A Rasch Primer:
The Measurement Theory of Georg Rasch

Ronald Mead

March 2008

Psychometric Services Research Memorandum 2008–001

Mead, R.J. (2008) *A Rasch primer: the measurement theory of Georg Rasch*.
Psychometrics services research memorandum 2008–001.
Maple Grove, MN: Data Recognition Corporation.

Table of Contents

A Rasch Primer: The Measurement Theory of Georg Rasch

Abstract

This is an introduction to the measurement philosophy of Georg Rasch. It has three sections: measurement, models, and methods. The most popular motivation for using Rasch's models is that they are extraordinarily easy to use, compared to the Item Response Models (IRT). The *Part III Methods* section will show the arithmetic, clearly exposing just how easy the analysis phase of Rasch analysis is.

The *Part II Models* section introduces a variety of models that have what it takes to be a Rasch model. This diversity of expressions is another good argument in Rasch's favor. These include the well-known dichotomous, partial credit, and rating scale models. Others, like the multi-faceted models, are less well known but can be very useful. Some, like Fischer's linear logistic test models (*LLTM*), are better known outside the US than inside.

Part I Measurement concentrates on the most compelling motivation for using Rasch's models: following Rasch's principles and applying his methods leads to *measurement*, and measures are the appropriate fodder for analysis. It attempts to draw, as sharply as possible, the distinction between Rasch measurement and the more complex IRT models, without criticizing IRT directly. In contrast to IRT, Rasch's interest was how best to extract all information from the data *relevant* to the construct of interest, not how to reproduce the data most precisely.

Overture

This apology of the measurement theory of Georg Rasch is arranged in three movements with different tempos (measurement, models, and methods) but a recurring theme. The Measurement section attempts to show what Rasch meant by the word and to draw as sharply as possible the distinction between Rasch measurement and the more complex Item Response Models (*IRT*), without criticizing IRT directly. To re-enforce that distinction, we will follow the traditional Rasch notation reasonably well, when equations are unavoidable. We will also reserve the term *IRT* for the models that are not Rasch models, because that term seems to describe the intent of those models, with their focus on fitting the data, while the older term *latent trait models* fits better with the Rasch perspective, with its focus on the attribute to be measured.

Our concern throughout is the efficacy of Rasch measurement, how it works under ideal conditions, which can hardly be controversial. When the data conform to Rasch's principles, *i.e., the data are based on agents that are equally valid and reliable and are not subject to interference from extraneous attributes of the objects*, the models have the power to encompass and extend the best of Thurstone and Guttman. This leads to measurement, as most understand the word, and sets the stage for the more important tasks of taking and analyzing measures.

Most of the discussion surrounding Rasch has focused on *effectiveness*, how the models function when confronted with real responses from real people to real tests, questionnaires, surveys, checklists, and other instruments, some put together with little or no thought for their suitability for measurement. In this very practical world, Rasch analysis seems to mean running data through Rasch calibration software. The conclusion that Rasch models are robust, i.e., do pretty well in this real world, should not be taken as justification to continue doing what we've been doing.

There are two commonly cited motivations for using Rasch's models. The most popular being they are extraordinarily easy to apply, compared to the IRT models. Useful results can be gotten with relatively small samples and the estimation algorithms converge readily unless the data are pathologically bizarre. Part III of this treatise will show the arithmetic, clearly exposing just how easy Rasch analysis is to use.

This is our first trouble with Rasch: solutions to measurement problems are too simple to be worth publishing.

Part I focuses on the more compelling motivation for using Rasch's models: *following Rasch's principles and applying his methods lead to measurement, and measures are the appropriate raw material for analyses*. We will note again, and emphasize repeatedly, that applying Rasch's methods means more than running data through a Rasch calibration program.

This is our second trouble: building instruments that meet Rasch's requirements is hard.

Part II describes a variety of models that have the prerequisites to be Rasch models. Some of these, dichotomous, rating scale, and partial credit, are well-known and widely used. Others, e.g., multi-faceted models, are less well known and should be more widely used. Some, e.g., Fischer's linear logistic test models (*LLTM*), are widely known and used outside the US.

There is nothing new here; the majority of the entries in the reference list are between 1960, when Rasch's *Probabilistic Models for Some Intelligence and Attainment Tests* (Rasch, 1960) was published, and 1980, when it was republished shortly after Rasch's death. There is more here about rocks, archery, and oral reading than about multiple-choice items. We attempted, not at all successfully, to avoid mathematics, but those seeking rigorous explanations of estimation methods or fit statistics will need to look elsewhere (e.g., Smith & Smith, 2004; Fischer & Molenaar, 1995). This is not the manual for any Rasch computer package; it will not explain what *WinSteps*, *RUMM*, *ConQuest*, or *LPCM-WIN* is actually doing. Finally, this is not a cookbook for applying a special case of IRT models, although we do embrace the notion that Rasch models are very special indeed.

