Journal of Montessori Research
2017, Volume 3, Issue 1

Determining the Measurement Quality of a Montessori High School Teacher Evaluation Survey

Anthony Philip Setari and Kelly D. Bradley

University of Kentucky

***Keywords**: evaluation, survey, Erdkinder, high school*

Abstract. The purpose of this study was to conduct a psychometric validation of a course evaluation instrument, known as a student evaluation of teaching (SET), implemented in a Montessori high school. The authors demonstrate to the Montessori community how to rigorously examine the measurement and assessment quality of instruments used within Montessori schools. The Montessori high school community needs an SET that has been rigorously examined for measurement issues. The examined SET was developed by a Montessori high school, and the sample data were collected from Montessori high school students. Using a Rasch partial credit model, the results of the analysis identified several measurement issues, including multidimensionality, misfit items, and inappropriate item difficulty levels. A revised version of the SET underwent the same analysis procedure, and the results indicated that measurement issues persisted. The authors suggest several ways to improve the overall measurement quality of the instrument while keeping the Montessori foundation. Additional validation studies with a revised version of the SET will be needed before the instrument can be endorsed for full implementation in a Montessori setting.

The number of Montessori high schools has increased across the United States and is now approximately 121 (National Center for Montessori in the Public Sector, 2017). Building on the popularity of Montessori education experiences in the early grades, Montessori high schools offer students with Montessori backgrounds an opportunity to continue their experience at the high school level. Students new to Montessori education also have the opportunity for a nontraditional high school experience. Underlying much of the Montessori high school philosophy is the principle that students have multifaceted cognitive, social, emotional, and moral experiences. These experiences provide the enrichment that facilitates the Montessori valorization process, explained well as "the process of becoming a strong and worthy person" (Donahoe, Cichuki, Coad-Bernard, Coe, & Scholtz, 2013, p. 18; Mayes & Williams, 2013), which is the primary intent of a Montessori adolescent education (R. Miller, 1990; J. P. Miller, 2010).

The role of the Montessori high school teacher in the valorization process is critical although poorly defined (Barker, 2011; Montessori, 1973). Unlike the depth of detail that Maria Montessori provided for education at the Early Childhood and Elementary levels, specifics on the Montessori high school experience are comparatively lacking (Barker, 2011). Further complicating the issue is that unique programs on Montessori adolescent pedagogy are offered by a range of institutions; some of the most recognizable include the Cincinnati Montessori Secondary Teacher Education Program, the Hershey Montessori School, the Houston Montessori Center, and the Montessori High School at University Circle in Cleveland. This ambiguity, along with the variety of Montessori adolescent education methods, has led to much uncertainty

Published online May 2017

about how to establish a true Montessori adolescent experience, as Dr. Montessori would have envisioned it, and to the development of highly variant Montessori high schools (Barker, 2011; Kahn, 2011; Kahn & Pendleton, 2007).

Currently, Montessori high schools do not have a widely used means of measuring the quality of their teachers. In an attempt to evaluate the performance of their teachers, the administrators at the Montessori high school in this study developed a teacher evaluation form, also known as a student evaluation of teaching (SET), to implement in their school. Basing their ideas on principles promoted by the North American Montessori Teacher Association ([NAMTA], 2015), the administrators developed a teacher evaluation instrument composed of 19 items. Partnering with the study's authors, the administrators sought to determine the quality of the instrument and identify ways to improve it.

The purpose of this study was to conduct a validation analysis on a Montessori high school SET, thereby demonstrating how the Montessori community can begin to rigorously examine the quality of measurement and assessment instruments implemented in their schools. To address the primary purpose of this study, we developed three research questions about the quality of the SET measurement instrument: (a) How well did the SET measure teacher effectiveness? (b) How well did the individual SET items measure teacher effectiveness? and (c) How well did the ability to endorse items on the SET align with an established model of teacher effectiveness?

Background

Montessori High Schools

The core of Montessori secondary educational philosophy is taken from Dr. Montessori's (1973) work, *From Childhood to Adolescence: Including Erdkinder and the Function of the University*, in which she proposed that adolescents be educated through an *Erdkinder* system. Meaning "children of the earth," Erdkinder was to be a largely unstructured environment in which adolescents worked and lived together in a farm setting (Barker, 2011). In addition to cognitive outcomes, the goal of Erdkinder is to develop students' social, emotional, and moral characteristics by cultivating social skills, emotional self-awareness, and introspective reflection (Mayes & Williams, 2013; Montessori, 1973; R. Miller, 1990; J. P. Miller, 2010). The development of these characteristics is believed to prepare students to be independent and successful in their postsecondary lives.

Although Dr. Montessori's (1973) foundational text indicated that teachers play a unique and critical role in Erdkinder, the specific expectations for teachers' actions were vague. For example, Dr. Montessori (1973) wrote, "teachers must be young, of open minds, ready to take an active part in the life of the school and to contribute personally" (pp. 124–125), although the specific details of how teachers were to achieve these ends were not detailed. Dr. Montessori (1973) further explained that teachers should facilitate students' learning, work to cultivate an appreciation of content knowledge in students, and be caring individuals. However, beyond encouraging students to learn content material through their farm work, Dr. Montessori provided scant details on how teachers were to reach these goals. In one of the few others instances where she directly addressed the issue of adolescent education, Dr. Montessori (2011) argued that a true Erdkinder teacher "has a real personality, a feeling heart, and takes keen interest in her pupils; one in whom children recognize a source of inspiration and upon whom they can rely" (p. 55), again failing to provide details of how to realize these goals in the school setting. Dr. Montessori's silence on how to transfer these teacher traits into a school environment has allowed for a great deal of variation in the instructional behaviors of Montessori secondary teachers (Kahn, 2011).

