International Journal of
Education Policy & Leadership

IJEPL
Volume 16(2)
2020

Get the Most from Your Survey: An Application of Rasch Analysis for Education Leaders

Lauren P. Bailes & Ratna Nandakumar, *University of Delaware*

Abstract

High-quality measurement tools are critical to school improvement efforts. Education researchers frequently employ surveys in order to assess a host of variables associated with school improvement. This article asserts that Rasch modeling techniques enhance the quality of a measurement tool because they comprise elements of both qualitative and quantitative research approaches, and because Rasch modeling corrects the erroneous conclusions that result from the errors associated with ordinal response scale data. This article illustrates, with specific attention to the needs of education leaders and researchers, how the Rasch measurement model gauges the usefulness of survey instruments. This study illustrates the benefits of Rasch modeling using the scale that measures teacher external political efficacy (TEPE). Findings show that a set of four items captures this domain well.

Lauren P. Bailes & Ratna Nandakumar. (2020). Get the Most from Your Survey: An Application of Rasch Analysis for Education Leaders. *International Journal of Education Policy & Leadership 16*(2). URL: http://journals.sfu.ca/ijepl/index.php/ijepl/article/view/851 doi: 10.22230/ijepl.2020v16n2a857

Introduction

Since the advent of the No Child Left Behind Act (2001), education laws have enshrined data-based decision-making as a key tenet, from the level of federal policy to the instructional choices of individual teachers. High-quality measurement tools are thus critical to school improvement efforts. Surveys frequently provide education researchers with tools to efficiently and thoroughly assess a host of variables associated with effective schools (Polikoff, 2014), so designing and deploying surveys are not the exclusive purview of researchers. School leaders, teachers, and other practitioners have been conscripted into the roles of data collectors, analysts, and data-based decision-makers (Galloway & Lesaux, 2014; Hamilton, Halverson, Jackson, Mandinach, Supovitz, Wayman, & Steele, 2009; Marsh & Farrell, 2015). The needs of schools and the pace of decision-making in education mandate a set of research techniques that is both robust and expedient—survey data simple enough to be broadly employed across education stakeholders while also offering a host of nuanced insights to skilled analysts. However, classical testing theory (CTT) demands a lot from survey designers, participants, and analysts: large sample sizes, sample completeness, the uniform perception of the intervals between response items, and some sophistication with regard to reliability and validity. Conversely, Rasch modeling ameliorates the challenges associated with survey research in instructional and policy contexts and offers a way to increase the data capacity of education leaders without the additional burden of extensive statistical or methodological training.

Instructors, leaders, and policymakers may opt for surveys or questionnaires to illuminate the ways in which teachers feel supported, schools engage community stakeholders, principals report their own efficacy for the leadership tasks, and states respond to federal policy initiatives (Berk, 2005; Freiberg, 1998). Surveys are an efficient tool by which to gather a good deal of information quickly. Survey data techniques, then, are built into the fabric of education research and evaluations because data derived from surveys offer powerful insights regarding educational institutions, actors, and leaders. However, survey users—be they instructors, leaders, or policymakers—may be constrained by the limitations associated with survey research, including small sample sizes, incomplete responses, self-reporting bias, and the difficulties of knowing what to do with collected data. So, when survey users ignore assumptions about the limits of survey use and the nature of survey data, those insights are compromised. Consider, for example, a survey that asks about the frequency of family engagement with schools using the response categories never, seldom, occasionally, and always. Thomas Knapp (1990) argues that the central categories (seldom and occasionally) are conceptually much closer than never and seldom. It could even be argued that the two middle categories should be reversed. A respondent may elect to represent his or her participation as seldom despite engaging more frequently than a respondent selecting occasionally, and these respondents' interpretations of the available choices are not likely to emerge in the analyses once the responses are coded as 1, 2, 3, and 4. Assumptions about the ordinal nature of item response options, the equality of intervals between items, and item weight may be particularly pernicious when ignored, especially in the contexts of educational leadership and school improvement. The complexities of using, evaluating, and applying survey tools often fall to education leaders and those charged with directing educational and political organizations. Without a corps of skilled analysts who both value data and use it to steer organizations, the current data-rich education landscape may outpace the education leaders who rely on data-based insights.

This demonstration study, then, is designed for an audience of education practitioner-scholars for whom data-based decision-making and data leadership are priorities (Datnow & Park, 2014). Alex Bowers (2017) posits that graduate training in education leadership prepares candidates for four distinct capacities within what is often a single role: practicing administrator, educational quantitative analyst, research specialist, and data scientist. Courses on quantitative methods within such graduate programs emphasize data literacy and utilization, so training in psychometric techniques is both critical and increasingly common. Additionally, Tommaso, Agasisti, and Bowers (2017) call for further emphasis on psychometric techniques as a sophisticated and necessary skill set for the demands of school leadership and data-informed decision-making across school systems. Therefore, this study assumes some familiarity with psychometric techniques. This article also offers an additional battery of methods for use when issues such as small sample sizes and unequal intervals between answer choices may preclude the use of conventional psychometrics, or that can be used in addition to conventional analysis.

Rasch theory is an established tool for measure development in several sectors of education research. Rasch-developed tools are anchors of both measuring and communicating aspects of student achievement and school improvement. Examples include the Lexile reading levels, the Chicago Consortium's work on school improvement, and the Australian Council for Education Research's system for communicating to parents the nuances of student achievement across content areas (Boone & Scantlebury, 2006; Bryk, Camburn, & Louis, 1999; Stenner, 1996). This article asserts that Rasch modeling techniques enhance the quality of a measurement tool because they comprise elements of both qualitative and quantitative research approaches. Additionally, Rasch modeling corrects the erroneous conclusions that result from the errors associated with ordinal response scale data. The use of Rasch modeling in measure development offers several benefits for researchers, practitioners, and policymakers in educational leadership: shorter and more efficient surveys, less time spent on survey administration, higher-quality survey data, ongoing monitoring of the performance of frequently used measurement scales, and ease of interpretation and clear direction for measure revision or action.

