaching*, *15*(1), 31-43.

Dunn, K. E., Airola, D. T., & Garrison, M. (2013a). Concerns, knowledge, and efficacy: An application of the teacher change model to data driven decision-making professional development. *Creative Education*, *4*(10), 673.

Dunn, K. E., Airola, D. T., Lo, W. J., & Garrison, M. (2013b). Becoming data driven: The influence of teachers' sense of efficacy on concerns related to data-driven decision making. *The Journal of Experimental Education*, *81*(2), 222-241. https://doi.org/10.1080/0020973.2012.699899

Dunn, K. E., Airola, D. T., Lo, W. J., & Garrison, M. (2013). What teachers think about what they can do with data: Development and validation of the data driven decision-making efficacy and anxiety inventory. *Contemporary Educational Psychology*, *38*(1), 87-98.

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

Green, J. L., Schmitt-Wilson, S., & Versland, T. M. (2016). Teachers and data literacy: A blueprint for professional development to foster data driven decision making. *Journal of Continuing Education and Professional Development*, *3*(1), 14-32. https://doi.org/10.7726/jcepd.2016.1002

Hair, J., Anderson, R., Tatham, R., & Black, W. (1998). Multivariate data analysis. Prentice Hall.

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using student achievement data to support instructional decision making*. U.S. Department of Education.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R.Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 195–216). Scientific Software International.

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

Jimerson, J. B., Choate, M. R., & Dietz, L. K. (2015). Supporting data-informed practice among early career teachers: The role of mentors. *Leadership and Policy in Schools*, *14*(2), 204-232.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood*. National Educational Resources.

Kane, M. T. (2016). Validation strategies: delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. Raymond & T. M. Haladyna (Eds.), *Handbook of Test Development* (Vol. 2, pp. 64-80). New York, NY: Routledge.

Kline, P. (2014). *The New Psychometrics: Science, Psychology and Measurement*. Routledge.

Lavery, M. R., Kruse, L., Krupa, E., Bostic, J., & Carney, M. (2020). Argumentation surrounding argument-based validation: A systematic review of validation methodology in peer-reviewed articles. *Educational Measurement: Issues and Practice*. Advanced online publication.

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education, 60*, 366-376.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519-530.

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, 11*(3), 320-341.

McDonald, R. P. (2013). *Test theory: A unified treatment*. Psychology Press.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling, 7*(4), 557-595.

Onwuegbuzie, A. J. (2004). Academic procrastination and statistics anxiety. *Assessment & Evaluation in Higher Education, 29*, 3-19.

Pierce, R., Chick, H., & Gordon, I. (2013). Teachers' perceptions of the factors influencing their engagement with statistical reports on student achievement data. *Australian Journal of Education*, *57*(3), 237-255.

Pierce, R., Chick, H., Watson, J., Les, M., & Dalton, M. (2014). A statistical literacy hierarchy for interpreting educational system data. *Australian Journal of Education, 58*(2), 195-217.

Piro, J. S., Dunlap, K., & Shutt, T. (2014). A collaborative data chat: Teaching summative assessment data use in pre-service teacher education. *Cogent Education, 1*(1).

Prenger, R., & Schildkamp, K. (2018). Data-based decision making for teacher and student learning: a psychological perspective on the role of the teacher. *Educational Psychology*, *38*(6), 734-752.

Reeves, T. D., & Chiang, J. L. (2017). Building pre-service teacher capacity to use external assessment data: An intervention study. *The Teacher Educator*, *52*(2), 155-172.

Reeves, T. D., & Chiang, J. L. (2018). Online interventions to promote teacher data-driven decision making: Optimizing design to maximize impact. *Studies in Educational Evaluation, 59*, 256-269.

Reeves, T. D., & Chiang, J. L. (2019). Effects of an asynchronous online data literacy intervention on pre-service and in-service teachers' beliefs, self-efficacy, and practices. *Computers & Education*, *136*, 13-33.Reeves, T. D., Summers, K. H., & Grove, E. (2016). Examining the landscape of teacher learning for data use: The case of Illinois. *Cogent Education*, *3*(1).

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more (Version 0.5–12). *Journal of Statistical Software, 48*(2), 1-36.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507-514.

Schumacker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling*. Lawrence Erlbaum.

Steiger, J. H., & Lind, J. C. (1980, June). *Statistically-based tests for the number of common factors* [Paper presentation]. Annual Spring Meeting of the Psychometric Society, Iowa City, IA.

Volante, L., & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Canadian Journal of Education, 30*(3), 749-770.

Walker, D. A., Reeves, T. D., & Smith, T. J. (2018). Confirmation of the Data Driven Decision-Making Efficacy and Anxiety Inventory's (3D-MEA) score factor structure among teachers. *Journal of Psychoeducational Assessment, 36*(5), 477-491.

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonmoral missing data. *Sociological Methodology, 30*(1), 165-200. https://doi.org/10.1111/0081-1750.00078