# Rasch Analysis for Instrument Development: Why, When, and How?

**William J. Boone***
Department of Educational Psychology, Miami University, Oxford, OH 45056

## ABSTRACT

**This essay describes Rasch analysis psychometric techniques and how such techniques can be used by life sciences education researchers to guide the development and use of surveys and tests. Specifically, Rasch techniques can be used to document and evaluate the measurement functioning of such instruments. Rasch techniques also allow researchers to construct "Wright maps" to explain the meaning of a test score or survey score and develop alternative forms of tests and surveys. Rasch techniques provide a mechanism by which the quality of life sciences–related tests and surveys can be optimized and the techniques can be used to provide a context (e.g., what topics a student has mastered) when explaining test and survey results.**

## INTRODUCTION

A range of statistical techniques such as factor analysis, calculation of Cronbach's alpha, point biserial correlations, and computing a raw score total are commonly used to develop instruments (tests, surveys) for educational research. These approaches have been used to evaluate the strength of the inferences drawn from instruments and to compute respondents' (e.g., student, teacher) performances. Rasch analysis is a psychometric technique that was developed to improve the precision with which researchers construct instruments, monitor instrument quality, and compute respondents' performances. Rasch analysis allows researchers to construct alternative forms of measurement instruments, which opens the door to altering an instrument in light of student growth and change. Rasch analysis also helps researchers think in more sophisticated ways with respect to the constructs (variables) they wish to measure. Some life sciences education researchers are already using Rasch techniques (e.g., Reeves and Marbach-Ad, 2016), but many continue to use instrument development and validation approaches that rely on classical test theory.

The purpose of this article is to provide a brief introduction to selected whys, whens, and hows of using Rasch techniques so that Rasch techniques become more widely used in the life sciences education research community. I start by briefly introducing the importance of carefully measuring with a test or survey and outlining the mathematical errors common to test and survey analysis conducted using non-Rasch techniques, which can be avoided by using Rasch analysis. I then describe quality-control steps inherent to Rasch that can improve the quality of measurement instruments. I conclude by explaining how to use Rasch techniques to better communicate research findings and outlining the steps that should be taken to develop different forms of a test.

## PROBLEMS WITH THE ANALYSIS OF TEST DATA AND SURVEY DATA

To appreciate the importance of Rasch techniques, we first need to think about what it means to measure a variable, such as the knowledge of a student or the attitude of a teacher. A researcher must begin by defining the single variable to be measured. Consider a concrete example of measuring the height of a flower, which can be measured along the continuum of a meter stick (Figure 1). By focusing on measuring only

**Erin Dolan,** *Monitoring Editor*
Submitted April 3, 2016; Revised September 13, 2016; Accepted September 19, 2016
CBE Life Sci Educ December 1, 2016 15:rm4
DOI:10.1187/cbe.16-04-0148

*Address correspondence to: William J. Boone (boonewjd@gmail.com).

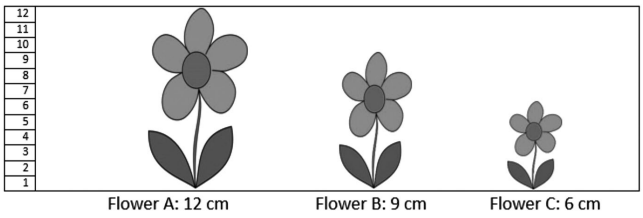


**FIGURE 1. Thinking about linear measurement. A meter stick being used to make linear measures and compare the height of three flowers.**

one variable, a researcher can make comparisons with confidence. For example, how do the heights of flowers A, B, and C in Figure 1 compare? Without a carefully developed measurement instrument that captures the parameters of one variable and one variable only, it is very difficult if not impossible to make meaningful comparisons. Another strength of a meter stick is its linear scale. This means that, if the difference between the height of flower A and the height of flower B is 3 centimeters, and the difference between the height of flower A and the height of flower C is 6 centimeters, an observer can confidently state that the ratio of the differences in height is 1:2. If the scale is not linear, then an observer could not make such an assertion. The concept of linearity is one of the most fundamental ideas for understanding why Rasch theory is an important tool for researchers.

