Learning to Teach for Social Justice—Beliefs Scale: An Application of Rasch Measurement Principles

Larry H. Ludlow, Sarah E. Enterline, and Marilyn Cochran-Smith

The authors illustrate how a Rasch model can guide the development of a new affective measurement instrument—the Learning to Teach for Social Justice—Beliefs scale. The results provide strong evidence of a meaningful continuum of attitudes about teaching for social justice ranging from those easier to endorse to those more difficult to endorse.

The idea of preparing teachers to teach for social justice is prevalent in a loosely related collection of teacher preparation programs, partnerships, grassroots teacher and community groups, and other initiatives in the United States and elsewhere. Despite national attention, however, there is considerable variation in meanings of the phrase teacher education for social justice, and, in general, this has not been a well-theorized term (North, 2006). Very generally speaking, however, most definitions (e.g., Adams, Bell, & Griffin, 1997; Cochran-Smith, 1999, 2004; Michelli & Keiser, 2005; Oakes & Lipton, 1999; Sleeter, 1996; Villegas & Lucas, 2002; Zeichner, 2003) have in common explicit recognition of the marked disparities in educational opportunities, resources, achievement, and long-term outcomes between minority and low-income pupil groups and their White, middle-class peers. This is coupled with the position that teachers have the potential to be both educators and activists committed to the democratic ideal and to reducing the inequities in American society. Teacher education for social justice, then, is teacher preparation deliberately designed to provide the social, intellectual, and organizational contexts to foster teaching for social justice in schools accommodating students in kindergarten through 12th grade (K–12).

Teaching for social justice in K–12 schools has as its primary consideration promoting pupils' learning (academic, social, emotional, and civic) and enhancing pupils' life chances, including challenging the structures, curriculum, labels, and school arrangements that limit or inhibit life chances. This agenda builds on a wide-ranging body of scholarship, practice, and grassroots efforts, including multicultural theory and pedagogy; research on effective practices in diverse classrooms; critical analyses of education and society; research on culture, language, and identity; organization at the grassroots community level to change schools; and theories related to the role of education in democratic societies. Teaching for social justice builds on and requires knowledge (i.e., knowledge of content, pedagogy, learners, cultures, schooling, communities, as well as knowledge of self), interpretive frameworks (i.e., ways of understanding and acting on the events and processes of schooling based on...
the integration of knowledge with beliefs, values, ethics, moral commitments, and attitudes), and practices (including subject-specific pedagogies and strategies for supporting the learning process of English language learners [ELL], pupils with special needs, and pupils from a range of socioeconomic backgrounds). Teaching for social justice also involves teacher commitment to being part of larger social movements by working as advocates and activists for their pupils.

In this article, we assume that teaching for social justice is a legitimate and measurable outcome of teacher education. The purpose of this article is to present evidence of the extent to which this assumption has been met. Specifically, we present the psychometric characteristics of the Learning to Teach for Social Justice—Beliefs scale. This includes the operational definition of the construct, the item development and pilot testing procedures, item analysis results of both classical test theory (CTT) and item response theory (IRT; Rasch) procedures, and evidence of discriminant validity.

METHOD

Variable Definition

Learning to teach for social justice is conceptualized in terms of six core components: teachers’ knowledge, skill, and interpretive frameworks; teachers’ beliefs, attitudes, and values; classroom practice and pedagogy; community participation; teachers’ learning in inquiry communities; and promoting pupils’ academic, social–emotional, and civic learning. In developing the scale to measure this variable, the Survey Team of the Boston College—Teachers for a New Era (BC–TNE) project (Ludlow et al., 2007) came to a common understanding that any variable, even one as complex as learning to teach for social justice, may be conceptualized as a continuum along which people differ. In the academic setting addressed in this study, this means that teachers would differ in the extent to which they understand, accept, and are prepared to teach in ways consistent with the social justice principles just described.

Measurement Models

The current field of psychometrics relies on two primary measurement models, CTT and IRT. CTT is based on the simple yet powerful concept that an individual’s observed score, defined as the total score on some measurement instrument, is made up of two unobservable, theoretical components: a true score and an error score (X = True + Error; Guilford, 1950; Lord & Novick, 1968; Spearman, 1904). Although the true score is never actually known, it is possible to generate estimates of the extent to which measurement error affected the observed score, thereby reducing the extent to which the true score is captured by the observed score. Hence, great effort is expended to estimate and reduce measurement error because the more measurement error can be reduced, the more confidence there is that the observed score accurately represents the true score.

The basic psychometric tools of CTT include factor analysis, reliability analyses (e.g., test–retest, internal consistency, and inter- and intrarater reliability), and validity analyses (e.g., content, variable, discriminant, divergent, predictive, consequential).

The limitations of CTT (e.g., the ability estimate of a person is dependent on the difficulty of the items, the standard error applies equally to all ability levels, item discrimination can be too high) have been widely recognized (Brennan, 2001; Hattie, Jaeger, & Bond, 1999; Masters, 1988; Traub, 1997; Wainer, 1986) and have led many investigators to use the principles underlying IRT. The IRT models differ, however, in how the probability of a specific response to a specific item is estimated (Hambleton, Swaminathan, & Rogers, 1991; Lord & Novick, 1968).
In general, IRT models are differentiated by the number of parameters associated with various item-specific characteristics. These characteristics are generally referred to as item difficulty, item discrimination, and item pseudoguessing parameters. IRT models that take each of these characteristics into account are referred to as one-parameter (1-PL, or the Rasch model), two-parameter (2-PL), and three-parameter (3-PL) logistic models, respectively (van der Linden & Hambleton, 1997).