Without consistency in the expectations for Montessori high school teachers, standardized teacher evaluation across the Montessori community has lagged, as Montessori high school administrators cannot refer to a key set of practices to determine if their teachers are demonstrating Montessori best practices. Across the Montessori secondary education spectrum, the large American Montessori organizations, such as the American Montessori Society (AMS) or NAMTA, do not indicate a standardized Montessori SET

for use in Montessori high schools throughout the United States. An endorsement of an SET from an organization such as AMS or NAMTA, along with summary statistics on rates of usage in schools and results, would strongly indicate that such an instrument had been developed and widely implemented in the Montessori secondary community; however, no such endorsements or statistics are provided by these two leading organizations.[1]

Without a set of psychometrically sound standardized evaluation instruments to implement across Montessori secondary schools, the Montessori secondary community is unable to evaluate its teachers and schools for consistency and quality. Although the Montessori secondary community struggles with evaluation issues, these issues are also pervasive at the earlier Montessori grade levels. The extent to which the overall Montessori community faces challenges with evaluation issues indicates great potential for psychometric instrument development and validation to address these issues.

Student Evaluations of Teaching

SETs are commonly used to evaluate schools (Kulik, 2001; Wright & Jenkins-Guarnieri, 2012). Historically, SETs have had several purposes: to capture student perspectives on their experiences with teachers and administrators for improvement purposes, to aid other students interested in a course or a specific instructor, and to gain information for academic research (Marsh & Dunkin, 1992). In order for SETs to be effective, an understanding of which factors make for a high-quality instructor must be established. Feldman proposed such factors in his work, *The Superior College Teacher from the Students' View* (1976). Although the model was intended for postsecondary schools, the factors also apply to other levels of education, including secondary schools. In Feldman's model, the three factors that produced a quality teacher were *presentation*, *facilitation*, and *regulation*. Presentation referred primarily to a teacher's course material delivery and was the easiest issue for teachers to address. Facilitation referred to the work teachers completed with students within the context of their interactions, which was largely influenced by the instructor's personality. Regulation referred to the administrative processes of a course, such as implementing a fair grading policy, and was the most challenging factor for teachers to influence. Although Feldman's model did not address all the factors of being a teacher, it provided a foundation for examining teacher performance in the classroom.

The Rasch Model

The Rasch model is a psychometric technique commonly used to conduct validation analyses on tests and surveys and is closely compared to a one-parameter item response theory model (Bond & Fox, 2007; de Ayala, 2009). Key features of the Rasch model include assigning difficulty levels to items and ability levels to respondents (Bond & Fox, 2007). In assigning these levels, researchers and policy makers better understand the degree of difficulty of a measurement instrument (e.g., a survey or assessment) and the ability of respondents to endorse the items. For example, an instrument featuring many items with low difficulty levels is expected to result in many respondents demonstrating high ability levels; in contrast, an instrument with a large amount of high difficulty level items would result in many respondents showing low ability levels (Bond & Fox, 2007). To extend this example to an SET, higher ability students would rate their teachers more positively, and lower ability students would rate their teachers more negatively. The presence of too many easy-to-endorse items promotes artificially positive endorsements of a teacher, and the presence of too many hard-to-endorse items promotes artificially negative endorsements. Developing an instrument with a range of item difficulty levels helps assure that the constructed instrument can assess

[1] NAMTA (2015) does provide examples of evaluation procedures and instruments; however, it does not indicate that these materials have been psychometrically validated for measurement quality. The organization also does not provide information regarding the usage of these resources by Montessori high schools.

or evaluate fairly and as intended, as it simultaneously takes into account respondents' varying ability levels.

When an instrument uses rating scales or Likert-type data, such as in survey research, polytomous forms of the Rasch model are used (Bond & Fox, 2007). If there is a reason to believe that respondents interpreted the response categories differently (e.g., if response categories changed midway through a survey and there were concerns that respondents did not notice this change in categories), then the Rasch partial credit model (PCM) is recommended over other polytomous models (Bond & Fox, 2007). The formula for the Rasch PCM model (Wright & Masters, 1982) is

$$\frac{\exp(\beta_n - \delta_{ik})}{\phi_{nik} = 1 + \exp(\beta_n - \delta_{ik})},$$

where ϕ_{nik} is the probability that person *n* will respond to item *i* with response *k*. $\beta_n - \delta_{ik}$ is the ability (β) of person *n* subtracted from the difficulty (δ) of moving to the *k* rating of item *i*. When interpreting the item difficulty levels in the Rasch PCM, item difficulty levels demonstrate the point on the item threshold at which endorsing a category above the point is equal to endorsing a category below the point.[2] Both person ability and item difficulty estimates are reported on a logit scale, which allows for comparisons of interval level growth, with reported logit scales commonly running from -3.0 to 3.0 (de Ayala, 2009; Toland 2014; Wright, 1993). In practical terms, difficulty level in response to a survey item is connected to a respondent's ability to endorse, or positively rate, that item. Thus, item difficulty levels with negative logit items are easier to endorse than are positive logit items. For person ability levels, respondents with negative logit scores are less able to endorse items than are respondents with positive logit scores.[3] Examining and interpreting these logits are important components of determining the quality of a measurement instrument in a Rasch validation approach.

In addition to item difficulty levels and person ability levels, item and person reliability estimates are reported in a Rasch PCM analysis. Item reliability is a means of determining whether the analysis contained a sufficient sample size to develop item difficulty estimates that accurately reflect the item's difficulty level (Linacre, 2015). Person reliability is a means of determining if the instrument included a sufficient number of items to accurately identify the person ability levels of respondents in the sample. An estimate of .80 is considered sufficient for both item reliability and person reliability (Linacre, 2015).