It is tempting to use raw survey and test data immediately, because there is so much linear data that researchers can immediately manipulate with simple mathematics. For example, the difference in running times between four runners can be confidently compared, the costs of six houses can be confidently compared, and so forth, because time and money are both linear. Yet psychometricians agree that errors exist in analyses that make use of raw test scores to compare students. To understand this concern, let us think about an exam that is scored on a scale of 0–25 points. A researcher might be tempted to treat the exam scale as linear and just "add up" the raw scores of different students to compare their levels of achievement. One problem of just adding up the number of correctly answered items and using that number to compare students is that it is highly unlikely that all test items are of equal difficulty. Therefore, a sum of raw scores cannot be used to achieve accurate comparisons of student performance. Consider the results of a test in Figure 2. Twenty-five multiple-choice items were presented to ninth-grade students. Imagine that the test covered a single variable (ninth-grade biology knowledge). Twenty of the items were well targeted to what ninth graders should know about the topic. However, the remaining five items were incredibly difficult, because they were at an introductory college level.

If a researcher simply summed and compared the scores of the students, he or she might assert that the difference in knowledge between Elizabeth and Henry (24 – 19 = 5) and between Pete and Johnny (10 – 5 = 5) are the same. However, this mathematical procedure contains a fundamental error, because the researcher ignores the differences in difficulty across the items. Elizabeth was able to answer a number of the highly difficult test items. Henry, Pete, and Johnny were unlikely to have successfully answered any of the five highly difficult test items. This means that the difference between Elizabeth's and Henry's knowledge levels is much greater than the difference in knowledge levels of Pete and Johnny. The seminal introduction to Rasch analysis, *Best Test Design* (Wright and Stone, 1979), discusses these issues in detail.

Now let us consider an example that illustrates a related problem with survey data. Figure 3 presents a commonly used rating scale of strongly agree (SA), agree (A), disagree (D), and strongly disagree (SD). A code of 4, 3, 2, and 1 is used as shorthand in a spreadsheet to indicate which response was selected for each survey item (e.g., SA is a 4, A is a 3). Figure 3 highlights one problem with immediately conducting statistical analysis with numerically coded respondent rating-scale answers. If a researcher conducts an immediate mathematical procedure with the rating-scale data, the researcher is assuming that the size of the jump from a strongly agree to agree is the same as the size of the jump from agree to disagree. The researcher can indeed argue that strongly agree represents more agreement than agree, and that agree represents more agreement than disagree, and so on. However, the researcher cannot immediately assume that the size of the jump between rating categories is equal.

Figure 3 also presents an additional issue with rating scales. Not only may the steps between adjacent rating categories be unequal, but the pattern of steps may differ from item to item. When the numerical answers to survey items are coded (e.g., SA = 4, A = 3, D = 2, SD = 1), it can be very tempting to immediately conduct mathematical analyses with those numbers. The only certainty is that, given a specific survey item, a rating of strongly agree means more agreement than a rating of agree, and so on through disagree to strongly disagree. Figure 3 shows

| Student | Score | Total Possible |
|---|---|---|
| Elizabeth | 24 | 25 |
| Henry | 19 | 25 |
| Pete | 10 | 25 |
| Johnny | 5 | 25 |

**FIGURE 2. Example test scores. The raw test scores of four ninth-grade students who completed the same 25-item test. Twenty items were appropriate for ninth-grade students, but five test items were at college level.**

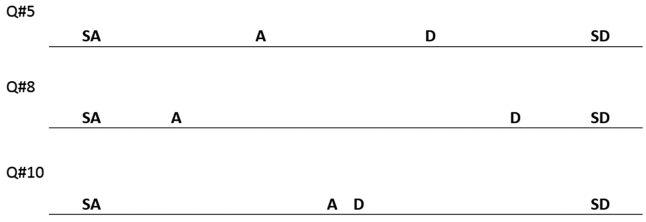


**FIGURE 3. Example survey rating scale. For the Q#5 scale, the "jump" between each of the ratings is equal. For the second (Q#8) and third (Q#10) scales, the "jump" from each rating to the next rating is not equal. Furthermore, the way the rating scale functions across the items is not identical. All that a researcher can assert is that the rating scale is ordinal (SA > A > D > SD) for each item.**

the potential unequal spacing of rating-scale categories for three survey items. In the field of psychometrics, researchers refer to such survey data as "ordinal" data. This means that an analyst can express the order of the response as follows: If Olive's answer to survey item 2 was "strongly agree" and Jin-Yung's answer to survey item 2 was "agree," then we know only that Olive's answer to item 2 exhibits a higher level of agreement than Jin-Yung's answer to item 2.