The more significant distinction between these models, however, is that they differ in their fundamental purposes. Rasch models are used as confirmatory tests of the extent to which scales have been successfully developed according to explicit a priori measurement criteria. These criteria include the requirements that (a) items define a unidimensional continuum in the domain and (b) items follow a strictly hierarchical ordering in their definition of the domain. If the responses of study participants to the scale items suggest misfit to these criteria, particularly regarding hierarchical ordering, then the items are examined for the purpose of strengthening them—the Rasch measurement model is not discarded or modified. In contrast, 2-PL and 3-PL models are designed to maximize the extent to which item response variation can be accounted for—they are statistical models subject to reexpression in any way that reduces residual variation. Hence, 2-PL and 3-PL models will always fit any data set better than a Rasch model.

From a Rasch measurement perspective, better fit is not a sufficient reason for choosing an IRT model. Rasch models are preferred because they dictate the way analysts think about and subsequently construct measurement instruments. When the data fit the model, the continuous scale is analogous to a linear ruler that is invariant in terms of level of ease or difficulty of accomplishing the task for any individual appropriate for testing. With regard to measuring the variable learning to teach for social justice, a Rasch model was used, not because it fit the data better than any other model, but because when the data fit the model, teacher candidates could be ordered along a continuum based on their endorsements of simpler to more complex beliefs.

Although this notion may seem obvious to some, to measure a human characteristic well, one must know a great deal about it. This is true whether one's interests are in functional ability (Coster, Haley, Ludlow, Andres, & Ni, 2004; Coster, Ludlow, & Mancini, 1999; Haley, Ludlow, & Coster, 1993; Ludlow & Haley, 1992), test anxiety (Ludlow & Guida, 1991), reading ability (Ludlow & Hillocks, 1985), attitude toward math (Ludlow & Bell, 1996), rater effects (Ludlow & Haley, 1996), self-efficacy (Gable, Ludlow, & Wolf, 1990), job satisfaction (Ludlow & Lunz, 1998), or masculine stereotypes (Ludlow & Mahalik, 2001). To speak about measuring a variable is really to question what it means to be characterized as high, low, or moderate on a variable. Specifically, what kinds of tasks separate a person with high math ability from a person with low math ability, what kinds of behaviors characterize a high functioning person from a low functioning person, what distinguishes a teacher who possesses a high degree of commitment to teaching for social justice from one who does not have this commitment?

If one believes that a variable can be operationally hypothesized a priori in a continuous and hierarchical manner, then one can assert that tasks (items) can be created that represent those levels of the variable. Furthermore, if these items can successfully be constructed, then there is an opportunity to both conceptually and literally locate and position a person in relation to that variable and then describe the types of items, or tasks, that are most closely associated with that person's score (which defines that person's position or location) on the instrument (the variable being measured).

Underlying this concept of a variable is the assumption that the variable is unidimensional. This assumption means that there is an attempt to measure one characteristic of a person at a time. This is done not because humans are unidimensional but because to understand behavior well often means that the behavior first must be broken down into simple constructs. Once these simple constructs are clearly understood, it may then be possible to
build more complex models for the purpose of generating a more comprehensive picture of the person. Multidimensional Rasch models exist, but they have not yet made their way into standard practice (Embretson, 1991). They are, however, becoming more frequently used in computer-adaptive testing in the medical realm; for example, using a multidimensional computer-adaptive test to achieve greater levels of precision and efficiency than are possible with separate unidimensional assessments (Haley, Pengsheng, Ludlow, & Fragala-Pinkham, 2006).

The assumption is made that every variable stretches across a continuum of simple-to-complex tasks, or levels of knowledge, or affective characteristics, or cognitive abilities. Based on this assumption, a very deliberate hierarchical arrangement of tasks (like a ladder) can be constructed along which a person is located. This person could also be expected to advance along that continuum (or up the ladder). In a very real sense, a hypothesized structural model of behavior (or perception, attitude, or ability) is being constructed.

After this time-consuming, literature-exhausting effort to define a variable, which commences in item development, the instrument is administered to an appropriate sample of study participants (usually it is first administered as a series of pilot studies). The item responses are then analyzed to determine the extent to which the instrument functioned as intended, or the extent to which the data fit the model. This means the solution is examined to see if there is evidence to support the existence of a unidimensional variable stretching across the a priori defined continuum. It also means examining the extent to which study participants have given responses that are consistent with their estimated locations on the variable, and it means examining items to determine the extent to which these items provoked responses consistent with the locations of these same participants on the variable. This issue of person and item consistency is addressed through various goodness-of-fit statistics.

In this article, we do not argue that there is one and only one way to statistically model a vector of person-level item responses, nor that the best way to measure the variable learning to teach for social justice is through use of these statistical techniques. Our purpose is to illustrate that Rasch measurement principles offer an opportunity to construct measurement instruments that can be useful for understanding skills, abilities, attitudes, perceptions, and behaviors in ways that are fundamentally different from those of instruments designed to be fit by 2-PL IRT models.

**Pilot Analyses**

From a set of more than 200 potential items related to social justice that we gathered from the literature on teaching for social justice, a subset of 25 items that seemed to possess some degree of content validity was administered to 284 graduating teacher education students. The CTT analyses produced low reliabilities, low variances in the item and total scores, and poor item–total correlations. The Rasch analyses produced poor data-to-model fits and no indication of a meaningful continuum that made any theoretical sense with regard to the item or teacher candidate estimates (i.e., location estimates; see The Model section). These results were not entirely unexpected. The 5-year BC–TNE project had a short window of time available for the development of its first exit survey; therefore, either a rough survey with recognizable flaws would be administered or no survey would be administered to the 1st year’s graduating class.