An additional set of item level estimates are developed, known as infit and outfit estimates, which provide insight into the quality of individual items in an instrument (Bond & Fox, 2007). Item infit and outfit *z* scores, reported as *t* statistics, are expected to fall within the range of -2.0 to 2.0, indicating that an item functions appropriately, thus suggesting that lower ability level respondents were less likely to endorse the item than were higher ability level respondents (Bond & Fox, 2007). In contrast, inappropriately functioning items function inconsistently, where lower ability level respondents may be more likely to endorse the item than may higher ability level respondents. These infit and outfit estimates are a means to identify issues with individual items.

Beyond the item level and person level estimates, the Rasch PCM includes estimates that help determine the *dimensionality* of the instrument (Bond & Fox, 2007; de Ayala, 2009). Dimensionality refers to the instrument's measurement of a latent trait, so unidimensionality indicates that the instrument measures a single latent trait (e.g., evaluating only teacher performance instead of both teacher performance and school climate). Determining unidimensionality requires examining the results of the reported principal

[2] More advanced validation techniques examine threshold functioning, which is beyond the scope of this article. Threshold examination is of particular interest when a rating scale has a large number of response categories (Bond & Fox, 2007).

[3] In the context of surveys for evaluation purposes, ability level refers to a respondent providing a high or low endorsement of the subject (Linacre, 2015; Nardi, 2006).

components analysis (PCA) of Rasch residuals (Linacre, 2015). A PCA of Rasch residuals returning a first contrast with an eigenvalue below 2.0 indicates the instrument is unidimensional. However, a first contrast with an eigenvalue at or above 2.0 means that it must be determined whether the mapping of these residuals showed items of the same facet type. This clustering of items with the same facet would suggest the presence of a second latent trait, likely one matching the facet of the clustering item and thus indicating multidimensionality. Although instruments can still function when they are multidimensional, a unidimensional instrument provides users with results that intentionally measure a single concept.

Methodology

The study methodology was designed to examine how the SET instrument functioned both at the item level and as a complete instrument through a survey validation framework. Two analyses were conducted. The first analysis was a Rasch PCM analysis that examined the Montessori high school SET in its original form. Appendix A includes the original SET, and each item is labeled with a facet that corresponds to one of Feldman's SET facets (1976). Montessori high school administrators collected data for the first analysis in the fall semester of 2014; data included responses from the 27 students who attended the study school. Students completed an SET for multiple teachers in the school, increasing the overall number of responses included in the analysis. After the first analysis, the authors reviewed the results with the SET creators and suggested possible revisions.

The second analysis examined the revised form of the Montessori high school SET using a Rasch PCM analysis. The revised form of the SET can be found in Appendix B. Similar to the original SET, each item is labeled with a facet that corresponds to one of Feldman's (1976) SET facets. School administrators collected data for the second analysis in the spring semester of 2015; data included responses from the same sample of students who provided data in the fall semester of 2014 and who were used in the initial analysis.

The survey validation framework used in this study guided the estimates examined as a result of each Rasch PCM analysis. The validation framework for this study was similar to that used by Royal and Elahi (2011) and Bradley, Sampson, and Royal (2006). These frameworks included examining estimates of instrument unidimensionality, item reliability, person reliability, item fit, and the spread of item difficulty levels. The analysis began by determining whether the item and person reliability estimates were at or above the suggested .80 level, which would indicate that the estimates developed by the Rasch PCM analysis can be confidently interpreted for the purposes of determining the quality of the measurement instrument (Bond & Fox, 2007). Determining reliability was followed by determining dimensionality, which required examining the PCA of the Rasch residual results for unidimensionality as determined by contrast estimates and factor loadings (Linacre, 2015). This was followed by examining item fit, which included determining if the infit and outfit *t* statistics were between -2.0 and 2.0. The analysis concluded by examining the spread of item difficulty levels and determining how this ordering compared to the theoretical item ordering of Feldman's (1976) model. To support that the instrument is measuring the proposed latent trait, the Presentation facet items should be the easiest to endorse, followed by the Facilitation facet items, and finally the Regulation facet items. If the item difficulty levels matched Feldman's (1976) model, then it would indicate that the instrument's items were at appropriate levels for the SET.

For this study, all analyses were conducted using Winsteps (Version 3.92.1; Linacre 2016). The first analysis included data from 106 student ratings, and the second Rasch PCM analysis included data from 105 student ratings. To protect the anonymity of participants, no demographic variables or student identifiers were included in the dataset, so these elements were excluded from both analyses.

Results

We begin with describing the outcomes of the analysis conducted on the initial SET. The first analysis results are followed by a detailed description of how the authors shared the results with the SET developers. We conclude with details of the analysis results from the revised SET.

First Analysis

Results of the Rasch PCM analysis indicated several measurement issues with the initial SET. Framing the interpretability of these results, both the item reliability estimate (.84) and the person reliability estimate (.86) were satisfactory, indicating the analysis included both a sample size and number of items sufficient to confidently interpret the generated estimates (Linacre, 2015). Next, the researchers examined the instrument's dimensionality results. Initial results indicated the instrument was not unidimensional; the results of the PCA of the Rasch residuals estimated the eigenvalue of the first contrast to be 3.0, above the 1.9 recommendation of Linacre (2015) and indicating that item loadings needed to be examined. The item loadings of the first contrast are reported in Table 1. The first contrast had a large representation of items from the Presentation facet with positive loadings (seven out of eight items). Unidimensionality could not be assumed, given the clustering of Presentation items with positive loadings in the first contrast, which indicated the instrument had issues with appropriate measurement. The instrument likely measured Presentation as a full dimension, rather than as a facet of the intended teacher-quality dimension.