Just as all test items cannot be assumed to exhibit the same difficulty, all survey items should not be assumed to be equally agreeable. For example, a 4 (strongly agree) in response to item 8 of a survey should not be assumed to indicate the same level of agreement as answering a 4 (strongly agree) to item 10 of a survey. To understand this issue, let us consider the highly cited Science Teaching Efficacy Beliefs Instrument (STEBI; Enochs and Riggs, 1990). This instrument includes 13 survey items that define a self-efficacy scale for preservice elementary teachers. One STEBI item is "I will continually find better ways to teach science," and another STEBI item (following reverse coding) is "I will be very effective in monitoring science experiments." Preservice elementary teachers on average have a weaker foundation in science than in other content disciplines. Thus, it may be easier for a preservice elementary teacher to answer "strongly agree" to the item concerning finding better ways to teach science in comparison to answering "strongly agree" to the item involving monitoring science experiments. Just as test items cannot be assumed to have the same level of difficulty, survey items cannot be assumed to have the same level of "agreeability."

Rasch techniques offer a way to avoid these pitfalls and make use of raw test scores and rating-scale data to compute linear "person measures." The term "person measure" is the name of the Rasch scale number that expresses the performance of a test taker or scale respondent. Specifically, Rasch analysis allows researchers to use a respondent's raw test or scale scores and express the respondent's performance on a linear scale that accounts for the unequal difficulties across all test items. Rasch techniques involve corrections for a number of psychometric issues (e.g., rating scales are ordinal, not all survey items mark the same part of the variable) so that accurate person measures can be computed.

## THE RASCH MODEL

Figure 4 is a commonly used schematic that summarizes the core mathematical and theoretical concepts of the Rasch model, which were first developed by the Danish mathematician Georg Rasch (1960; see Appendix A in the Supplemental Material for a summary of selected Rasch terms). The single vertical line represents the construct to be evaluated by a test. Along this vertical line is a notation regarding the ability level of a student Oli along the variable. Also, three test items are plotted along the variable. Each item is located in a position that indicates the level of difficulty or ease of each item with regard to the variable. Of the greatest importance is that each item along the variable exhibits a probability of the respondent (with a specific ability level) correctly answering each item. An item exhibiting difficulty higher than the ability level of the respondent will have a lower probability of being correctly answered than an item of difficulty below the ability level of the respondent. In the case of our schematic, Oli will have a 50% chance of correctly answering item 2, less than a 50% chance of correctly answering item 1, and greater than a 50% chance of correctly answering item 3.

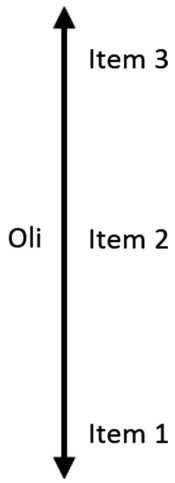


**FIGURE 4. Rasch measurement schematic. To measure, an analyst must 1) consider a single construct (represented by the vertical line); 2) consider the parts of the variable marked by different test items; 3) understand that a test taker will be located at some point along the variable; and 4) understand that the probability of a respondent answering a test item correctly can be expressed.**

Figure 5 depicts the Rasch mathematical model for dichotomous test items. The model is based on an appreciation that, to make measurements in the case of right/wrong test items, researchers must consider the difficulty of each test item along a variable and the overall ability level of a test taker with respect to the variable. Georg Rasch's model specifies that, when a respondent ($B_n$ on the left side of the equation) answers an item ($D_i$ on the left side of the equation), this relationship will be expressed by the natural log of the respondent correctly answering the item ($P_{ni}$) divided by the probability of the respondent not correctly answering the test item ($1 - P_{ni}$). Thus, the Rasch mathematical model (for right/wrong tests) makes use of a single variable, the location of a respondent along the variable, and the location of test items along the variable.