As a result of the poor performance of the 25 items on this first exploratory exercise, a more serious, conceptual, theory-driven approach was taken in the development of the Learning to Teach for Social Justice–Beliefs scale. For example, focus group exercises were conducted with students in our undergraduate and graduate classes. These exercises consisted of prompts such as, “What would be the characteristics of a teacher who is (or, is not) effective at teaching for social justice?” “How would you construct a scale to measure social justice?” and “What is a social justice item that anyone with even the slightest sense
of social justice would agree to?” We conducted these exercises in research methods, test construction, and psychometrics classes. They were introduced specifically during lectures addressing issues of operational definitions, item writing, reliability, construct and discriminate validity, instrument development, pilot testing, and data analysis. We also conducted them in graduate classes addressing philosophies and practice of education.

Those exercises produced a new set of 20 potential items that were subsequently piloted in different research methods classes. Students responded to each item using a 5-point, Likert-type scale that ranged from strongly disagree (1) to strongly agree (5). More importantly, they were asked to comment on the clarity of each statement (e.g., were any terms or phrases confusing, did double-barreled items exist, did the scale responses seem appropriate, did the item itself seem related to social justice?).

These 20 items were analyzed using both CTT and Rasch measurement models. This second round of analysis led to additional changes in the items. Specifically, 8 of those 20 items revealed flaws (e.g., low item–total correlations, confusion in wording, poor fit) that eliminated them from future surveys.

The analyses were also discussed in interdisciplinary BC–TNE team meetings, which provided an additional layer of complexity to our task of attempting to measure the variable learning to teach for social justice. The most interesting aspects in the development of the instrument at this point were discussions that addressed the desired direction of scoring on the items. For example, would a teacher who is better prepared to teach for social justice strongly agree or strongly disagree with the statement, “Teachers should be ‘color-blind’ when it comes to working with students in the classroom”?

**Current Operational Version of the Learning to Teach for Social Justice Scale**

The final set of 12 social justice items resulting from these pilot analyses has been integrated into the BC–TNE Entry Survey, Exit Survey, One-Year Out Survey, Two-Year Out Survey, and Three-Year Out Survey system as the Learning to Teach for Social Justice scale (for details of other measures included in the surveys, see Ludlow et al., 2007). These surveys are completed by the teacher candidates and graduates at crucial times in their education and practice. The 12 items are identical on the Entry and the Exit Surveys, and they are worded such that teacher candidates are asked about their beliefs regarding teaching for social justice. This version of the scale is the Learning to Teach for Social Justice–Beliefs scale. The concepts and subjects addressed in the 12 items remain the same for the One- Two- and Three-Year Out Surveys; however, the items are reworded to ask teachers about their specific classroom practices regarding social justice. This version of the scale is the Learning to Teach for Social Justice–Practices scale. The analysis in this article addresses only the Learning to Teach for Social Justice–Beliefs scale used in the Entry and Exit Surveys.

The Learning to Teach for Social Justice–Beliefs scale (SJ) items are presented in Appendix A. For each item, respondents are asked to answer using a 5-point, Likert-type rating scale in which 1 = strongly disagree, 2 = disagree, 3 = uncertain, 4 = agree, and 5 = strongly agree. These items are positively and negatively worded. The positively worded items (Items SJ1, SJ2, SJ4, SJ7, and SJ8) were conceptualized and written as statements that address relatively less controversial aspects of teaching for social justice. We expected that students would agree with these items.

The inclusion of negatively worded items (Items SJ3, SJ5, SJ6, and SJ9–12) was deliberate. That is, we expected that it would be relatively easy to endorse positively worded statements with which most students should have some minimal level of experience, even in their first semesters of college. The negatively worded items, however, were intended to address concepts and experiences that only an experienced teacher candidate would have encountered. Moreover, to prevent response bias resulting from students responding in a socially desirable positive manner, these items were written in such a way that careful consideration of each
item's meaning and intent would be required. Students were expected to disagree with these items. These items were reverse scored (R) so that higher scores (e.g., scores closer to 5) would correspond with a stronger commitment to teaching for social justice.

Once the items are coded in the appropriate direction, a higher total score corresponds to a higher level of commitment to teaching for social justice. We expected these scoring categories to generate wide variability in participant responses. Furthermore, we included a middle category choice of uncertain because we expected to see changes in the use of this category as teacher candidates progressed through their preparation. In other words, we expected that entering students would not fully understand some of the ideas and concepts that these items in the scale addressed because of their presumed lack of exposure to the concepts associated with the variable learning to teach for social justice. Hence, one measure of program effectiveness would be a reduction in the frequency with which the uncertain category was chosen by graduating candidates on the Exit Survey.

The first administration of these two surveys (Entry and Exit) to the two cohorts of student candidates and graduates now define a baseline for the remaining years of the BC–TNE project. Because an underlying goal of our collective research at Boston College is to transform the teacher preparation program, the analyses will be longitudinal and will provide evidence that feeds back into the program. Hence, the psychometric analyses reported in the current study were intended to ensure that the Learning to Teach for Social Justice—Beliefs scale was invariant across time of testing and experience with teaching. This property of invariance is a necessary condition in subsequent efforts to measure change within student candidates and graduates in the teacher preparation program.

