THE NEW AND IMPROVED Consortium RASCH MEASUREMENT MODEL PRIMER

Requirements of a good measurement system. When developing surveys to accurately measure peoples' beliefs and attitudes, there are several characteristics of the measurement system that are essential. They should be required to have: (1) measures of the attitude of each person ("person measure") on a linear, unbounded scale; (2) item difficulties that are placed on the same scale as person measures; (3) evidence that there is unidimensionality of the constructs being measured; (4) estimates of the precision of each person measure (person standard error), which also provides an overall reliability of the measure; (5) ways to assess construct validity of the measures; (6) ways to assess the internal validity of the measures; and (7) a way to handle missing responses. Raw-score measurement models suffer from several major disadvantages. First, raw-score models treat response categories as if they are linear when in fact they are not. Second, raw-score models do not give individual standard errors (usually only overall standard errors, or average standard errors). Third, raw-score models provide no mechanism to gauge how well the data fit the model. And finally, raw-score models do not handle missing data well, if at all.

The Rasch measurement model. Each of these concerns is addressed in our choice to use a Rasch model (Wright and Masters, 1982) to test the validity and reliability of the measures we will develop. The Rasch measurement model employs the principles of Item Response Theory (IRT) to analyze test and questionnaire data. On an assessment on which a person answers questions either right or wrong, Rasch determines the probability of a person responding correctly or incorrectly to each item. The basic principle of the Rasch model derives from this relationship: a higher-ability person will have a greater probability of getting a given item correct than a lower ability person; a given person will have a higher probability of getting an easier item correct than getting a harder item correct. With surveys that are scored on a multi-point rating scale, the model is extended so that instead of modeling the probability of getting a particular item correct, we model the probability of responding in a particular category in contrast to responding in any *lower* category on that item (for example, choosing "agree" vs. "strongly disagree" or "disagree" on a 4-point scale). Therefore, on a survey, "difficulty" means that it is harder to agree with (or endorse) the item statement. Items that are most difficult are those that have lower odds of people responding positively to them. For example, the following items are currently asked of students:

How safe do you feel:

- *In the hallways and bathrooms of the school.*
- *Outside around the school.*
- *Traveling between home and school.*
- *In your classes.*

Response categories: Not safe, Somewhat safe, Mostly safe, Very safe

Rasch analysis has found that the most difficult item to endorse in Safety was the second item above; relatively few students report that they feel safe outside around the school. The easiest item to endorse was the last item; most students feel very safe in their classes. Below is a further description of how Rasch addresses each of the measurement model requirements listed above.

(1) Rasch places person measures on a linear ability/attitude scale: Survey item response categories are not linear. For example, the difference between "strongly disagree" and "disagree" on a 4-point Likert scale is *not* necessarily the same as the difference between "disagree" and "agree" and so it would be a mistake to treat the category codes as numbers and do arithmetic with them, as raw score measurement models do. Instead, Rasch creates linear measures from the counts of responses in

categories, which are presented in logits.[1] The analysis produces a score for each person, called a **person measure** (the overall measure of a person's ability/attitude across all items in a measure). It also produces a score for each item, called an **item difficulty** (the overall measure of the difficulty of endorsing each item).

(2) Rasch places person measures and item difficulties on the same scale: Rasch places both person measures (person attitude) and item difficulties on the same scale thereby permitting us to directly make inferences about a person's performance relative to the scale of items. In the case of questionnaire data, this enables us to easily predict a person's responses to all the items in a measure given the person's score on the measure. For example, we can say that a person with a score of 1.0 is expected to agree with the two hardest items and strongly agree with the other items. This enables us to concretely characterize a person's attitude more meaningfully than just saying "her measure on Teacher Influence was 1.0". So, although the logit measures are not meaningful in themselves, being able to state expected responses enables us to concretely describe any measure value.

(3) Rasch establishes unidimensionality of measures: Of primary importance is the notion that measures are unidimensional. If the scale is unidimensional, we are measuring people on a single unitary concept on which people exhibit more or less of that construct. We develop items that manifest differing degrees of that construct, and then combine the items to form a scale. For example, to assess the degree to which teachers collaborate with each other, we would ask a set of questions about teacher collaboration that are intended to be easier or more difficult to agree with. Measures that assess multiple concepts are difficult to understand or to use in analysis. Moreover, without unidimensionality, it becomes very difficult to verify the reliability and validity of the measure. Rasch ensures this assumption using various methods, as described in #6 below.