Table 1

Item Level Estimates for Initial SET

Item	First contrast loading	Measure	*SE*	Infit mean-square	Outfit mean-square	Infit *t*	Outfit *t*
p1_i	.22	.14	.13	.90	.73	-.6	-1.4
p2_i	.40	-.05	.13	.80	.91	-1.3	-.3
p3_i	-.37	-.90	.17	1.40	1.23	1.8	.8
p4_i	-.22	-.26	.14	.87	1.06	-.7	.4
p5_i	.12	.60	.12	.94	.99	-.4	.0
p6_i	.37	.58	.12	1.85	1.95	4.7	4.3
p7_i	.03	-.15	.13	.72	.66	-1.8	-1.6
p8_i	.60	.04	.13	1.31	1.17	1.8	.8
p9_i	.31	.30	.12	.94	.84	-.4	-.8
p10_i	-.03	.09	.13	1.66	1.42	3.6	1.9
p11_i	-.52	-.30	.14	1.07	.76	.5	-1.0
f1_i	-.01	.31	.12	.56	.63	-3.4	-2.2
f2_i	-.27	.14	.13	.79	.82	-1.4	-.9
r1_i	.68	.32	.12	1.31	1.44	1.9	2.1
r2_i	-.04	-.27	.14	.95	.86	-.2	-.5
r3_i	-.47	-.17	.13	.81	.70	-1.1	-1.4
r4_i	-.58	.09	.13	.77	.82	-1.5	-.9
r5_i	-.54	-.04	.13	.92	.89	-.5	-.5
r6_i	-.61	-.46	.15	.74	.57	-1.5	-1.9

Note. Items in the Presentation facet begin with the prefix "p"; Facilitation items begin with the prefix "f"; Regulation items begin with the prefix "r."

The results at the item level were mixed. The *t*-statistic estimates, reported in Table 1, indicated that four items—p6_i, p10_i, f1_i, and r1_i—demonstrated an issue with misfit (Bond & Fox, 2007; Linacre, 2015). These estimates indicated that students with varying views of their teachers were likely endorsing teachers in a similar manner, and thus the misfit items misrepresented students' perspectives. Item difficulty levels were then examined to determine the presence of a range of item difficulty levels and to evaluate the ordering of item difficulty levels as compared to the model proposed by Feldman (1976).

The Wright map[4] in Figure 1 demonstrates the logit hierarchy of the item difficulty estimates. As Figure 1 shows, item difficulty levels overlapped greatly, suggesting a redundancy in item measurements. This result indicated that students were not asked to endorse items from a range of difficulty levels. Therefore, it is likely that all teachers, regardless of quality, were given similar ratings that prevented administrators from identifying high- and low-performing teachers. The difficulty levels also did not extend below -1.0 logit or above 1.0 logit, indicating that the instrument did not effectively measure respondents at the highest and lowest ability levels. This result also demonstrated that students were prevented from expressing highly positive or highly negative views of teachers, as there were no items that reflected these views. These initial item difficulty results indicate that the SET lacked an appropriate range of items with varying difficulty levels.

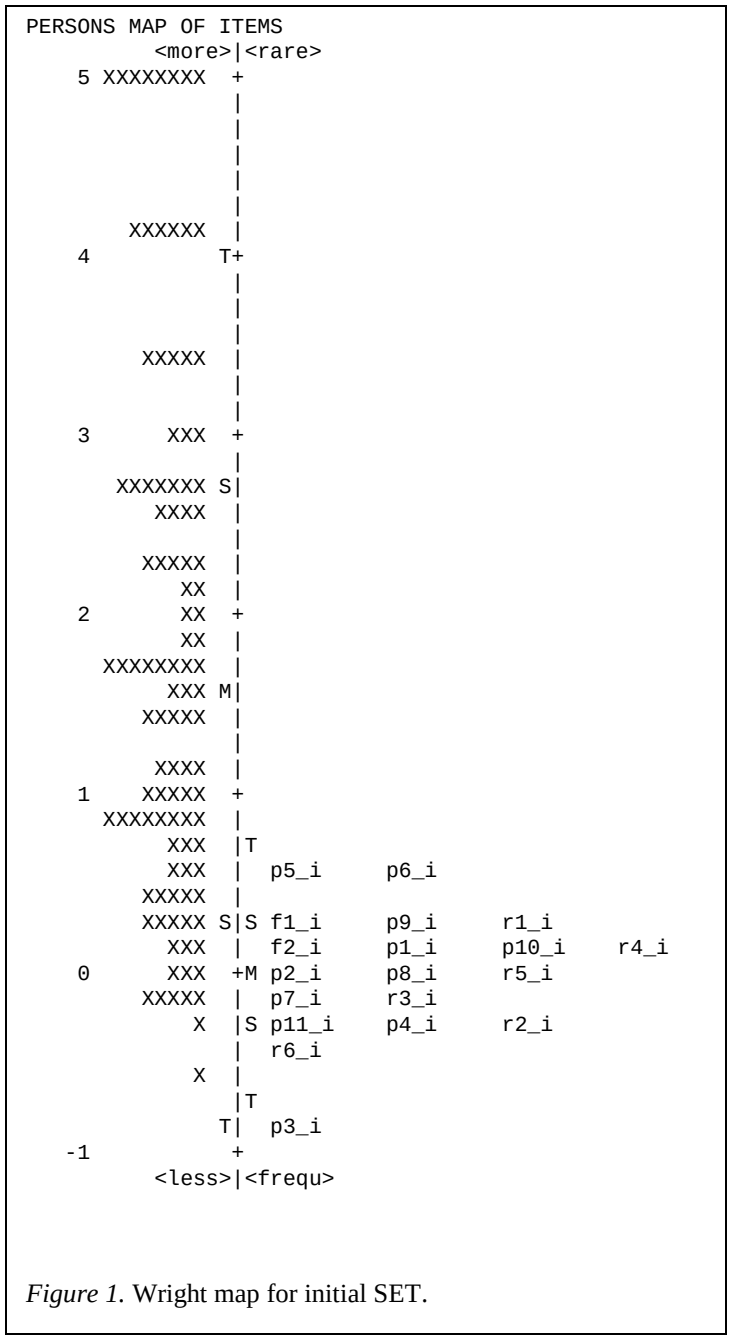

Figure 1. Wright map for initial SET.