In addition, we expected to find that the mean level of participants’ scores would be relatively low for the incoming class (and all incoming classes) as measured by the Learning to Teach for Social Justice—Beliefs scale in the Entry Survey. We hypothesized that the graduating class (and all graduating classes) would score higher than the entering class as measured by the Learning to Teach for Social Justice—Beliefs scale in the Exit Survey and that as the BC–TNE effect would become more pronounced over time, the mean score of the graduating class on this scale would rise. This pattern would establish the discriminant validity of the scale.

Chronologically, the 2005 Exit Survey was completed by 224 graduating teacher candidates in late spring 2005, and the 2005 Entry Survey was completed by 268 incoming undergraduate and graduate students in early fall 2005. The overall response rate was 88.2% and 90.0% for the Exit and Entry surveys, respectively.

CTT RESULTS

Entry Survey

For those in the incoming class who completed the Entry Survey, the responses to the 12 items of the Learning to Teach for Social Justice—Beliefs scale generated a Cronbach’s alpha of .77, and the item analysis revealed no negative point-biserial correlations. In addition, the items were factor analyzed for the purpose of providing one form of construct validity. Prior to the factor analysis, we checked the appropriateness of the correlation matrix for factoring. The Kaiser–Meyer–Olkin (KMO) statistic was high (.809), the determinant was nonzero, and Bartlett’s test of sphericity was significant, all of which were encouraging results.

Using principle axis factoring (thereby choosing to analyze only common variance) with a varimax rotation, two factors were extracted. The two factors accounted for 34.1% of the total variance. Factor 1 accounted for 17.7%; Factor 2 added 16.4%. These factors were significantly correlated with one another. Specifically, when an oblique rotation was performed, the factor correlation was .36.

The seven negatively worded items define Factor 1, the five positively worded items define Factor 2. Given the scale development specifications, this result was expected. The plot of
the varimax loadings is seen in Figure 1, Box 1. Does this result suggest that two scores should be reported for each person—one positive and one negative score? Or, based on the oblique rotation and the correlation between the two factors, is it possible that the solution can be interpreted as a single variable composed of two clusters of similarly worded items addressing relatively easy and more difficult to endorse aspects of that variable?

Recall that the factors in principal axis factoring are extracted in such a way that each eigenvector (or factor) accounts for the maximum orthogonal variance possible. Hence, the first factor mathematically defines a solution along which the items share the greatest common variance. The second extracted factor then defines a projection through the variable space that accounts for the next greatest amount of common variance. The unrotated factors are customarily rotated because the factor loadings are usually too numerous and too variable to comprehend when presented in a table, and plots of the unrotated factors are usually too indistinct to distinguish any meaningful pattern. That is not the case, however, with the data in this study.

Figure 2, Box 1 contains the unrotated solution for the Entry Survey data. Inspection of the unrotated factor loadings reveals that the three highest loadings on Factor 1 are Items SJ3R, SJ4, and SJ10R—all addressing diversity. (See Appendix A for wording of all items.) The next two items address English language issues (Items SJ5R and SJ6R), and the next two items address race issues (Items SJ1 and SJ2). These items consist of both positive and negative statements. This first unrotated factor provides evidence for the existence of a common Learning to Teach for Social Justice–Beliefs factor. The second unrotated factor then separates out the positive and negative statements (the small circle above the horizontal axis and the larger circle below the horizontal axis, respectively).

The question remains, however—do the circles represent substantive differences or trivial differences that are due to the direction of the phrasing? To answer this, the seven negatively worded items were rephrased as positively worded items. For example, Item SJ3R, “For the most part, covering multicultural topics is only relevant to certain subject areas, such as social studies and literature,” became “Covering multicultural topics is relevant to all subject areas, including math and science.” If responding to the negatively phrased items were simply a more difficult cognitive task, then analysis of this new scale would not show two factors, or clusters, of positive and negative items. Conversely, if the positive and negative items reflected substantive differences, as intended, then factors and clusters similar to the initial solution should result.

FIGURE 1
Varimax Solutions for the Data From the Entry and Exit Survey Learning to Teach for Social Justice–Beliefs Scale

*Note. Factor 1 = negatively worded items; Factor 2 = positively worded items.*
Accordingly, the new set consisting only of 12 positively worded items was administered to two graduate-level classes—one studying psychometrics, the other studying curriculum controversies. These students were not comparable to those who responded to the Entry and Exit Surveys. The data were factored as previously described and nearly identical clusters of items emerged. These results provide evidence that the two clusters of items share a common factor addressing learning to teach for social justice but are distinguishable in that they address different aspects of this variable.

Exit Survey

A similar analysis was conducted on the Exit Survey data for the spring 2005 graduating teacher candidates. These students were exposed to the inherent emphasis placed on social justice that characterizes Boston College, but they experienced minimum exposure to any demonstrable BC–TNE effect because the teacher education program was undergoing changes in curriculum and practica at that time. Their results serve as a baseline against which the results from all future graduating classes will be compared.

The responses of the graduating teacher candidates had a lower reliability estimate compared with the Entry Survey results (α = .71). Interestingly, the scale variance on the Exit Survey (28.9) was considerably less than the variance on the Entry Survey (37.9). Recalling that the magnitude of the Cronbach’s alpha is a function of the scale variance, this slight decline in alpha is understandable and meaningful. This reduction in scale variance may itself be understood as a program effect of having brought a degree of homogeneity to the students after 4 years of study.