The purpose of this article is to illustrate, with specific attention to the needs of education leaders, practitioners, researchers, and other decision-makers, how the Rasch measurement model gauges the usefulness of survey instruments. Recent research has attended particularly to the needs of education leaders in university preparation programs and suggests that "all decision-makers should be able to apply analytical thinking to the decisions they must make daily. In brief, general administrators should be trained to criticize and utilize analyses, rather than formulate them themselves" (Bowers, 2017, p. 78). This article contends that Rasch modeling must have a place in the lexicon and skill set of educational decision-makers. The techniques demonstrated here provide several advantages over conventional psychometric approaches to measurement and the core skills of CTT, which include exploratory and confirmatory factor analysis. When survey users make incorrect assumptions about the nature of data (especially regarding normality and intervals between answer choices), the resultant analyses may be so erroneous they are unusable. Employing Rasch measurement techniques is one way to ensure that analyses reflect the underlying data realities and are, in turn, applicable to the needs of the organization. Rasch modeling techniques offer an empirical test of construct validity that is sample-independent while still making use of both item and person scores to convert scores from ordinal to interval scales; this is not possible with CTT. Thus, this article is illustrative, not investigative, because demystifying this evaluative tool invites the use of further measurement testing—and substantive improvement—among educators.

The following section details the procedure for analyzing survey data using Rasch analysis. Subsequent to that introduction, it compares CTT and Rasch theory, with specific attention to the advantages of Rasch modeling for educational contexts. Then, for further comparison and illustration, two complete methods sections are provided using the same survey tool: the first examines a scale for teacher external political efficacy (TEPE) using CTT, the second examines the same scale using Rasch theory. The article concludes with a discussion of the scale and, more broadly, highlights the value of Rasch theory for education professionals who use survey data in the service of organizational improvement.

Description of Rasch analysis for survey data

The Rasch model for survey data

The rating scale model (RSM; Andrich, 1978; Wright & Masters, 1982) is one of the Rasch models specifically developed for analyzing data arising from a survey instrument where items follow a Likert-type response and the number of response categories is fixed across items. The RSM describes the log odds of a person (n) choosing the category (k) of item (i) as (Bond & Fox, 2015; Wright & Masters, 1982):

$$\text{logit} = B_n – D_i – F_k$$

In the above equation, B_n denotes a person's ability level, also known as person measure; D_i denotes the difficulty (or agreeability) level of the item, also known as item measure; and F_k denotes the step difficulty (or threshold) level of category k of the item. The ability refers to what is being measured on the survey instrument—in this case, teacher external political efficacy—and the difficulty level of an item refers to the degree of popularity or endorsability of the item. The step difficulty level of an item's category choice refers to the point on the latent scale where the probability of choosing category k and category k-1 are equal. For example, an anxiety item may have four response categories such as, strongly disagree (SD), disagree (D), agree (A), and strongly agree (SA). In this case, there will be three threshold categories. The first threshold is the step from SD to D; the second threshold is the step from D to A; and the third threshold is the step from A to SA. The number of steps (thresholds) is equal to one less than the number of response categories.

In conventional analysis of Likert-type scale data, it is assumed that all items are of similar difficulty and step measures between categories are equidistant. In the

Rasch model, items may differ in difficulty, and step measures need not be of equal distance. For example, an increase in anxiety from "strongly agree" to "disagree" need not be the same as from "disagree" to "agree."

Trevor Bond & Christine Fox (2015) provide a roadmap for conducting Rasch analyses for the purpose of developing high-quality instruments, particularly with respect to reliability and validity. Despite the thoughtful process involved in constructing the items to develop an unambiguous scale and choosing the appropriate number of response categories, a researcher must empirically investigate whether the respondents use the response categories as intended. Therefore, quality-control criteria must be met before fitting the RSM to data. Rating scale diagnostics involves examining the following: 1) item polarity, 2) category frequencies, 3) average measures, 4) step measures, 5) category fit, and 6) probability curves. These terms are described below.

Item polarity

Item polarity is investigated by the point-measure correlation. It is the correlation of the item with the overall measure of the underlying construct. Ideally, these correlations should be positive and moderate to high, because negative or near-zero values indicate that an item is not consistent with the underlying construct.

Category frequencies

Category frequencies comprise the number of respondents who chose each response category. A minimum of 10 responses per category is recommended in order to satisfactorily proceed with Rasch analysis. Categories with low frequencies are problematic because they do not provide enough information to estimate threshold values. Therefore, infrequently used categories indicate redundant categories and should be collapsed into adjacent categories. The shape of the frequency distribution for each category is also important, as regular distributions such as uniform, normal, and slightly skewed are preferred over highly skewed distributions.

Average measures

Average measures refer to the average measure for people in the sample who chose a particular response category. These averages should increase across the rating scale. For example, the average person measure for all those choosing category D should be higher than those choosing the category SD. This is because under the model assumptions, it is presumed that the higher the person measure, the higher the rating on the item.

Step measures (or step calibrations)

Step measures are the intersections of response category functions. They are difficulty estimates for choosing one response category over another (e.g., how difficult is it to choose "strongly agree" over "disagree"). Step measures should increase with category level. They should increase by about 1.4 logits or more (but less than five) in order to show sufficient distinction between categories.