When comparing the item ordering to Feldman's (1976) model, additional issues with item difficulty levels became apparent. As Figure 1 demonstrates, there was no clear indication of an item difficulty ordering based on facet. Although Presentation items should be the easiest to endorse, with item difficulty levels ideally at the low end of the negative range, Presentation items appeared throughout the item difficulty range. After the Presentation facet, the Facilitation items should be the next-most difficult items to endorse, according to Feldman's (1976) model. The two Facilitation items fell into appropriate item difficulty levels, with both at the moderately-difficult-to-endorse level, .14 (item f2_i) and .31 (item f1_i). However, the Facilitation items were at similar levels as several items from other facets, indicating the SET did not contain the appropriate items at the moderately-difficult-to-endorse level; according to Feldman's (1976) model, the non-Facilitation facet items should not be at this level. According to Feldman's model, Regulation items should have been among the most challenging items to endorse. However, only two of the Regulation items had item difficulty levels above the 0.0 logit (i.e., more difficult to endorse; Bond & Fox, 2007). The comparison of item difficulty estimates

[4] Wright maps are a means of reporting Rasch model results (Bond & Fox, 2007). The left side of the map indicates person ability levels, and the right side of the map indicates item difficulty levels. The numbers in the middle of the map are the logit scale for the person ability and item difficulty estimates.

to Feldman's model further indicated issues with the initial SET. Overall, the results demonstrated that several measurement issues in the initial SET needed to be addressed.

Revision Process

After conducting the first analysis, the researchers discussed the results with the Montessori high school administrators who developed the initial SET. The discussion included a detailed explanation of how well the overall instrument functioned in its ability to evaluate teachers, as well as an explanation of how well individual items functioned. Based on these results, the researchers suggested several ways to improve the instrument, including revising or dropping the misfit items (i.e., items p6_i, p10_i, f1_i and r1_i) and revising current items to be much easier to endorse or much harder to endorse.

The researchers also recommended how to enhance the instrument to assure measurement quality. These recommendations included revising items (a) that were double-barreled, (b) that may have used terms unfamiliar to respondents, and (c) that included clauses that could be interpreted differently by respondents (Nardi, 2006). For example, a double-barreled item, such as item p4_i "Lesson topics are clear and concise," uses a conjunction that may cause a respondent to answer both parts of the question or only one part of the question. Using unfamiliar terms, such as "differentiation of instruction" in item f1_I, could confuse respondents who were unfamiliar with the term. The problem with the use of descriptive clauses in items, such as in item p11_i "Has a good rapport with students, based on mutual respect," is that a student's response may change because of the wording of the clause. In this example, a teacher may have a good rapport with students, but it may not be based on mutual respect, possibly leading to inconsistent measurement. The researchers recommended revising items that included any of the three identified measurement issues.

The researchers also suggested revising the SET scale. The initial SET scale used estimated percentages of time as the response categories for students. The researchers identified two measurement issues with this scale: (a) the ability of students to assign temporal percentages to a teacher's efforts, and (b) the practical impossibility for teachers to simultaneously engage in all behaviors all of the time, as an increase of any one behavior would likely lead to a decrease of other behaviors. There also was an issue with percentages overlapping on the scale, allowing students to endorse the same percentage on two different parts of the scale. Therefore, researchers suggested that the scale be revised to ask about infrequency and frequency or disagreement and agreement.

During this meeting, the administrators asked many questions about the findings and the researchers' recommendations. Administrators also discussed their concerns about the revision or removal of items, which the researchers noted. Based on the results and administrator feedback, the researchers revised the SET. Finally, the administrators incorporated their own revisions to the instrument and implemented the revised instrument with their students in a scheduled evaluation.

Second Analysis

Results of the Rasch PCM analysis on the revised SET (SET-R) indicated the instrument still had measurement issues, despite the revisions. The reliability estimates were above the preferred level of .80, with person reliability at .87 and item reliability at .86 (Linacre, 2015). These reliability estimates indicated that the Rasch estimate results could be confidently used by the researchers to answer the research questions. Dimensionality of the SET-R was then examined. The results of the PCA of Rasch residual estimates showed that the first contrast had an eigenvalue of 2.2, indicating that the instrument was likely not unidimensional; however, the item loadings, found in Table 2, did not indicate the presence of an additional factor (Linacre, 2015). The facets of the positive and negative item loadings were mixed and did not cluster on a single facet for either loading; clustering would have indicated the presence of a second dimension. Although the first contrast eigenvalue was above 1.9, the lack of item loadings by facet indicated the instrument could be considered unidimensional.