The KMO statistic was high (0.755), the determinant was nonzero, and Bartlett’s test of sphericity was significant. Principal axis factoring with a varimax rotation extracted two factors explaining 29.6% of the total variance. Factor 1 (negatively worded items) accounted for 17.5% of the variance; Factor 2 (positively worded items) accounted for 12.1%. The oblique rotation generated a factor correlation of 0.15. Similar rotated and unrotated factor loadings were found in comparison to the Entry Survey. The rotated plot is presented in Figure 1, Box 2; the unrotated plot is presented in Figure 2, Box 2. In Box 2, the unrotated Factor 1 defines the Learning to Teach for Social Justice—Beliefs scale, while the unrotated Factor 2 again separates the positive (less complex) and negative (more complex) items.

The factor structures for the Entry and Exit Surveys are similar in appearance (whether one chooses to use the rotated or unrotated solutions), with the negatively worded items...
loading highest on one factor and the positively worded items loading highest on the other factor. However, it is important to test whether these results are statistically equivalent across the two administrations. For this purpose, coefficients of congruence (CC) were computed. The CC is a measure of the degree to which the factor loadings on corresponding administrations are similar (Cureton & D’Agostino, 1983). Because these factors were derived using the same factoring procedures and were guided by the simple-structure criterion, this comparison is appropriate. The CC will equal 1.0 if the factor loadings are identical, and Cureton and D’Agostino suggested that the CC should be at least .90 if the two administrations produce the same factor. The coefficient is expressed as

$$CC = \frac{\sum l_i - \sum t_i}{\sqrt{\sum l_i^2 \sum t_i^2}}$$

where $l_i$ corresponds to the factor loadings from the first solution, $t_i$ corresponds to the factor loadings on the same factor from the second solution, and $n$ represents the number of items.

The CC for the positively worded factor was .95; for the negatively worded factor, it was .97. These comparisons meet the .90 criterion—the same factors were extracted across samples and levels of experience.

Having satisfied ourselves that the classical item analysis results suggest similar levels of stability and consistency within the scale across the two groups of teacher candidates, we next performed a simple, independent means $t$ test on the total scores. The two means differed significantly—graduating students obtained statistically significantly higher scores than did entering students. These findings are consistent with our expectations and serve as evidence of the scale’s discriminant validity.

**RASCH MODEL RESULTS**

**The Model**

The model used in this study is known as the Rasch rating scale (Andrich, 1988; Wright & Masters, 1982, p. 49). This model is appropriate when the scoring categories have rigorously defined scoring rubrics that transcend, or do not depend on, the characteristics of the specific items. The same analysis was performed on the two survey administrations. The WINSTEPS (Version 3.64.2) software package was used (Wright & Linacre, 1998, see Appendix B).

$$\pi_{nx} = \frac{\sum_{k=0}^{m} \frac{1}{e^{\sum_{k=0}^{m}}}}{e^{\sum_{k=0}^{m} \frac{1}{e^{\sum_{k=0}^{m}}}}}$$

The $\pi_{nx}$ is the probability of person $n$ responding in category $x$ to item $i$ where $\delta_i$ is the location (generically referred to as item difficulty) of item $i$ on the variable, $\tau_j$ is the location (threshold parameter) of the $k$th transition from one scoring category to the next for the $m+1$ rating categories, and $\beta_m$ is the parameter for an individual’s level of commitment to teach for social justice. These location estimates are reported in a metric referred to as logits (Ludlow & Haley, 1995). These estimates simultaneously portray the structure of the learning to teach for social justice—beliefs variable and the location of each teacher candidate along the variable.

**Scale Structure**

Figure 3 contains the separate variable maps for the 2005 Entry and 2005 Exit Surveys. In each map, the items are ordered from easiest to endorse (bottom of the map) to hardest to
<table>
<thead>
<tr>
<th>Entry SJ Survey</th>
<th>Exit SJ Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>(n = 268)</em></td>
<td><em>(n = 224)</em></td>
</tr>
<tr>
<td>&quot;Logits&quot; Persons</td>
<td>&quot;Logits&quot; Persons</td>
</tr>
<tr>
<td>&quot;Items&quot;</td>
<td>&quot;Items&quot;</td>
</tr>
<tr>
<td>5 +</td>
<td>5 +</td>
</tr>
<tr>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>4 +</td>
<td>4 +</td>
</tr>
<tr>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>3 +</td>
<td>3 +</td>
</tr>
<tr>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>2 ### +</td>
<td>2 ### S+</td>
</tr>
<tr>
<td>.### SJ12R</td>
<td>.### SJ12R</td>
</tr>
<tr>
<td>.### SJ11R</td>
<td>.### SJ11R</td>
</tr>
<tr>
<td>1 +</td>
<td>1 +</td>
</tr>
<tr>
<td>.### SJ5R</td>
<td>.### SJ5R</td>
</tr>
<tr>
<td>.### SJ3R</td>
<td>.### SJ3R</td>
</tr>
<tr>
<td>.### SJ10R</td>
<td>.### SJ10R</td>
</tr>
<tr>
<td>0 ### S+</td>
<td>0 ### S+</td>
</tr>
<tr>
<td>.### SJ2</td>
<td>.### SJ2</td>
</tr>
<tr>
<td>.### SJ7</td>
<td>.### SJ7</td>
</tr>
<tr>
<td>#</td>
<td>#</td>
</tr>
<tr>
<td>-1 + +SJ1</td>
<td>-1 +</td>
</tr>
<tr>
<td>SJ4</td>
<td>SJ1 SJ4</td>
</tr>
<tr>
<td>-2 +</td>
<td>-2 +</td>
</tr>
</tbody>
</table>

**FIGURE 3**

Variable Maps for the Data From the Entry and Exit Survey Learning to Teach for Social Justice–Beliefs Scale

*Note.* SJ = Learning to Teach for Social Justice–Beliefs scale; R = reverse-scored item; M = mean candidate estimate; S = standard deviation of estimates. Each "#" is 3 candidates. Each "+" is 1 candidate.
endorse (top of the map). The teacher candidates are ordered from lowest scoring (bottom of the map) to highest scoring (top of the map) in terms of learning to teach for social justice beliefs. As demonstrated by the easiest to endorse items, it is easiest to strongly agree with Items SJ1 and SJ4 (examine one’s own beliefs and incorporate diverse cultures), followed by Items SJ8 (think critically about government policies), SJ2 (racism and inequity openly discussed), and SJ7 (challenge school arrangements). Note that these are the positively worded items identified in the factor analyses. (See Appendix A for exact wording.)