Category fit

The Infit and Outfit measures for each category demonstrate that persons' use of categories is appropriate. Values between 0.6 and 1.4 are considered good for the RSM data (Linacre & Wright, 1994). Values beyond this range indicate noise in the measurement process. Such categories warrant further examination, and may indicate that collapsing adjacent categories would result in a better overall fit.

Probability curves

A probability curve is a visual probability function for each response category. Each response category should be the most probable choice across some region of the latent construct continuum.

All the above diagnostic criteria, when assessed together, provide useful information regarding how to revise a rating scale to increase the reliability and validity of the measure. After quality control criteria have been satisfied, a researcher can then fit the RSM to data and further assess the degree of model fit to data.

WINSTEPS (Linacre & Wright, 2000) is a software program that is widely used for conducting a Rasch analysis of various Rasch models. In addition to providing diagnostic statistics, a WINSTEPS analysis provides model estimates and their standard errors, various reliability estimates, and an item map for examining the construct validity of the test. These are described in detail below.

Item and person fit indices

The item fit index provides an indication of how well an item contributes to the construct being measured by the test in a meaningful manner. Fit statistics are also useful in assessing the unidimensionality of data. The model assumes that an item has a greater probability of yielding a higher rating for a person with higher ability than a person with lower ability (Smith, Conrad, Chang, & Piazza, 2002). Similarly, a person has a higher probability of responding to an easier item than a relatively more difficult item. Fit indices indicate how well an item conforms to the model assumptions. WINSTEPS provides two types of fit indices for persons and items: Infit and Outfit. The Infit item measure is more sensitive to unexpected responses of persons close to the item difficulty estimate, whereas the Outfit measure is sensitive to outliers. For details about these fit statistics refer to Benjamin Wright and Mark Stone (1979). Infit and Outfit values, when expressed as mean-square statistics, have an expected value of one (Wright & Linacre, 1994). Values much lower than one indicate a lack of adequate variability in the data. Values much greater than one indicate excessive variability. Values ranging between 0.6 and 1.4 are considered good fit for self-reporting RSM data (Smith et al., 2002; Wright & Linacre, 1994).

Person-item map

The Rasch model converts raw measures of item agreeability and person scores into interval measures in logit units. These logit measures for items and persons are used in constructing the person-item map. Item measures are plotted from easiest (most agreeable) at the bottom to most difficult (least agreeable) at the top. Person measures are plotted from least able (least agreement with items) at the bottom to most able (most agreement with items) at the top. This map is a very useful tool to investigate the effectiveness and utility of the instrument to the sample, including construct validity as described below.

Distribution

A person distribution is expected to be normally distributed or skewed in one direction. In an item distribution, items are expected to be uniformly distributed (comparable to marks on a ruler) or clustered.

Targeting

Is the test too easy or too hard for the sample? This question is examined by contrasting person and item distributions. For example, one might examine the mean performance of persons relative to the mean of items.

Predictive validity

This is applicable to the person distribution. Are the people ordered as we would expect based on other information about them? For example, we expect healthier or more educated people to have higher measures of a certain construct. Predictive validity indicates whether or not those expectations are borne out by the data.

Construct validity

This is applicable to item distribution and is an assessment of whether the items are ordered as we would expect based on the intended measurement. For example, in an arithmetic test, is division generally more difficult than addition? Or, in a scale measuring independent living, is climbing stairs more difficult than preparing food?

Test construction comparison

This article now turns to a comparison of CTT and Rasch modeling with regard to test construction. In a typical survey or test development, the first step is to define the construct underlying the test items (note that most survey instruments measure a single construct). The next step is to construct an initial pool of items, usually at least twice the number of items aimed at the final version of the test. Items should range on the latent construct from lower levels to higher levels (least difficult to most difficult to endorse). At this stage, content validity checks of items (Furr & Bacharach, 2014) must be carried out through extensive reviews and revisions. Next, pilot data are collected on a representative sample.

In conducting data analysis to empirically validate the items, the two approaches—CTT and Rasch modeling—differ in detail, as shown in Table 1. Rasch analysis provides more thorough and rich information that goes beyond what CTT provides. For instance, the person-item map may serve as a valuable tool in reviewing and revising items in terms of targeting the sample. In addition, Rasch measurement is robust against incomplete or missing data, provides fit measures for items and persons, and provides standard errors for items and persons. A Rasch analysis can yield a survey instrument that has fewer well-targeted items than CTT, yet is highly reliable and valid. Furthermore, item and person measures are linear, comprise an interval scale, and are sample- and test-independent. Table 1 presents a comparison of CTT and Rasch modeling processes (Wright, 1992).

Table 1. Comparison of classical test theory (CTT) and Rasch model (RM)

Characteristic	CTT	RM
Data	Must be "complete"	Robust against missing data
Sample size	The larger the better	Reliable estimates can be obtained with as few as 30 subjects
Number of items	The larger the better	A smaller number of targeted items can provide more reliable measures than a larger number of items
Item analysis	Sample-dependent item difficulty and corrected item-total correlation	Sample-free item measures, test-free person measures, point-measure correlation, individual item- and person-fit statistics
Reliability	Cronbach's alpha coefficient	Person reliability index, item reliability index, item separation index, and person separation index
Construct validity	Factor analysis	Item-person map Meaning of scale
Ordinal ranking	Linear positioning on the construct that is	explicitly defined by the item content Additivity
Non-linear	Linear	Precision
Average precision over	all persons	Quantified by standard errors for all persons and all items
Accuracy	Unknown	Quantified by fit statistics
Analysis	Unsuited for statistical analysis	Ideal for statistical analysis