Table 2

Item Level Estimates for Revised SET

Item	First contrast loading	Measure	*SE*	Infit mean-square	Outfit mean-square	Infit *t*	Outfit *t*
p1_r	.08	-.66	.19	.79	.81	-1.5	-1.1
p2_r	.73	-.52	.18	1.00	1.03	.0	.2
p3_r	-.24	-.93	.20	1.29	1.13	1.8	.7
p4_r	.37	.19	.17	.71	.77	-2.0	-1.5
p5_r	-.31	.12	.17	.98	.97	-.1	-.1
p6_r	-.43	.26	.17	1.22	1.24	1.4	1.5
p7_r	-.41	.41	.17	.83	.81	-1.2	-1.2
f1_r	.62	-.23	.18	1.26	1.21	1.6	1.2
f2_r	-.02	-.70	.19	.67	.81	-2.4	-1.0
f3_r	.30	.57	.17	1.51	1.65	3.0	3.5
f4_r	-.41	-.38	.18	.86	.82	-.9	-1.1
f5_r	.40	.17	.17	.70	.79	-2.1	-1.4
f6_r	-.14	.12	.18	.89	.84	-.7	-1.0
f7_r	-.30	-.13	.18	1.29	1.48	1.8	2.5
f8_r	.02	-.13	.18	.72	.70	-2.0	-2.0
r1_r	.06	.03	.17	.88	.94	-.8	-.3
r2_r	-.21	1.20	.16	1.61	1.63	3.6	3.4
r3_r	-.18	.52	.17	.69	.78	-2.2	-1.4
r4_r	.18	.09	.18	.89	.88	-.7	-.7

Note. Items in the Presentation facet begin with the prefix "p"; Facilitation items begin with the prefix "f"; Regulation items begin with the prefix "r."

The item level estimates, reported in Table 2, indicated issues with the measurement of specific items on the SET-R. Of the 19 items, six indicated a misfit issue, according to their infit or outfit *t*-statistic estimates (i.e., items f2_s, f3_s, f5_s, f7_s, r2_s, and r3_s; Bond & Fox, 2007). These outcomes showed that students with both more and less favorable perceptions of their teacher were likely similarly endorsing the misfit items. The item difficulty levels further indicated an issue with the instrument, as the spread of difficulty levels was not wide, and thus the instrument could not distinguish well between students with more favorable and students with less favorable perceptions of their teachers. As is evident in the Wright map in Figure 2, the item difficulty levels clustered in the moderate range, between -.93 and 1.20. Additional items at the easier-to-endorse and more-challenging-to-endorse levels would need to be added to increase the instrument's ability to measure the full range of student perceptions.

The last component of the analysis examined the order of the SET-R items with their difficulty levels to determine if they were aligning by facet with Feldman's (1976) model. As Figure 2 shows, the item difficulty levels did not order by facet in this way because the Presentation and Facilitation items were at similar difficulty levels; that is, the Facilitation items were not consistently more challenging to endorse than the Presentation items. The Regulation facet items were among the more difficult items to endorse, although the items overall fell within the moderately difficult logit range (Bond & Fox, 2007). Additionally, students completing the SET-R were able to endorse Regulation items at a similar level as Presentation items, although Regulation items should be more challenging to endorse. These results indicated that the SET-R did not contain items with appropriate difficulty levels according to Feldman's (1976) model and needed additional revision.

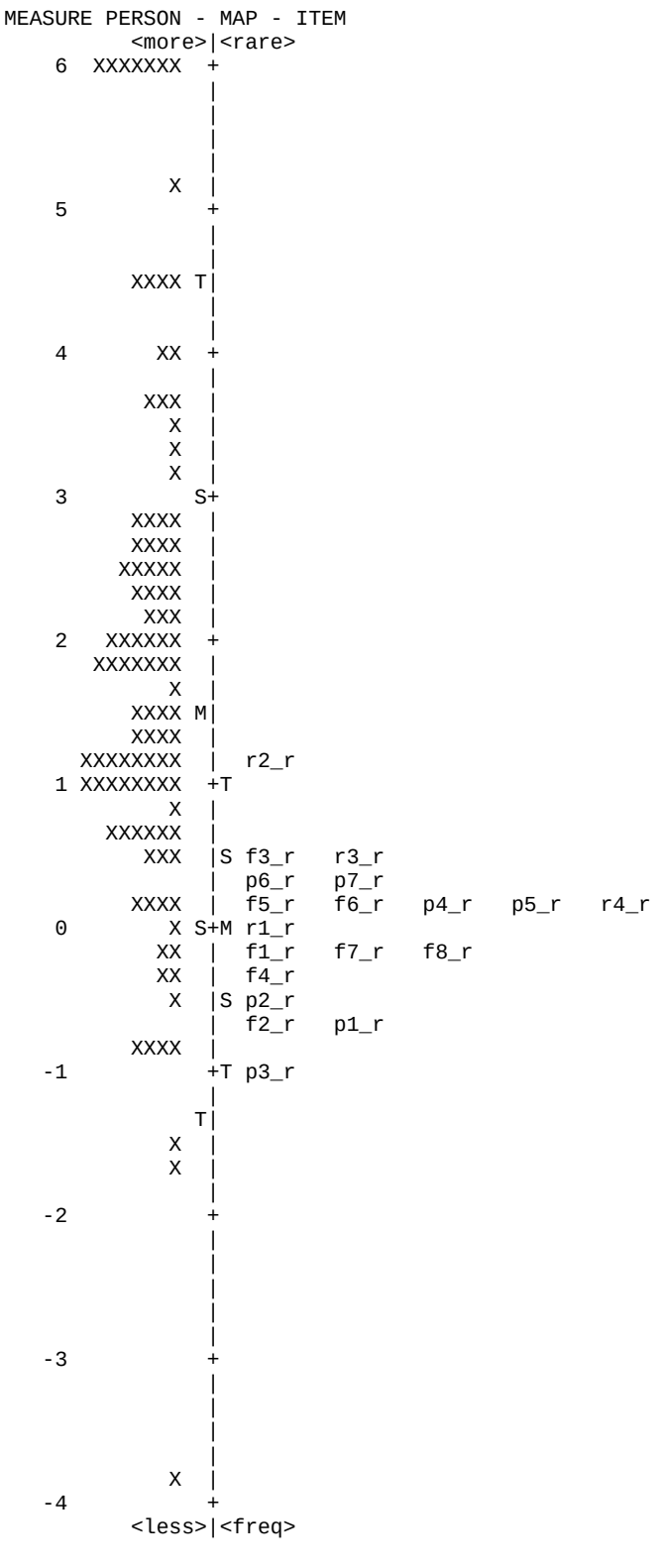

Figure 2. Wright map for revised SET.