## APPLYING RASCH THEORY TO INSTRUMENT DEVELOPMENT AND DATA ANALYSIS

### Instrument Conceptualization and Design

Rasch analysis is both mathematics and theory. To understand how Rasch theory can guide instrument development, let us consider a biology education research project in which a researcher plans to administer a 25-question multiple-choice biology knowledge test to students. The researcher will, in essence, create a "meter stick" that will be marked by the 25 test items in order to compare students' knowledge. Some items

$$B_n\text{-}D_i = \ln (P_{ni}/1\text{-}P_{ni})$$

**FIGURE 5. The dichotomous Rasch model. $B_n$ is the ability of the test taker along the variable; $D_i$ is the difficulty of a test item; $P_{ni}$ is the probability of the test taker correctly answering a specific test item; and $1 - P_{ni}$ is the probability of a test taker incorrectly answering a test item.**

will exhibit a low level of difficulty, and these items will mark the easier end of the meter stick. Other items will exhibit a middle level of difficulty, marking the middle of the meter stick. Still other items will exhibit a high level of difficulty, marking the high end of the meter stick. Generally, our researcher should work toward presenting a range of "test-item difficulty" to students. This idea is similar to a meter stick for measuring the height of the flowers (Figure 1). Practically speaking, we can make only a limited number of marks on the meter stick. Thus, if we do not know the length of what we are meanning an equal distribution of marks along the meter stick provides optimal measurement opportunity.

The next step in applying Rasch theory is for our researcher to predict the location of marks (item difficulty) along the meter stick for specific test items. This means that the professor must use his or her understanding of what is being measured and, ideally, research on student biology knowledge to make predictions of item difficulty (where items fall on the meter stick). This use of theory to make predictions is central to measurement and Rasch analysis. If test developers cannot make the predictions, then the test developers do not understand what is being measured and cannot discern the meaning of one student performing better or worse than another student. For example, studies of student understanding of evolutionary change support a theory that students will have 1) more difficulty explaining evolutionary change of plants in comparison to animals; 2) more difficulty understanding between-species change in comparison to within-species change; and 3) more difficulty understanding loss of variables in comparison to gain of variables (Nehm *et al.*, 2012). This information can be used to formulate test items that span the meter stick of student understanding of evolutionary change.

The same Rasch techniques can be applied when developing a survey instrument. For example, if a researcher wishes to collect survey data on teachers' confidence in teaching biology, the researcher must be able to predict which survey items tap different ranges of confidence. Items should be included that would be agreeable even to teachers with low levels of confidence (e.g., "I will be able to plan a biology lesson"), and items should be included that are agreeable to only the most confident teachers (e.g., "I would feel at ease if the department chair wanted to observe my teaching"). In this example, the two items mark different parts of the variable "confidence."

Following the thoughtful construction of the measurement instrument, our researcher should collect pilot data, conduct a Rasch analysis of the pilot data, and then refine the instrument, for instance, by adding or removing items or changing the rating scale to have more or fewer rating-scale steps. Two exemplary steps taken in a Rasch analysis to evaluate the functioning of an instrument are outlined below. Many Rasch software programs can be used. Winsteps (Linacre, 2015), the most widely used Rasch software, is user-friendly, and the author of the program provides guidance and assistance to users.

### Using a Rasch Wright Map to Evaluate the Strengths and Weaknesses of an Instrument

To further understand the power of Rasch analysis for instrument development and improvement, we now consider a Wright map, which is named in honor of the University of Chicago's Benjamin Wright, who worked closely with Georg Rasch. A Wright map makes use of the fact that the difficulty of test items can be computed, and those test-item difficulties are expressed using the same linear scale that is used to express a student's performance—the person measure. In the case of a test, a Wright map allows researchers to evaluate how well the test items are defining a variable. A Wright map also allows researchers to compare the predicted order of item difficulty with the actual order of item difficulty in a data set. Such comparisons facilitate an assessment of construct validity by providing evidence that the instrument is measuring in a way that matches what a theory would predict. Wright maps open, multiple avenues for researchers to evaluate the inferences that can be confidently made through use of an instrument. I will provide an overview of selected Rasch analysis techniques, which are described in detail in *Rasch Analysis in the Human Sciences* (Boone *et al.*, 2014).