Advantages of Rasch modeling

Rasch modeling need not replace other psychometric methods used to investigate the construct validity of an instrument, such as exploratory factor analysis (EFA) and confirmatory factor analysis (CFA), but may instead augment them. For example, EFA can be utilized to investigate the degree to which the items in the tool tend to measure the same latent construct. It does not, however, measure the ease or difficulty with which individuals in the sample can endorse items within the measurement scale. That is, it does not give the researcher information about how participants who have high levels of the underlying construct affirm or endorse different items than those who have lower levels of the construct. Further, EFA does not indicate at which end of the endorsability spectrum the scale may lack items that, if included, may better measure the latent construct. Rasch measurement assesses both the difficulty of an individual item and the capability (ability level) of a person on the underlying construct based on his or her responses to those items. The researcher, then, has several tools that would not be available using other psychometric techniques. Specifically, Rasch techniques offer opportunities "to test [whether] the items form a unidimensional variable (by examining statistically idiosyncratic responses), to calibrate the magnitude of differences among items on an interval scale, and measure each person on the newly created variable" (Fox & Jones, 1998, p. 30).

The development of a survey instrument generally involves item analysis (such as item difficulty or item popularity, item-total correlation), a reliability coefficient such as a Cronbach's alpha, and factor analyses. Classical psychometric techniques such as exploratory and confirmatory factor analysis are sample dependent. That is, item indices such as validity and reliability depend heavily upon the sample the data were collected from. The degree of homogeneity or heterogeneity in the sample can cause significant differences in the item indices. Similarly, person scores, usually obtained by summing scores or calculating means across items, differ depending on the difficulty (or popularity) of items and even the number of response items used on the survey. Rasch measures for items and persons are sample free. Item measures, as a result, are valid beyond the particular sample at hand and person measures are valid beyond the particular set of items used.

The most common approaches to measurement and scale development in education (e.g., EFA, CFA) do not address the underlying assumptions of parametric analyses. Assumptions include approximately equal intervals, random sampling from a defined population, and independent samples. The research is clear: educational settings tend to violate these assumptions. For example, teachers and students in a single school are more likely to be similar to each other than they are to be similar to their counterparts in another school. Further, the distribution of some characteristics—such as dropout rates—tend to cluster in schools rather than distribute equally across organizations. Parametric analyses, then, are limited in that they require data be measured on an equal interval scale. However, Likert-type items are not truly measured on an interval scale. Rasch modeling, addresses these and other issues of measurement because it converts data into an interval scale and makes it suitable for statistical analyses. It provides a common metric for items and persons, which is not a feature of conventional psychometric techniques for measure development and use. Finally, item and person measures can be located as points on the scale that define the underlying trait of the continuum.

With regard to reliability, in addition to Cronbach's alpha, Rasch modeling provides person and item separation indices. Rasch analysis provides measures for each item and each person. The person measure denotes each person's perceived attitude toward the underlying trait measured by the items. Person fit further indicates whether the person's responses to items are consistent with the model. The item measure denotes the item's position on the underlying trait continuum. Item fit indices also indicate whether or not the item fits well with the rest of the items in defining the underlying trait measured by the items. Item and person estimates are determined such that they maximize the fit of the data to the model. The person separation index is another index of reliability. It adjusts the standard deviation for the measurement error in computing the reliability and denotes the ability of a test to discriminate among various categories of person abilities.

There are further advantages to these person and item fit indices that are not present in factor analyses: fit indices also allow identification of socially desirable responses to items by checking whether the fit index for an item is too good. Finally, Rasch output includes an item-person map which juxtaposes items and persons (as they are on the same scale) and allows researchers to investigate the construct validity. This also provides a means of examining and assessing the utility of items for the sample and how well items are targeted to the sample. Gaps in the scale are evident and may be remediated by adding items to the measurement tool (Bond & Fox, 2015).

Analytic method: Classical test theory

This section describes the instrumentation, data collection protocol, sample, and primary analytic method employed in our demonstration of classical test theory techniques for survey analysis.

Instrument

Campbell, Gurin, and Miller (1954) were the first to use the term "political efficacy." (p. 187) Political efficacy was originally conceptualized as a multidimensional construct measured by a single, four-item scale. As the construct underwent a series of theoretical evolutions and validity studies, Robert Lane (1959) determined that it comprised two dimensions: an internal and an external dimension to feelings of political effectiveness. He defined the external dimension as, "the image of democratic government as responsive to the people" (p. 145) and the internal dimension as one's own ability to marshal and deploy the skills for political engagement. Political efficacy has been empirically associated with a host of political beliefs and behaviors, including such desirable outcomes as democratic consensus-building practices and political empowerment in some marginalized populations (see, for example, Emig, Hesse, & Fisher, 1996). In the same way that beliefs about political efficacy account for individuals' beliefs about the public domain and public officials, the concept of teacher political efficacy may help educators better understand constituents' beliefs about school systems and about the personnel involved in governance and decision-making. This emergent tool (Bailes, 2016), which measures external political efficacy among educators, hails from research traditions in both political science and social cognitive theory. The current study uses only the teacher external political efficacy (TEPE) scale. This tool assesses the ways in which teachers feel sufficiently skilled to interact with education policy systems (see Table 4 for the items that comprise TEPE).