Discussion

The primary purpose of this study was to examine an SET in use in a Montessori high school and, in doing so, show the Montessori community how to rigorously examine the quality of measurement and assessment instruments used in their schools. The results indicated that the SET-R needed additional reworking before it could be confidently used for evaluation purposes. We answered the first research question (i.e., "How well did the SET measure teacher effectiveness?") by examining the dimensionality of the instrument and determining if there was a spread of item difficulty levels for students to endorse. According to the results, the SET-R can be interpreted as unidimensional, suggesting the instrument is measuring the concept of teacher effectiveness. However, the lack of item spread showed that the instrument was incapable of measuring the wide range of person ability levels. Additional revision is needed before the instrument can provide effective measurement of students' perceptions. We answered the second research question (i.e., "How well did the individual SET items measure teacher effectiveness?") by examining the items for misfit issues. As the results showed, the SET-R included six misfit items. These misfit items indicated that, although most items were capable of measuring teacher effectiveness, additional item revision is needed to assure all items provide effective measurement. To answer the third research question (i.e., "How well did the ability to endorse items on the SET align with an established model of teacher effectiveness?"), we examined the facets of the items in relation to their item difficulty levels. Comparing the item difficulty levels and their facets to Feldman's (1976) model demonstrated that the Presentation and Facilitation facet items did not have the expected item difficulty levels. The Regulation items were among the more

difficult items to endorse, although their overall difficulty levels were not at the highest levels that Feldman's (1976) model proposed (Bond & Fox, 2007). The answers to these research questions demonstrated that, to assure the instrument is of high quality, additional work on the SET-R is needed.

Conclusion

The authors conclude that the SET-R needs additional revisions. Possible revisions include dropping or revising the misfit items, as well as assuring that items at low and high difficulty levels are included on the instrument. After these initial revisions are made, the instrument will begin to better measure the views of students with both higher and lower perceptions of their teacher. Altering the misfit items will also aid in assuring the instrument is measuring a unidimensional trait, as these misfit items are likely interfering with the clarity of the instrument's overall measurement (Bond & Fox, 2007). An additional possible revision includes removing items in the same facet at similar difficulty levels, as these items with similar difficulty levels are providing duplicative measurements of the same concept. For example, items p6_r and p7_r have similar difficulty levels, .26 and .41 respectively, and both measure elements of presentation. Removing either p6_r or p7_r would reduce the number of questions a student has to answer but still capture the student's perception of a teacher's presentation quality.

Additional revisions of the SET-R and continued validation studies will ultimately yield a high-quality instrument that can be widely implemented in Montessori high schools. The results from this instrument could collect data that would allow Montessori stakeholders and administrators to determine the quality of their teachers and make informed decisions about the future, thus ensuring the best educational experiences for students. Furthermore, the development of this high-quality instrument would demonstrate to the Montessori community that its schools and teachers can be evaluated in a quantitative manner that aligns with its values. We hope the validation process described here has shown the Montessori community how to rigorously examine current measurement instruments and the value of such examination.

Limitations

This study had two primary limitations. First, the Rasch analyses would have benefitted from a larger student sample, which would have permitted the development of more accurate estimates (Bond & Fox, 2007; Linacre, 2015). Second, because the development of the SET items was not based on a set of general principles that is accepted by the Montessori high school community—which arguably does not exist—it may not be accepted by the wider Montessori audience (Barker, 2011; Kahn, 2011). The extent to which the SET items reflect Montessori views on desirable Erdkinder teacher traits is unclear, as the items were developed from outcomes pertaining to a certain Montessori secondary-school philosophy and then modified for inclusion on an evaluation instrument (Barker, 2011). Without an extensive and well-developed model for Erdkinder teacher effectiveness, the items can only be compared to non-Montessori–specific models of teacher effectiveness, and thus their reflectiveness of Montessori values cannot be confirmed. Additional work in the area of Erdkinder standards would enhance these schools' ability to develop evaluation instruments and systems that clearly reflect a unified Montessori vision of Erdkinder education, as instruments such as the SET in this study could be compared to those agreed-upon standards. The limitations of this study can be addressed by collecting data from a wider sample of students and by confirming the appropriateness of the instrument's items with members of the Montessori community.

AUTHOR INFORMATION

†Corresponding Author

Anthony Philip Setari† is a recent doctoral student graduate of the University of Kentucky and can be reached at anthony.setari@uky.edu.

Kelly D. Bradley is a professor in the College of Education at the University of Kentucky.

References

American Montessori Society. (2015). *Montessori schools*. Retrieved from http://amshq.org/Montessori-Education/Introduction-to-Montessori/Montessori-Schools

Barker, D. (2011). A historical look at Montessori's Erdkinder. *Communications, 1-2,* 96–112.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York, NY: Lawrence Erlbaum Associates.

Bradley, K. D., Sampson, S. O., & Royal, K. D. (2006). Applying the Rasch rating scale model to gain insights into students' conceptualisation of quality mathematics instruction. *Mathematics Education Research Journal, 18,* 11–26. doi:10.1007/BF03217433

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

Donahoe, M., Cichucki, P., Coad-Bernard, S., Coe, B., & Scholtz, B. (2013). Best practices in Montessori secondary programs. *Montessori Life, 25*(2), 16–23.