Figure 6 depicts a Wright map that plots the items in an instrument according to their order of difficulty. On the right side of the Wright map, the 25 items of the test are presented from easiest (item 2, bottom) to most difficult (item 30, top). The items are plotted in terms of item difficulty computed using Winsteps and the Rasch model formula. A "logit" scale is used to express item difficulty on a linear scale that extends from negative infinity to positive infinity. For many analyses, item difficulties will range from –3 logits to +3 logits.

Our researcher should now review the ordering of test items along the variable and compare the predicted ordering of items to the observed ordering of items. If the ordering matches what is predicted from theory, strong evidence is provided that the researcher has a good concept of what is being measured. If the pattern of item difficulty exhibits some major divergences from the prediction, then the researcher must stop and consider why the differences occurred. Is there something about the theory that needs to be revised?

Next, the researcher could evaluate how well the 25 items mark the meter stick. Are there gaps in the location of marks? If two students should fall in the gap (i.e., between marks), a researcher would not be able to differentiate the students. Are there locations where numerous marks are in the same location of the meter stick? Having test items mark the same location of the meter stick is, in essence, wasting a mark. It is better to remove one of the test items and shorten the test. The item could be removed and replaced with a new item that fills a gap. In Figure 6, readers can observe a good distribution of items from easiest to most difficult. However, some marks are located at the same spot or close together (e.g., items 31 and 36). Also, some parts of our meter stick are bare and need marks (e.g., between items 17 and 18 and item 7).

Wright maps are also valuable because they exhibit a plot not only of the items but also of the respondents. On the left or "person" side of the Wright map, an "X" is used to plot each of the 75 test takers. The higher the person measure, the better the test performance. The lower the person measure, the poorer the test performance. The six people (six "X's" in the top row) who have a person measure slightly below 2.0 logits are the highest-performing test takers for this measure. Analysis of the Rasch person measures provides researchers with a tool to evaluate the quality of their instruments. For example, does the ordering of the person measures make sense? In other words, are those students a

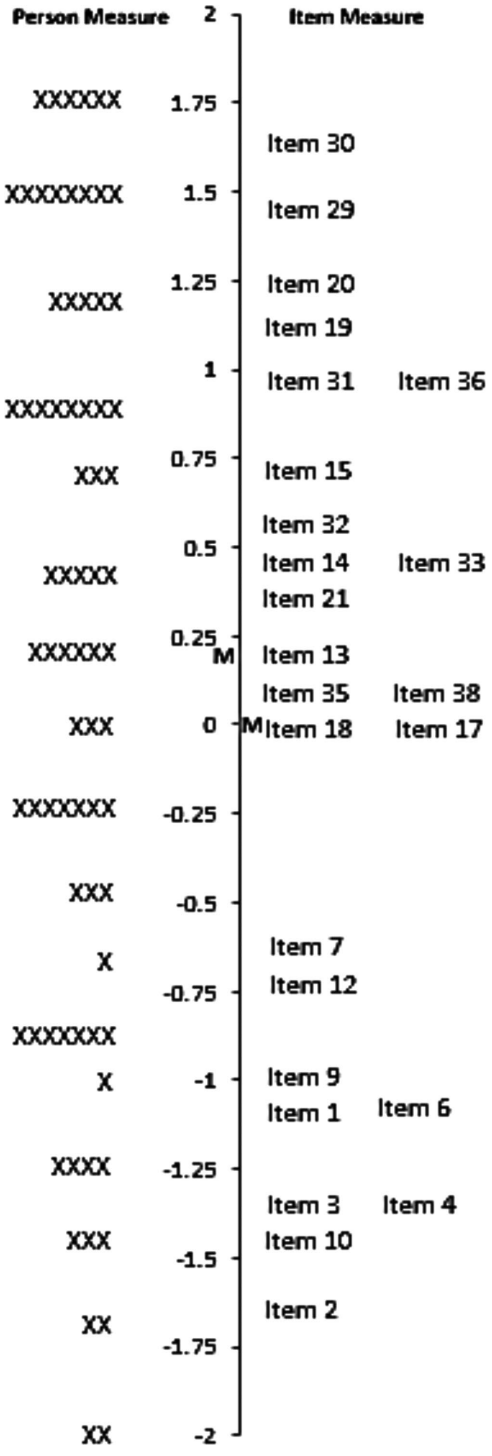


**FIGURE 6. Example Wright map. A Wright map can allow researchers to quickly identify strengths and weaknesses of an instrument. For example, are some test items measuring the same part of the variable? Are there portions of the tested variable that are missing test items? Investigating the location and distribution of test items on a Wright map is akin to reviewing the marks placed on a meter stick.**

researcher might predict to be high performers indeed high performers? Are students expected to be low performers indeed low performers?