Data collection

Data for this study were collected in traditional public schools in an urban district of a Midwestern state and in public charter schools that serve as alternatives to the traditional public school system in the metro area. Schools were recruited if they included grades 3, 4, and 5. Of the 106 schools recruited, 53 participated (response rate = 50%). Individual teachers were eligible to participate in the survey if they were at least 0.5 full-time employees (FTE) and taught in academic content areas. Response rates at the level of individual teachers were not calculated because researchers were not given access to employment records at each school in the sample. The survey tool used in this study assessed a diverse array of constructs; subscales were divided evenly between two forms (Form A and Form B) so that adjacent respondents did not use the same form. Participating schools were given a $50 gift card in return for their cooperation.

Teacher political efficacy was comprised of nine items; four items assessed external political efficacy (TEPE) and five items measured internal political efficacy (TIPE). Teachers reported their political efficacy using a Likert-type scale scored from one to six (1 = strongly disagree, 6 = strongly agree). Only data measuring TEPE (Form A) are used in this analysis.

Sample

Of the 53 schools that participated in the study, three schools had to be dropped (one had to be dropped because the students were exclusively in middle grades; one because none of the teachers responded to the same form as the items of interest to the current study, namely Form A; and one because it did not report third grade scores). This left a total sample of 50 public elementary schools. From the 50 schools, a total of 412 teachers, all of whom were at least .5 FTE, responded to Form A of the survey. Note that all school-level variables were standardized (M = 0, SD = 1). Descriptive statistics for teacher- and school-level variables are detailed in Table 2. Of the 412 participants, approximately 96 percent of the teachers were female, and the majority were white (93%). Most teachers attained a master's degree (57.9%) and have been in their current school between zero and three years (57.7%).

Table 2: Descriptive statistics for the total score on teacher external political efficacy

Level 1	*N*	Minimum	Maximum	Mean	Standard deviation
Total TEPE	412	−1.84	2.38	.00	1.00
Gender (females)	396	.00	1.00	.85	.36
Minority status (whites)	384	.00	1.00	.21	.40

Exploratory factor analyses and internal consistency

Exploratory factor analysis was performed on data from TEPE (four items, n = 394) to determine the underlying structure of the data. Three criteria were used to determine the number of components to retain: scree plot, the number of eigenvalues greater than one, and parallel analysis. All three criteria suggested retaining one component. Following this, principal axis factor analysis was conducted to fit a one-factor model to the data and extract factor loadings. Factor loadings of the TEPE scale are reported in Appendix A. The internal consistency reliability coefficient, Cronbach's alpha, for the four-item TEPE scale was 0.87. These results are consistent with the results obtained in the initial analysis of scale development of Thompson (1994) as well as with Bailes's (2016) development of the educator-specific tool.

Analytic method: Rasch analysis

This article now turns to an illustration of survey analysis using Rasch theory and techniques.

Diagnostic analysis

The WINSTEPS software program (Linacre & Wright, 2000) was utilized to conduct a Rasch analysis of TEPE data using the rating scale model (RSM; Andrich, 1978). Each TEPE item was measured on a six-point Likert scale where 1 denotes strongly disagree (SD), 2 denotes disagree (D), 3 denotes somewhat disagree (SWD), 4 denotes somewhat agree (SWA), 5 denotes agree (A), and 6 denotes strongly agree (SA).

WINSTEPS results produce estimates of person measure (perceived Teacher External Political Efficacy) for each teacher, and item measures (item endorsibility) for each item. In addition, results include fit indices for items and persons, and information on reliability and the construct validity of the TEPE scale.

As a first step, diagnostic tools were analyzed to assist in the identification of potential problems with the functioning of the rating scale. As described in the previous section, these include point-measure correlations, evaluating category counts, average measures, step measures for the transition between adjacent categories, and the category Infit and Outfit mean-square statistics. The diagnostic information regarding functioning of the six-point scale is presented in Tables 3 and 4 and in Figure 1.

Table 3: Category counts, average measures, Infit mean-square statistics, and step measures for the six-point scale

Category label	Category meaning	Total count	Average measure	Infit MNSQ	Outfit MNSQ	Step measures
1	Strongly disagree	229	–3.31	1.05	1.02	None
2	Disagree	359	–1.98	0.88	0.84	–3.37
3	Somewhat disagree	371	–0.69	0.82	0.82	–1.35
4	Somewhat agree	324	0.46	0.95	0.96	0.00
5	Agree	245	1.66	1.13	1.20	1.27
6	Strongly agree	98	2.86	1.39	1.28	3.45

Table 3 shows diagnostic information for item categories. The Total Count column lists observed counts for each category. The minimum count (98) for category 6 is far higher than the required minimum of 10 responses. As expected, the average measures for each category listed in the Average Measure column are ordered, progressing from -3.31 logits for the lowest category to 2.86 logits for the highest category. The Infit and Outfit mean square statistics for all categories are within the expected range of 0.6 to 1.4, indicating that all response categories are functioning appropriately.

Table 4: Items statistics: measure, Infit and Outfit mean squares (MNSQs), and point-measure correlation

Item description	Item measure	Infit MNSQ	Outfit MNSQ	Point-measure correlation
I think school leaders generally care about what teachers like me think.	–1.17	1.25	1.22	0.81
I know education policymakers share my concerns about schools.	0.44	0.97	0.95	0.84
In general, people who run our schools consider my best interest when they make decisions.	0.77	0.87	0.89	0.84
I generally assume school politics are fair and school politicians try to do the best by teachers.	–0.04	0.88	0.88	0.85