(4) Rasch provides a person standard error: Because Rasch incorporates all responses by all respondents simultaneously, it produces a measure for *each person* (see #1 above). This person measure has an associated standard error, indicating the precision of this measure. The person standard error depends on how many items a person responds to (more data produces better precision), and how extreme the measures are (measures that are very low or very high will have the least precision). Calculated on a large number of data, persons who are near the average item difficulty will have low standard errors, meaning that we are more confident that the measure we have calculated for that person is very close to the person's actual attitude. The impact of large person standard errors can be reduced in two ways: (a) increase the precision estimates of person measures by adding more items, and/or (b) use these person standard errors in analyses to adjust for variation in measurement error. The average squared person standard errors also contribute to the calculation of the reliability of the measure: The lower the standard errors, the higher the reliability.

(5) Rasch provides a way to verify construct validity: We define construct validity as the degree to which the order of the empirical item difficulties agrees with the conceptual difficulties of the items. Using the **item difficulty** scores (which order items based on their difficulty within a measure; see #1 above), one can ascertain whether the items included in a particular measure match the conceptual difficulty of the items (e.g., an item you believe many teachers will agree with has a lower difficulty level than an item that you believe few teachers will agree with). The alignment of item difficulties with conceptual difficulties indicates that we are measuring what we really want to measure.

[1] Logits are log-odds units. For a probability *p* (on a scale of 0 to 1, i.e., bounded at both ends), the odds is $O = \frac{p}{1-p}$, which ranges from 0 to +infinity (bounded at the bottom). Taking the natural logarithm of the odds gives the logit, which is on a scale of –infinity to +infinity (completely unbounded).

(6) Rasch provides a way to ensure internal validity: We define internal validity as having measures that are unidimensional, meaning that they are measuring one, and only one, concept (see #3 above). The Rasch model calculates an expected response for each person to each item, and produces fit statistics indicating the degree to which people and items are acting in accordance with expectation. For example, a person with a high person measure will be expected to score highly (or on a survey, endorse more items), especially items that are more easily endorsed by everyone. The difference between the expected response and the observed response for that person is the residual. The **person fit statistic** is then just an aggregation of all the residuals from all items for that person. A person with a poor fit statistic is likely someone who has responded randomly. We are less likely to believe a person measure with a large misfit statistic. We can inflate the standard error of the person to reflect this uncertainty. Analogously, the **item fit statistic** is calculated from an aggregate of the residuals for all people to that item.[2] This helps determine whether there are items measuring a concept *other than* the one being assessed by the remaining items in that measure (indicating that we should perhaps reject the presumption of unidimensionality). You can increase the internal validity of your measures by removing items that are not related to the concept being measured, or add in other items that enhance the definition of the concept. Rasch models also provide **point-biserial correlations**, indicating how much the responses to each item within a measure are correlated with the overall measure. We use the item fit statistics and point-biserial correlations to verify that our measures only include items that are measuring the degree to which people endorse a single, underlying concept.

(7) Rasch allows for missing data: Rasch also handles missing data with aplomb. Given a set of items by which we measure people, each person can be given any selection of items to respond to without any bias or loss of accuracy of measurement. Rasch assigns an unbiased score to every person who answers at least one of the items in a measure.

Lastly, Rasch also provides step difficulties of the response categories in a measure. **The step difficulty** tells you the relative difficulty in responding in a particular category in contrast to the previous category. This can highlight whether particular response categories are being relatively unused or over-used. If this occurs, the developer might want to collapse response categories or revise the language in a subsequent version.

[2] The fit statistic we use, the mean-squared residual, varies between 0 and infinity. It can be interpreted as the ratio of the observed variation in the responses to the expected variation. Values greater than 1.0 indicate more than expected variation. In other words, a value of 1.25 means that there is 25% more variation than expected (either by a particular person or on a particular item). Very large values can indicate responses resulting from harmful extraneous factors such as noise, bias, or multidimensionality. We generally treat values less than 1.30 (30% more variation than expected) as acceptable, but there is always some leeway and room for interpretation. Values less than 1.0 indicate less than expected variation. This is not necessarily a bad situation, but can mean that the item is very similar to other items in the measure and is not providing much useful, additional information.