Finally, because the Wright map provides both person measures and item measures on the same linear scale, researchers can determine how well the test items are distributed with regard to the ability level of test takers. A technique for accomplishing this is to evaluate how close the mean item measure ("M" on the right side of the Wright map) is from the mean person measure ("M" on the left side of the Wright map). For this data set, the mean items and mean persons are very close to each other. This arrangement suggests good test-item targeting. Also, this means that the range of test items presented to the students is appropriate for this group of respondents. In other words, the test items are not too difficult or too easy for the students.

### Additional Rasch Strategies for Evaluating Instrument Quality

A number of additional Rasch steps can be taken to evaluate the quality of a measurement instrument. One technique is to evaluate of the "fit" of items to the Rasch model (Boone *et al.*, 2014). One way to consider the topic of fit is that items at the more difficult end of the variable should be harder to correctly answer than items at the easy end of the continuum. This should be true for all students answering a set of items regardless of their ability levels. If items do not fit the model, they may measure more than one variable. It is critical to identify and possibly remove such items, as the goal of an instrument should only be to measure different parts of a single variable. In a Rasch analysis, identification of items that do not contribute to useful measurement can be accomplished by reviewing "fit" statistics (e.g., MNSQ Item Outfit, MNSQ Item Infit) for each test item. If an item does not clearly fit, often it is best to remove the item from the test and replace it with a new item. There are many reasons why an item may not fit (Wright, 1991). An item may not fit because it is difficult for the sample of students but is unexpectedly answered correctly by a number of poor-performing students. An item may be a misfit because it is an easy item that is unexpectedly answered incorrectly by high-performing students. A commonly used rule of thumb is to evaluate the Outfit MNSQ statistic for each item to determine if it exceeds 1.3. If so, the item might be misfitting the Rasch model and may be operating in a manner that is not useful for measurement.

Another technique to evaluate instrument quality is to review person-fit statistics in order to flag respondents who exhibit unusual answering patterns. For instance, patterns may suggest that a student concentrated for a period of time and then did not concentrate. Patterns can be detected that suggest a student wildly guessed when taking a test. The important point is that researchers who are first learning about Rasch should be aware of a number of data quality-control steps that are taken to evaluate the consistency of students' answers. For example, if students exhibit unusual answering patterns those students might not be included in an analysis.

## DEVELOPING DIFFERENT TEST FORMS

Because Rasch techniques are built upon the goal of measuring a single variable, researchers can build different forms of a test. Test responses can therefore be expressed on a single scale that is independent of the test form completed. For example, a 25-item test can be constructed for a Fall administration, a

| Form A | Form B | Form C |
|---|---|---|
| Unique Item ID Number | Unique Item ID Number | Unique Item ID Number |
| 1 (Only Form A) | 4 (Form A & B) | 18 (Form B & C) |
| 2 (Only Form A) | 5 (Form A & B) | 19 (Form B & C) |
| 3 (Only Form A) | 6 (Form A & B) | 20 (Form B & C) |
| 4 (Form A & B) | 7 (Form A & B) | 21 (Form B & C) |
| 5 (Form A & B) | 16 (Only Form B) | 27 (Only Form C) |
| 6 (Form A & B) | 17 (Only Form B) | 28 (Only Form C) |
| 7 (Form A & B) | 18 (Form B & C) | 29 (Only Form C) |
| 8 (Only Form A) | 19 (Form B & C) | 30 (Only Form C) |
| 9 (Only Form A) | 20 (Form B & C) | 31 (Only Form C) |
| 10 (Only Form A) | 21 (Form B & C) | 32 (Only Form C) |
| 11 (Only Form A) | 22 (Only Form B) | 33 (Only Form C) |
| 12 (Only Form A) | 23 (Only Form B) | 34 (Only Form C) |
| 13(Only Form A) | 24 (Only Form B) | 35 (Only Form C) |
| 14(Only Form A) | 25 (Only Form B) | 36 (Only Form C) |
| 15(Only Form A) | 26 (Only Form B) | 37 (Only Form C) |