At the top of each map are the hardest items to endorse (by disagreeing). Item SJ12R (prepare students for lives they are likely to lead) is harder to reject than Item SJ11R (success in school depends primarily on how hard students work). This is followed by Item SJ5R (assimilate immigrant children and ELL into American society) and Item SJ3R (covering multicultural topics is only relevant to certain subjects). In the middle of the distribution are Items SJ6R, SJ9R, and SJ10R (lower expectations for ELL, economically disadvantaged bring less and gain more, and it is not the teacher’s job to change society, respectively). It is encouraging and consistent with our scale development expectations that proceeding up the variable means addressing increasingly more complex, controversial, and debatable choices.

The Entry Survey item separation is 8.6, and the person separation is 1.8. For the Exit Survey, item and person separation are 8.4 and 1.5, respectively. These statistics take into account the standard deviation of the parameter estimates relative to the mean measurement error in those estimates (Wright & Masters, 1982). Because these statistics are partly a function of the sample size and number of items, respectively, these values represent reasonable spread in the item and person estimates.

Variable maps show the mean location for each item based on the item’s total score across all candidates. They do not reveal the location of different response categories or the level of response expected of a person to any item at a given location. The scale structure portraying the lowest and highest response category thresholds for the Entry Survey is presented in Figure 4.

In this variable map, the locations of the candidates are the same as in Figure 3. The column labeled “Mean Item Estimate” places the items in the same locations as shown in Figure 3. The column labeled “1st Threshold for Items,” however, places the items at positions defined as mean item estimate plus the 1st threshold estimate, in this case −1.71. This set of locations shows how probable it is for the candidates to respond with a 2 on the items. The column labeled “4th Threshold for Items” places the items at positions defined as mean item estimate plus the 4th threshold estimate, in this case 2.17. This set of locations show how probable it is for the candidates to respond with a 5 on the items. This map clearly shows that most candidates, at the time they completed the Entry Survey, are in the agree to strongly agree range on the familiar items addressing broad-based issues of equity and diversity. But they are in the uncertain to disagree range on the more focused political action items (recall that a 5 on the reverse-scored items corresponds to a strongly disagree response). This map will serve as a baseline to chart not only progress over time but also the strength of commitment associated with that progress.

The next step was to determine the invariance of the Learning to Teach for Social Justice—Beliefs scale by comparing the results from the two administrations. A close inspection of Figure 3 for the Entry and Exit Surveys reveals a similar order to the item location estimates. The Entry Survey estimates are slightly more bunched, whereas there is a slightly more uniform spread in the Exit Survey estimates. This is reasonable because the graduating students have more experience with these ideas and can make finer distinctions between them. It is also evident that the person locations (measures) are relatively higher on the Exit Survey scale—note the location of the mean candidate estimate (M). Again, this is reasonable because these students have learned these principles as part of their teacher education program.

A plot of the 12 pairs of item estimates from the Entry and Exit Survey solutions provides additional evidence that the structure of the Learning to Teach for Social Justice—Beliefs
<table>
<thead>
<tr>
<th>&quot;Logits&quot; Candidates</th>
<th>1st threshold</th>
<th>Mean Item estimate</th>
<th>4th threshold for Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>#</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ12R</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ11R</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ5R</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ3R</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ10R SJ6R SJ9R</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ2 SJ7 SJ8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>#</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ12R</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ4</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ11R</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ5R</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ3R</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ10R SJ6R SJ9R</td>
</tr>
<tr>
<td>0</td>
<td>.#####</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ2 SJ7 SJ8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ11R</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ5R SJ3R SJ10R SJ6R SJ9R</td>
</tr>
<tr>
<td>-2</td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ2 SJ7 SJ8</td>
</tr>
<tr>
<td>-3</td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SJ1 SJ4</td>
</tr>
</tbody>
</table>

**FIGURE 4**

Variable Map of Candidates, Learning to Teach for Social Justice–Beliefs Scale Items, and Lowest and Highest Response Thresholds: Entry Survey

Note. SJ = Learning to Teach for Social Justice–Beliefs scale; R = reverse-scored item. Each "#" is 4 candidates. Each "." is 1 to 3 candidates.

The scale is invariant across the two administrations (see Figure 5). The ranges and variances are comparable, with the Exit Survey estimates slightly more spread out. The Pearson product–moment correlation is .966. These results support the earlier factor analysis statistics, plots, and congruence coefficients.
Category Characteristic Curves

Category characteristic curves (CCCs) display the probabilities of responding to each category for any person on any item. From a measurement and instrument development perspective, the ideal pattern in the CCCs occurs when each response category is the expected response for some level of person-by-item interaction. This pattern occurs when the threshold estimates follow a strictly increasing order. As seen in Figure 6, the spread in the CCCs and category threshold estimates (where the CCCs intersect) is excellent for the Entry Survey. That is, when a candidate with a relatively low estimate of commitment to teaching for social justice answers an item with a relatively high estimate of commitment to teaching for social justice and the difference is, for example −1.0, the expected response is 2. Likewise, when a candidate with a relatively high estimate responds to a relatively lower level item and the difference in the estimates is, for example 2.0, the expected response is 5.