Figure 1 Response category probability curves

```
P      -+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-
R  1.0 +                                                             +
O      |                                                             |
B      |                                                             |
A      |1                                                            |
B   .8 + 11                                                        66+
I      |   1                                                     66  |
L      |    11                                                  6    |
I      |      1                                                6     |
T   .6 +       1                                   5555      66      +
Y      |        1   2222222                      55    55   6        |
    .5 +         1 2       22                   5        556         +
O      |         2*          2 333333   44444455          655        |
F   .4 +        2  1         3*      3*4     54          6   5       +
       |      22    1       3  2     4 3    5  44       6     5      |
R      |     2       11   33    22 44   33 5     4     6       55    |
E      |   22          1 3        *       *       44 66          55  |
S   .2 + 22            3*        4 2     5 3        *              5 +
P      |2             3  11    44   22 55   33    66 44             5|
O      |           333     1144      5*2      3366     444           |
N      |       3333      44441111 555   222 66663333      4444       |
S   .0 +*************************************************************+
E      -+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-
       -5    -4    -3    -2    -1     0     1     2     3     4     5
        Person [MINUS] Item MEASURE
```

Figure 1 shows the category response curves for each of the six categories of the survey. The x-axis represents the ability and the y-axis represents the probability of endorsing an item-category. There are six response curves corresponding to each of the six response categories of items on the survey. For example, the 2s correspond to the second response category, "Disagree." This bell-shaped curve of 2s peaks in the ability range from -3.4 to -1.4. This means, on average, teachers with TEPE measures between -3.4 to -1.4 are more likely to endorse this category. Similarly, teachers with TEPE values between 1.3 and 3.4 are more likely to endorse category 5, "Agree." What is important to observe is that each response category is the highest preferred choice across some region of the ability scale (TEPE measures). This is important because category choices should be well defined and mutually exclusive (Linacre, 1999). Based on Figure 1, one can conclude that all six item categories (SD, D, SWD, SWA, A, SA) are being utilized by the respondents. If this were not the case (Bond & Fox, 2015, see Figure 11.2, p. 254), one would want to collapse the redundant categories and reanalyze until the data exhibits mutually exclusive item categories and thereby improves the test reliability (Linacre, 1999; Wright & Stone, 1979; Updyke & Lewandowski, 1997).

This information, while technical, offers value to practitioner-analysts in at least two ways. First, it alerts them to overlapping response categories. If, returning to the above example of TEPE, respondents conflate the categories somewhat disagree and disagree, then practitioners are unable to distinguish the needs of respondents in those categories. For a host of surveys important to survey users in schools, survey parsimony is desirable. A diagnostic tool like that pictured in Figure 1 may result in a greater degree of survey precision. Further, Figure 1 provides visual information regarding how respondents with varying degrees of the measured construct are most likely to respond. Again, considering the earlier example: practitioner-analysts may identify the intersection of a respondent who is highly likely to demonstrate external political efficacy and the probability that the same respondent will select a specific response option. Because each response category is the highest preferred choice across some region of the ability scale—in this case, the ability of a teacher to report external political efficacy beliefs—those analyzing the survey are able to determine the region of the ability scale at which individuals endorse a different response option. The resultant intervention also becomes clearer: increasing a teacher's external political efficacy is equivalent to moving that respondent from endorsing one region of the survey range to endorsing another region of the survey range.

The last column on Table 3 shows step measures. These values denote an increase in TEPE from one category to the next. For example, the increase in TEPE measure required to go from SD to D is 3.4 units, whereas the increase in TEPE measure required to go from D to SWD is only about 2 units (3.37-1.35). Hence, the distances between the categories are not equally spaced.

Figure 2. Item-person map of external political efficacy

```
------------------------------------
 MEASURE    Person - MAP - Item
               <more>|<rare>
    5            .#  +
                  #  |
                     |
                  .  |
                  .  |
                     |
    4                +
                .##  |
                     |
                     |
                    T|
                  .  |
    3                +
                  .  |
                     |
               ####  |
                     |
                     |
    2       .######  +
                  .  |
                     |
            ####### S|T
                  .  |
            .######  |
    1                +
                     |  TEPE4_13
         ##########  |S
                     |  TEPE2_11
          #########  |
                     |
    0   .##########  +M TEPE3_12
                     |
          #########  |
                    M|
                     |S
         .#########  |
   -1             .  +
          .########  |  TEPE1_10
                     |
        .##########  |T
                     |
                     |
   -2    .#########  +
                  .  |
                    S|
         .#########  |
                  .  |
                  .  |
   -3         .####  +
                     |
                  .  |
                     |
                     |
             ######  |
   -4                +
                     |
                    T|
                     |
                .##  |
                     |
   -5        .#####  +
               <less>|<frequent>

EACH "#" IS 3: EACH "." IS 1 TO 2
```

The point-measure correlations of all four items on the TEPE scale are listed in the last column of Table 4. All the correlations are high positive values ranging from .81 to .85, indicating that all items contribute to the underlying trait of teachers' external political efficacy.

Rasch assessment of item fit and reliability

Following the diagnostics analysis, the fit of the RSM model to data was examined. The assumption unidimensionality was investigated using the item fit statistics. Table 4 shows item measures and item fit statistics. It can be seen that for all items both Infit and Outfit values were within the range of 0.6 to 1.4, providing support for a unidimensional construct of TEPE. Item measures ranged from -1.17 to 0.77, suggesting that items are catering to individuals in the middle range of the scale (see Figure 2). The person reliability estimate (conceptually equivalent to Cronbach's alpha) was 0.87, the same as that obtained by conventional analysis. The person separation index was 2.1, indicating that items are able to discriminate persons along the scale of external political efficacy.

Rasch construct validity evaluation

The item-person map displayed in Figure 2 demonstrates the construct validity of the TEPE survey instrument in the sense of how well the items of the survey define the construct of external political efficacy and the representation of items across the ability continuum, as described below.