**FIGURE 7. Multiple test forms. An example of how item anchors can be used to link the measurement scale of different test forms. Four items (4, 5, 6, and 7) are common to forms A and B, allowing the two scales to be linked. Four items (18, 19, 20, and 21) are common to forms B and C, allowing all the test takers (regardless of the form completed) to be expressed on the same scale.**

25-item test with many new items can be constructed for a Spring administration to the same group of students, and student performance on the two tests can be confidently compared. To understand this issue, consider three forms of a 25-item biology test, Forms A, B, and C. Form A is to be administered in the Fall, form B is to be administered in the Winter, and form C is to be administered in the Spring. To measure growth in student knowledge over time, the three forms must be equally difficult. Otherwise, it will appear that student growth is greater, or less, than it actually is.

First, it is almost impossible to develop test forms that exhibit identical difficulty. Second, if students are learning, it would be advisable to have harder items presented in each subsequent test form. By using Rasch techniques, researchers can develop forms of a test with different mixes of items and express all of the forms on the same scale. Developing three forms of a test using Rasch requires the employment of "item anchors." Item anchors are common items that are presented across forms and serve as reference points, such that student performance can be expressed on a single scale regardless of test form completed. Figure 7 presents a schematic that displays a mix of test items for three forms that are linked through item anchors. Even though each test form includes a different mix of items, linking the measurement scale through common items makes it possible to express test-takers' performance on the same measurement scale. This prevents any differences in test-form difficulty from influencing the interpretation of differences in student results. A simple way to link or anchor the three tests with four common items to the same scale is to first conduct the Rasch analysis with the data from test form A. When the test form B data are collected and analyzed, anchor the four items common to the item difficulty values computed through the analysis of the test form A data. This allows the responses from form B to be measured on the same scale as responses to form A. The same procedure can be followed to link the form C scale to the form B scale, thereby linking the form C scale to the form A scale.

## COMMUNICATING RESEARCH FINDINGS

A final point to make for those who are considering using Rasch measurement is that Rasch analysis enables researchers to explain the meaning of a person measure using the landscape defined by the test items. When Rasch analysis is used, it is possible for any person measure (e.g., how well Isabella performed on a test) to determine which items one can predict that Isabella answered correctly and which items one can predict that Isabella did not answer correctly. Below is a brief introduction to the way in which Wright maps are currently being used to describe the meaning of a test-taker's performance.

Figure 8 provides a Wright map with the location of 10 items and the location of the mean person measure at a pretest time point (e.g., start of a semester) and a posttest time point (e.g., end of a semester). Given the mathematical properties of the Rasch formula, a researcher is able to extend a horizontal line for the pre and post mean measures. For the results presented here, the items *below* the pre average line are items for which there is a greater than 62% chance of the average pre person correctly solving the test item. The items *above* the pre line are those for which there is a less than 62% chance of the typical pre person correctly solving the test items. Thus, the items below the pre line are items a researcher can be fairly sure that the typical pre student will correctly answer, and the items above the pre line are the items that the typical pre student will incorrectly answer. This means that a researcher could both compute a pre measure for a group of respondents and explain what the meaning of the group measure is.

Now review the location of the post group measure. Items that are below the post average group measure (marking the average student post measure) are those items that a researcher

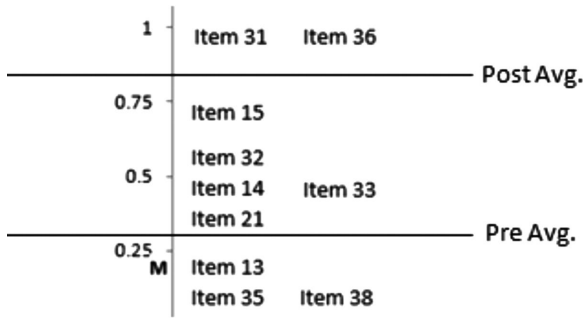


**FIGURE 8. Making inferences using Wright maps. A Wright map allows a researcher to explain the meaning of the growth observed from pre to post. The items falling between the pre and post lines help describe the growth.**

would predict the typical post student to answer correctly. The items above the post average line are items that a researcher would predict the typical post student to not answer correctly. Now let us imagine that a researcher conducted a statistical test comparing the pre group average measure to the post group average measure, and the researcher determined a statistically significant difference with a large effect size (Lenhard and Lenhard, 2015). The researcher can now explain the meaning of the change in student responses from pre to post. The meaning of the change can be described by the items that the average student could not correctly answer before the intervention, but the average student could correctly answer these items after the intervention. These are the items that lie between the pre average measure of the students and the post average measure of the students.