The pattern for the Exit Survey threshold estimates in Figure 7 is noteworthy because it differs from the Entry Survey estimate’s pattern. The Entry Survey threshold estimates were (−1.70, −0.50, 0.05, 2.20)—this is the desired pattern. The Exit Survey thresholds were (−1.30, −0.07, −0.40, 1.70)—here there is a disordinal pattern in the estimates. The middle
Category Characteristic Curves (CCCs) for the Data From the Learning to Teach for Social Justice–Beliefs Scale Entry Survey

Note. Threshold estimates: (−1.70, −0.50, 0.05, 2.20).

category of uncertain is not the expected response for any level of person-by-item interaction. This difference in the patterns is, however, a reasonable and expected result. It means that incoming students were relatively uncertain about some concepts and ideas when they entered the teacher education program. They may not have had much experience with these issues and may not have thought much about these specific issues. Graduates, in contrast, have spent up to 4 years in a program that is known for its commitment to social justice and have spent a great deal of time addressing these ideas. These graduating teachers are no longer as uncertain about their beliefs regarding learning to teach for social justice.

Fit Analysis

Rasch goodness-of-fit analyses rely principally on standardized residuals—the difference between the observed response and the response expected under the model (Wright & Stone, 1979). A positive residual results when a higher than expected response occurs; negative residuals result from lower than expected responses. Although a variety of statistical and graphical procedures are available for analyzing residuals (Ludlow, 1983, 1986), a standard first approach is to consider summary statistics in the form of standardized, weighted mean squares. Although these statistics do not have exact degrees of freedom and critical values, a rich history of experience has developed regarding their general properties and utility (see, for example, the work of Smith, 1991).

The Rasch goodness-of-fit analyses generally start with the standardized, weighted statistics because they are roughly analogous to t statistics and take into account the variance of the expected response (the so-called ZSTD INFIT in the WINSTEPS [Version 3.64.2] software). A very flexible criterion of +2 to +3 is often used initially. Because the version
Category Characteristic Curves (CCCs) for the Data From the Learning to Teach for Social Justice—Beliefs Scale Exit Survey

Note. Threshold estimates: (~1.30, -0.07, -0.40, 1.70).

of this statistic for items is easily inflated as the sample size grows greater than 500 or so people, the unstandardized version (the so-called mean square INFIT) is also checked. Here a flexible criterion of +1.3 is often used to flag potential problems. These two criteria are generally sufficient to reveal consistent unexpected responses that are either made to an item (item fit) or made by a person (person fit).

A review of the flagged fit statistics and the largest standardized residuals for the data in this study revealed that roughly 12% of the candidates in each administration produced noticeable unexpected responses, whereas only a few items provoked consistent unexpected responses. Although there were a few clear instances of extremely unexpected responses (e.g., a strongly disagree when a strongly agree response was expected), most of the unexpected residual variation occurred as a matter of degree. That is, there were frequent instances in which a strongly agree was expected but only an agree was provided. This was particularly apparent on the Entry Survey, where an agree or disagree response would be expected, but an uncertain response was provided. These situations, however, do not warrant any concern about poor data-to-model fit.

Measuring Change

Given that the purpose of developing the Learning to Teach for Social Justice—Beliefs scale is to measure changes in beliefs, we used the Entry Survey item estimates as anchors to reestimate the candidates' Exit Survey responses. Although this is not strictly necessary because the scale estimates are independent of the person estimates when the data fit the model, it is useful to measure differences (growth) in the entering and graduating teachers by placing them on the baseline scale defined by the incoming students.
Figure 8 contains the anchored estimates for the Exit Survey responses based on the Entry Survey estimates. The difference in the two groups is still evident but the advantage of this procedure is that the scale structure is fixed and we have an unambiguous interpretation of the distributional differences in terms of progress up the scale for respondents. For example, the mean response for the Entry Survey corresponds roughly to disagree with SJ11R: "Whether students succeed in school depends primarily on how hard they work." The mean response for the Exit Survey corresponds roughly to disagree with SJ12R: "Realistically, the job of a teacher is to prepare students for the lives they are likely to lead."

When the anchoring is based on the Entry Survey estimates, the candidates' estimates when they graduate are measures of how far they have gone relative to the Entry Survey.

<table>
<thead>
<tr>
<th>Entry Survey</th>
<th>Exit Survey (Anchored)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n = 268)</td>
<td>(n = 224)</td>
</tr>
<tr>
<td>&quot;Logits&quot;</td>
<td>&quot;Logits&quot;</td>
</tr>
<tr>
<td>Persons</td>
<td>Persons</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>-2</td>
<td>-2</td>
</tr>
</tbody>
</table>

**FIGURE 8**

Anchored Variable Maps for the Data From the Entry and Exit Survey Learning to Teach for Social Justice–Beliefs Scale

Note. SJ = Learning to Teach for Social Justice–Beliefs scale; R = reverse-scored item; M = mean candidate estimate; S = standard deviation of estimates. Each "#" is 3 candidates. Each "*" is 1 candidate.
On the other hand, a reasonable argument can be made that the anchoring should be performed based on the Exit Survey results. This approach would represent the desired outcome and the entering candidates would be measured in terms of how far they have to go (Ludlow, Haley, & Andres, 2005) or the extent to which they have to learn to teach for social justice. Using the Exit Survey estimates for anchoring on is also reasonable strictly from the scale definition perspective because the items at the Exit Survey form a clearer continuum. Both anchoring solutions will likely be applied over the course of the study with future administrations of these two surveys, because they each address a slightly different but meaningful question.