A vertical line separates person and item measures. Person measures, denoted by "#", are placed to the left of the vertical line and item measures along with item names are placed to the right of the vertical line. The M, S, and T on the left and right sides of the vertical line represent the mean, one standard deviation from the mean, and two standard deviations from the mean for persons and items respectively. Those items that are easier to agree with (or endorse) have lower values of item measure than items that are more difficult to endorse. Hence, items are arranged from bottom to top in decreasing order of endorsibility. The topmost item is the most difficult to endorse, and the bottommost item is the easiest to endorse. Persons, on the other hand, are arranged from least able to most able. Persons at the top of the line highly endorse all items, denoting that they agree with more items on the survey and, in turn, demonstrate more of the construct being measured (TEPE). Persons at the bottom of the line, relatively speaking, exhibit a lower endorsement of items and exhibit a lower TEPE measure. For example, item TEPE4_13 (I generally assume school politics are fair and school politicians try to do the best by teachers) is the most difficult to agree with, while the item, TEPE1_10 (I think school leaders generally care about what teachers like me think) is the easiest to endorse (see Table 4).

The item-person map indicates the order of items in terms of "difficulty" or endorsibility, as they empirically define the construct. It also indicates possible improvements and further refinements to the scale. Items are bunched in the middle of the scale, while respondents are spread across the ability scale from -5 to +5. This illustrates that this survey is good at accurately assessing the levels of TEPE for teachers whose measures are in the range of about one standard deviation from the mean. There are many teachers whose measures are well above this range. That is, these teachers' level of TEPE is much higher than can be measured with this set of items. Their level of TEPE cannot be accurately measured by this set of items. More items are needed at the top of the scale that can elicit higher levels of endorsability than TEPE4_13. Similarly, there are many teachers whose TEPE levels are much below the range of item measures. These teachers' levels of TEPE also cannot be accurately measured with these items. More items are needed that are easier to endorse than item TEPE1_10. A practitioner-analyst reviewing the results of this diagnostic tool would be advised to add more items that are more difficult to endorse in order to identify the most teachers who report the highest level of external political efficacy beliefs—those which are higher than the current survey can accurately measure. By doing so, the survey analyst not only identifies the teachers with the desired characteristics but may also observe the individuals who report high levels of the desired construct in order to intervene among individuals who report low levels of TEPE. These additional items would increase the representativeness of the content domain covered and improves measurement precision (Smith et al., 2002).

Summary and conclusion

Findings of the demonstration study showed that the TEPE construct is well captured by the set of four items. All four items significantly contribute to the scale, providing good reliability. These items differ in their level of endorsibility. As indicated on the item-map, some items are easier to endorse for this sample than others. Moreover, these items are able to discriminate teachers in the mid-range in their perception of TEPE, but not on either ends of the scale. The scale could be improved by the addition of items at both ends of the difficulty spectrum. All four of the items hover around the mean and extend only slightly beyond one standard deviation. In order to better discriminate among respondents and better understand the function of political efficacy among educators, the measure requires a balance of items that are both easy to endorse and hard to endorse. The construct validity of the scale is improved by having items spread across the scale according to their level of endorsability in order to differentiate teachers with differing levels of TEPE.

In this demonstration study, TEPE provides a useful illustration of the ways in which classical test theory and Rasch analysis differ. The scale used here as an example is emblematic of a host of perception measures that are used in education to assess such outcomes as school quality, stakeholder engagement, experiences of professional development, and equity. More broadly, this demonstration study illustrates the ways in which Rasch techniques provide scholars, practitioners, and policymakers with an additional tool for survey development. Throughout a Rasch analysis process, researchers have the opportunity to learn about and refine an instrument in ways not necessarily available in classical psychometric techniques. Because Rasch modeling is sample independent, it serves as an especially good alternative for those who are developing instruments in the field to improve their own organizations. Provided that users obtain 12–30 responses per response option, they do not need to be concerned with "completeness," power analysis, or surveying a sufficiently large sample (Makoul, Krupat, & Chang, 2007). School leaders and policymakers might find the Rasch method, then, particularly useful because they can rapidly develop instruments that are useful to their organizations or contexts without the costs associated with large-scale data collection.

Rasch analysis invites and values elements of both qualitative and quantitative approaches to data analysis. Researchers must thoughtfully consider the array of items in a given measurement tool and the degree to which the set of items completely captures the construct of interest. Additionally, judgment-based scales (such as TEPE, used in this example) are notoriously prone to be subject to poor measurement because, while they are typically ordinal, respondents do not interpret the original choices as separated by equal intervals. Items written too generally may not capture the extent of respondents' beliefs, and so relationships among constructs may be misrepresented. Following the application of Rasch analysis, an iterative process is initiated in which person and item measures indicates the ease or difficulty of the items and, in turn, whether the measure requires revisions. Rasch techniques, unlike the tools available in classical test theory, suggest to practitioner-scholars the precise revisions necessary to best understand the fullest range of the construct.