## CONCLUSION

Rasch techniques have greatly impacted the manner in which social science research makes use of tests and surveys (e.g., Panayides *et al.*, 2010). The Rasch framework offers procedures (see Appendix B in the Supplemental Material for a Rasch road map) for constructing and revising social science measurement instruments and documenting measurement properties of instruments (e.g., reliability, construct validity). Rasch techniques also enable researchers to make critical corrections when using raw test score data or survey data. Specifically, Rasch techniques allow nonlinear raw data to be converted to a linear scale, which then can be evaluated through the use of parametric statistical tests. In addition to the examples provided in earlier, there are Rasch steps that can be used to investigate additional important instrumentation issues (e.g., step ordering/step disordering, item reliability, person reliability, differential item functioning, and differential test functioning). The important point for beginning Rasch users to note is that creating a good test or survey starts with a theory about a variable of interest followed by a steps for evaluating how well the instrument appears to measure the chosen variable.

One of the most powerful aspects of Rasch measurement is that the technique allows the meaning of student measures and the meaning of group measures to be explained using the context of the instrument's items. For a test, if a group of students improves from pre to post, a researcher can explain the meaning of the change. For those who are interested in learning more about Rasch techniques, *Rasch Analysis in the Human Sciences* (Boone *et al.*, 2014) is a great starting book. More advanced texts include *Best Test Design* (Wright and Stone, 1979), *Rating Scale Analysis* (Wright and Masters, 1982), and *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (Bond and Fox, 2007). Many biology education researchers, two good examples being Eggert and Bögeholz (2010) and Jüttner *et al.* (2013), have used Rasch techniques and can serve as examples of how to go about this work.

## REFERENCES


Bond TG, Fox CM (2007). Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 2nd ed., Mahwah, NJ: Erlbaum.

Boone WJ, Staver JR, Yale MS (2014). Rasch Analysis in the Human Sciences, Dordrecht, Netherlands: Springer.

Eggert S, Bögeholz S (2010). Students' use of decision-making strategies with regard to socioscientific issues: an application of the Rasch partial credit model. Sci Educ 94, 230–258.

Enochs LG, Riggs IM (1990). Further development of an elementary science teaching efficacy belief instrument: a pre-service elementary scale. Sch Sci Math 90, 694–706.

Jüttner M, Boone W, Park S, Neuhaus BJ (2013). Development and use of a test instrument to measure biology teachers' content knowledge (CK) and pedagogical content knowledge (PCK). Educ Assess Eval Account 25, 45–67.

Lenhard W, Lenhard A (2015). Calculation of Effect Sizes, Bibergau, Germany: Psychometrica. www.psychometrica.de/effect_size.html (accessed 1 September 2016).

Linacre JM (2015). Winsteps Rasch Measurement (computer software), Beaverton, OR.

Nehm RH, Beggrow EP, Opfer JE, Ha M (2012). Reasoning about natural selection: diagnosing contextual competency using the ACORNS instrument. Am Biol Teach 74, 92–98.

Panayides P, Robinson C, Tymms P (2010). The assessment revolution that has passed England by: Rasch measurement. Br Educ Res J 36, 611–626.

Rasch G (1960). Probabilistic Models for Some Intelligence and Attainment Tests, Copenhagen: Danmarks Paedagogiske Institut.

Reeves TD, Marbach-Ad (2016). Contemporary test validity in theory and practice: a primer for discipline-based education researchers. CBE Life Sci Educ 15, rm1.

Wright BD (1991). Diagnosing misfit. Rasch Meas Trans 5(2), 156.

Wright BD, Masters GN (1982). Rating Scale Analysis, Chicago: MESA Press.

Wright BD, Stone MH (1979). Best Test Design, Chicago: MESA Press.