DISCUSSION

The purpose of this article was to explain the development of the Learning to Teach for Social Justice—Beliefs scale. Although the psychometric analyses could have stopped with the CTT results, we performed Rasch rating scale analyses for several important reasons. First, the items were intended to operationally define a unidimensional construct of learning to teach for social justice. This construct, in turn, was intended to define a continuum of questions corresponding to beliefs ranging from a weaker to a stronger commitment to teaching for social justice. Hence, the Rasch analyses served as a confirmatory test of the a priori hypothesized structure. Third, and most important, if the scale was invariant across the two surveys (as defined by similar item difficulty to endorse estimates), then it would be possible to measure how teacher candidates’ beliefs develop and progress from the time they enter their teacher education program and to their subsequent graduation to the classroom.

The Rasch rating scale analyses were of paramount importance in establishing the invariant structure of the Learning to Teach for Social Justice—Beliefs scale over the two groups of teacher candidates. The subsequent establishment of this structure now provides an opportunity to track meaningful positive changes in the perceptions, attitudes, and experiences around the practice of teaching for social justice as Boston College teacher candidates progress through their teacher education program. Success in constructing such a scale not only contributes to the efforts of Boston College to gather evidence of the effectiveness of its BC–TNE initiatives but also constitutes a major contribution to the literature on social justice and its measurement.

Teaching for social justice has been described, for the purpose of the research at Boston College, as a pedagogy that is intended to foster pupil’s learning and to help teachers to understand the social and institutional inequities that are embedded in our society. Through this scale, we have presented the idea that learning to teach for social justice is a measurable outcome of teacher preparation. The results provide evidence that this is indeed a reasonable task. These items are internally consistent, conceptually unidimensional, and define a meaningful and theoretically defensible continuum. Thus, future use of this Learning to Teach for Social Justice scale will provide valuable evidence of the impact of teacher education on learning to teach for social justice.

Based on the differences between the two surveys, we expect that entering students will always be at a similar point along this continuum because we are assuming that each class of incoming students will have similar entering characteristics. We then expect that graduating students will always be at a higher point because they have learned to teach for social justice as a result of their teacher preparation. Furthermore, we expect that as BC–TNE efforts are implemented in the coming years, graduating teachers will be exposed to a measurable BC–TNE program effect. This program effect will be observed through expected increases in each year’s graduating class, as compared with former graduates, regarding their commitment to teaching for social justice.
REFERENCES


APPENDIX A
Entry and Exit Survey Items of the Learning to Teach for Social Justice–Beliefs Scale

Respond to the following statements regarding your beliefs about teaching:

1. An important part of learning to be a teacher is examining one’s own attitudes and beliefs about race, class, gender, disabilities, and sexual orientation.
2. Issues related to racism and inequity should be openly discussed in the classroom.
3R. For the most part, covering multicultural topics is only relevant to certain subject areas, such as social studies and literature.
4. Good teaching incorporates diverse cultures and experiences into classroom lessons and discussions.
5R. The most important goal in working with immigrant children and English language learners is that they assimilate into American society.
6R. It's reasonable for teachers to have lower classroom expectations for students who don’t speak English as their first language.
7. Part of the responsibilities of the teacher is to challenge school arrangements that maintain societal inequities.
8. Teachers should teach students to think critically about government positions and actions.
9R. Economically disadvantaged students have more to gain in schools because they bring less to the classroom.
10R. Although teachers have to appreciate diversity, it's not their job to change society.
11R. Whether students succeed in school depends primarily on how hard they work.
12R. Realistically, the job of a teacher is to prepare students for the lives they are likely to lead.

Note. Likert-type rating scale for responses: 1 = strongly disagree; 2 = disagree; 3 = uncertain; 4 = agree; 5 = strongly agree. R denotes reverse-scored items.
APPENDIX B

WINSTEPS Control File for the 2005 Entry Survey Learning to Teach for Social Justice–Beliefs Scale

: this is a WINSTEPS specification control file
: a semi-colon means a comment
&INST
TITLE = "Social Justice: Entry 2005"
,Input Data Format
NAME1 = 1 ; column of start of person information
NAMLEN = 3 ; maximum length of person information
ITEM1 = 4 ; column of first item-level response
NI = 12 ; number of items = test length
XWIDE = 1 ; number of columns per response
PERSON = Person ; Persons are called ...
ITEM = Item ; Items are called ...
,Data Scoring
CODES = 12345 ; valid response codes
CLFILE = * ; label the categories in Table 3
1 Weak ; 0 in the data means "Strongly Disagree"
5 Strong ; 1 in the data means "Strongly Agree"

,User Scaling
UMEAN = 0 ; item mean - default is 0.00
USCALE = 2 ; measure units - default is 1.00
UDECIM = 2 ; reported decimal places - default is 2
MRANGE = 0 ; half-range on maps - default is 0 (auto-scaled)
&END ; enter the item names
sj1
sj2
sj3
sj4
sj5
sj6
sj7
sj8
sj9
sj10
sj11
sj12
,enter the data
END LABELS
1545534253411
255444442344
.
28343243243111
284544444544333

Note. WINSTEPS is a software package (Wright & Linacre, 1998).
Copyright of Measurement & Evaluation in Counseling & Development is the property of American Counseling Association and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.