The purpose of this study was to illustrate Rasch measurement techniques to educational leaders with the specific aim to enhance the construct validity of a survey instrument, and convert ordinal scores into interval scores that are sample- and item-independent. In this regard, the principles underlying a Rasch model are described and its application to survey instruments using the rating scale model is illustrated for the TEPE scale. As demonstrated here, Rasch techniques provide educational leaders an additional battery of testing and measurement tools that augment or supersede those available to them through more conventional measurement techniques. Education leaders often need data rapidly and thus need to create tools, collect data, conduct analyses, and determine organizational direction quickly. Rasch techniques allow them to do this: leaders may draft survey items without undue concern for establishing the intervals between items, and data collection need not represent completeness in the sample. Moreover, educational leaders who opt for Rasch analyses need not collect or use data that violate the assumptions of parametric analyses. Finally, education leaders who employ Rasch techniques will be more skilled navigators of the data-rich environments that currently typify schools and schooling and, therefore, more able to lead organizations to beneficial actions as a result of understanding those environments. Ultimately, education leaders are compelled to choose tools that allow them to lead organizations in ways that support organizational improvement and student achievement. Rasch techniques provide leaders with the speed, efficiency, and precision they need to make the best decisions with the fewest errors and the most confidence.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573.

Agasisti, T., & Bowers, A.J. (2017). Data analytics and decision making in education: Towards the educational data Scientist as a key actor in schools and higher education institutions. In *Handbook of contemporary education economics* (pp. 184–210). Northampton, MA: Edward Elgar Publishing.

Bailes, L. (2016). *A theoretical and empirical investigation into the meaning and measure of political efficacy and its application to education.* (Electronic thesis or dissertation). Retrieved January 2, 2020, from https://etd.ohiolink.edu/

Berk, R.A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17*(1), 48–62.

Bond, T.G., & Fox, C.M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences.* Hove, UK: Psychology Press.

Boone, W.J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple choice tests. *Science Education, 90*(2), 253–269.

Bowers, A.J. (2017). Quantitative research methods training in education leadership and administration preparation programs as disciplined inquiry for building school improvement capacity. *Journal of Research on Leadership Education, 12*(1), 72–96.

Bryk, A., Camburn, E., & Louis, K.S. (1999). Professional community in Chicago elementary schools: Facilitating factors and organizational consequences. *Educational Administration Quarterly, 35*(5), 751–781.

Campbell, A., Gurin, G., & Miller, W. (1954). *The voter decides.* Evanston, IL: Row, Peterson.

Datnow, A., & Park, V. (2014). *Data-driven leadership.* San Francisco, CA: John Wiley & Sons.

Emig, A.G., Hesse, M.B., & Fisher, S.H. (1996). Black-White differences in political efficacy, trust, and sociopolitical participation: A critique of the empowerment hypothesis. *Urban Affairs Review, 32*(2), 264–276.

Freiberg, H.J. (1998). Measuring school climate: Let me count the ways. *Educational Leadership, 56*(1), 22–26.

Furr, R.M., & Bacharach, V.R. (2014). *Psychometrics: An introduction* (2nd ed.). Thousand Oaks, CA: Sage Publishing.

Fox, C.M., & Jones, J.A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology, 45*(1), 30–45.

Galloway, E.P., & Lesaux, N.K. (2014). Leader, teacher, diagnostician, colleague, and change agent: A synthesis of the research on the role of the reading specialist in this era of RTI-based literacy reform. *The Reading Teacher, 67*(7), 517–526.

Hamilton, L., Halverson, R., Jackson, S.S., Mandinach, E., Supovitz, J.A., Wayman, J.C., & Steele, J.L. (2009). *Using student achievement data to support instructional decision making.* United States Department of Education. Retrieved January 2, 2020, from http://repository.upenn.edu/gse_pubs/279

Knapp, T.R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nursing Research, 39*(2), 121–123.

Lane, R.E. (1959). *Political life: Why people get involved in politics.* New York, NY: MacMillan.

Linacre, J.M., & Wright, B.D. (2000). WINSTEPS: Multiple-choice, rating scale, and partial credit Rasch analysis [Computer software]. Chicago, IL: MESA.

Linacre, J.M., & Wright, B.D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions, 8*(2), 350.

Makoul, G., Krupat, E., & Chang, C.H. (2007). Measuring patient views of physician communication skills: Development and testing of the Communication Assessment Tool. *Patient Education and Counseling, 67*(3), 333–342.

Marsh, J.A., & Farrell, C.C. (2015). How leaders can support teachers with data-driven decision making: A framework for understanding capacity building. *Educational Management Administration & Leadership, 43*(2), 269–289.

Polikoff, M.S. (2014). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education, 121*(2), 183–212.

Roderick, M. (2012). Drowning in data but thirsty for analysis. *Teachers College Record, 114*(11), 1–9.

Smith, E.V., Conrad, K.M., Chang, K., & Piazza, J. (2002). An introduction to Rasch measurement for scale development and person assessment. *Journal of Nursing Measurement, 10*(3), 189–206.

Stenner, A.J. (2006). *Measuring reading comprehension with the Lexile framework.* Durham, NC: Metametrics, Inc. Paper presented at the California Comparability Symposium, October 1996. Retrieved January 2, 2020, from http://www.lexile.com/DesktopDefault.aspx?view=re

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54,* 837–847.

Updyke, W. F., & Lewandowski, C. (1997). Post-hoc Rasch analysis of optimal categorization of an ordered-response scale. *Journal of Outcome Measurement, 1*(4), 286–304.

Wright, B.D. (1992). Scores are not measures. Rasch Measurement: Transactions of the Rasch Measurement SIG, *American Educational Research Association 6*(1), 208.

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

Wright, B.D., & Stone, M.H. (1979). *Best test design.* Chicago, IL: Mesa Press.

Appendix
Exploratory factor analysis for TEPE scale

Item description	Factor loadings
I think school leaders generally care about what teachers like me think.	0.712
I know education policymakers share my concerns about schools.	0.810
In general, people who run our schools consider my best interest when they make decisions.	0.835
I generally assume school politics are fair and school politicians try to do the best by teachers.	0.816