Feldman, K. A. (1976). The superior college teacher from the students' view. *Research in Higher Education, 5,* 243–288.

Kahn, D., & Pendleton, D. R. (2007). *The whole-school Montessori handbook for teachers and administrators*. Burton, OH: North American Montessori Teachers' Association.

Kahn, D. (2011). Eight pictures at an exhibition: A Montessori retrospective on the discovery of the adolescent. *Communications, 1-2,* 15–41.

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research, 109,* 9–25. doi:10.1002/ir.1

Linacre, J. M. (2015). *A user guide to Winsteps Ministeps Rasch-model computer programs: Program manual 3.90.0.* Available from Winsteps http://www.winsteps.com/manuals.htm

Linacre, J. M. (2016). Winsteps [Computer software]. Available from Winsteps http://www.winsteps.com

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 8, pp. 143–233). New York, NY: Agathon Press.

Mayes, C., & Williams, E. (2013). *Nurturing the whole student: Five dimensions of teaching and learning.* Lanham, MD: Rowman & Littlefield Education.

Miller, J. P. (2010). *Whole child education*. Canada: University of Toronto Press.

Miller, R. (1990). *What are schools for? Holistic education in American culture*. Brandon, VT: Holistic Education Press.

Montessori, M. (1973). *From childhood to adolescence: Including Erdkinder and the function of the university.* New York, NY: Schocken Books.

Montessori, M. (2011). Principles and practice in education. *Communications, 1-2,* 50–60. (Reprinted from First Lecture, Institute of Medical Psychology, London, November 10, 1936.)

Nardi, P. M. (2006). *Doing survey research: A guide to quantitative methods*. Boston, MA: Pearson Education.

National Center for Montessori in the Public Sector. (2014). *USA Montessori census*. Available from http://www.montessoricensus.org

North American Montessori Teachers' Association. (2015). *Curriculum downloads*. Available from http://www.montessori-namta.org/Curriculum-Downloads

Royal, K. D., & Elahi, F. (2011). Psychometric properties of the Death Anxiety Scale (DAS) among terminally ill cancer patients. *Journal of Psychosocial Oncology, 29,* 359–371. http://www.tandfonline.com/doi/abs/10.1080/07347332.2011.582639

Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence, 34,* 120–151. doi:10.1177/0272431613511332

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Wright, B. D. (1993). "Logits"? *Rasch Measurement Transactions, 7,* 228. Retrieved from https://www.rasch.org/rmt/rmt72e.htm

Wright, S. L., & Jenkins-Guarnieri, M. A. (2012). Student evaluations of teaching: Combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education, 37,* 683–699. http://dx.doi.org/10.1080/02602938.2011.563279

Appendix A

Initial Version—Student Evaluation of Teaching Questions by Category

Presentation
p1_i. Balances student-centered and teacher-centered instruction (i.e., direct instruction to large group, but also small group lessons and coaching of small groups and individuals).
p2_i. Actively teaches and coaches during class time but gives ample time for independent work (shelf work/project work).
p3_i. Has a thorough knowledge of course content.
p4_i. Lesson topics are clear and concise.
p5_i. Provides a variety of teaching methods on a regular basis.
p6_i. Provides several work/project options for students to choose for lessons.
p7_i. Facilitates smooth transitions between activities.
p8_i. Manages lessons so that they begin and end in a timely manner, leaving enough time for independent work.
p9_i. Lessons are engaging; using hands-on materials, real-life experiences, and encouraging discussion as much as possible.
p10_i. Encourages discussion in seminars and/or lectures.
p11_i. Has a good rapport with students, based on mutual respect.

Facilitation
f1_i. Understands how to use differentiation of instruction so that all students are challenged and supported.
f2_i. Asks questions that employ higher order thinking skills during lessons/discussions to promote thinking "outside the box."

Regulation
r1_i. Provides the opportunity for large blocks of work time.
r2_i. Provides access to curriculum and course objectives.
r3_i. Understands how to set up the necessary infrastructure for students to follow guidelines that create student success and a pleasant classroom environment.
r4_i. Employs and teaches students creative resolution techniques to resolve conflict in the classroom.
r5_i. Knows when to intervene to guide students who exhibit inappropriate behavior.
r6_i. Fosters a learning environment that encourages concentration, self-discipline, respect, and independence.

Rating Scale[5]
1 = 60% or less of the time
2 = 60%–70% of the time
3 = 70%–80% of the time
4 = 80%–90% of the time
5 = 90%–100% of the time

[5]An important concern about this instrument is that the rating scale overlaps in percentages across different rating levels (e.g., 60% is present in both a rating of 1 and a rating of 2).

Appendix B

Revised Version—Student Evaluation of Teaching Questions by Category

Stem
My Montessori teacher:

Presentation
p1_r. Explains course objectives.
p2_r. Allows time for independent work.
p3_r. Has a thorough knowledge of course content.
p4_r. Clearly explains the topic of lessons.
p5_r. Challenges students at all levels of learning.
p6_r. Uses a variety of teaching methods.
p7_r. Teaches engaging lessons.

Facilitation
f1_r. Provides large blocks of work time.
f2_r Provides individual attention to students.
f3_r. Provides options for students to choose their work.
f4_r. Ask questions that challenge students.
f5_r. Manages classroom time well.
f6_r. Encourages class discussions.
f7_r. Has a good relationship with students.
f8_r. Fosters a learning environment that promotes independence.

Regulation
r1_r. Sets clear classroom guidelines.
r2_r. Resolves classroom conflict with creative techniques.
r3_r. Corrects students who exhibit inappropriate behavior.
r4_r. Facilitates smooth transitions between activities.

Response Scale
1 = Strongly disagree
2 = Disagree
3 = Agree
4 = Strongly agree