Part I: Philosophy and Theory of Measurement

Measurement theory refers to a body of principles, ideas, rules, and techniques for quantifying some interesting aspect of an object. Typically, the intent is to make inferences based on the measures but analysis is a distinctly separate process from measurement. Measurement does not care if we simply collect and file the measures or use them to achieve world peace. It is concerned with the very narrow problems of specifying the group of objects, defining the interesting aspect of the objects, determining relevant evidence, and transforming that evidence into a measure. In simple words, it is about building rulers, like the one in Figure 1. The numbers themselves are meaningless until associated with objects we are interested in and mileposts that mean something.

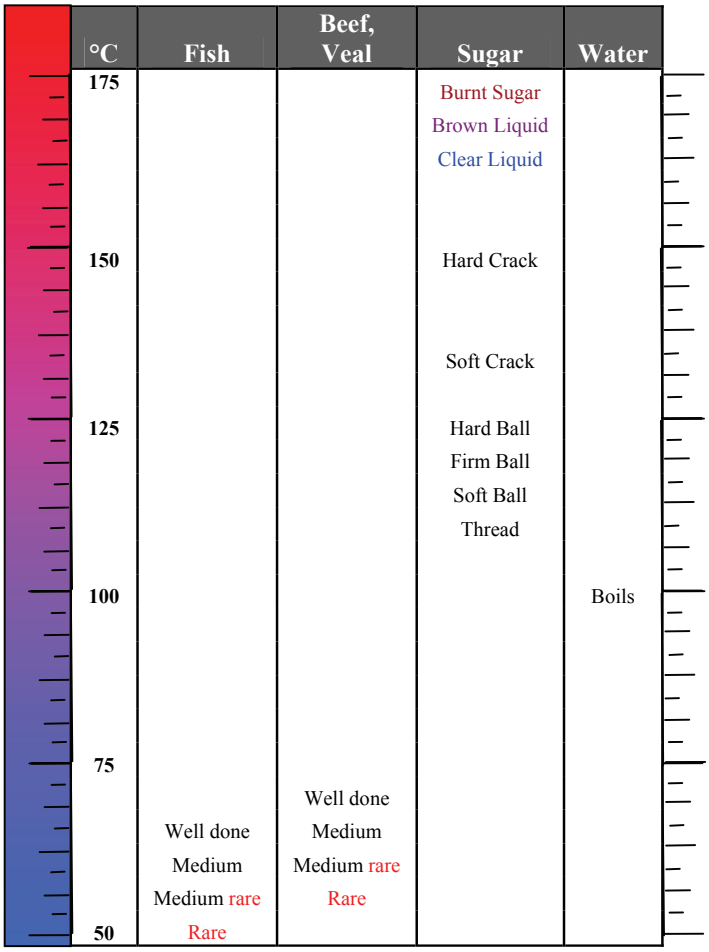

Figure 1: Celsius Temperature Guidelines

In psychological and educational measurement, the evidence comes from the object (e.g., person) interacting with the agent (e.g., item). A set of agents comprise an instrument (e.g., test.) The phrase *equally valid and reliable* used earlier to introduce Rasch's requirements, represents a major hurdle. In factor analytic terms, it suggests that the items all load equally on the first and only factor. In measurement terms, it means items that are truly interchangeable; after considering the single item parameter, we have absolutely no favorites among the items. Constructing instruments and banks of items that can claim this level of equality is the true challenge for measurement.

Building such instruments is valuable in its own right, helping define, refine, or discredit underlying theoretical notions (Andrich, 2004, pp. 171-174). Choosing the agents that we think are the best representatives of the idea forces serious consideration of what the idea really is and what constitutes acceptable evidence. Examining the results of the interactions with objects is almost always surprising and illuminating.

Once built, those instruments provide useful, meaningful measures of the objects that are truly of interest, objects like students and patients. The goal is to *understand*; not simply *explain* in the barren statistical sense, the item response matrix. A perfectly fitting model does not imply understanding. A stochastic model explains little beyond the state of our ignorance. The real work of measurement is not in estimating the parameters of the model that fits best, but in building the instrument that measures best.

If measurement is the goal, validity trumps reliability.

If Rasch measurement is not a special case of IRT, what is it?

Rasch Measurement is to Item Response Theory as Experimental Design is to General Linear Models. The math is easier and the inferences stronger but it takes more work and careful planning up front. Fisher (1947) shows us how to find cause and effect, the Holy Grail of science, given an appropriately designed experiment. Similarly from the Rasch perspective, the solutions for many vexing measurement problems are obvious and computationally trivial, given an *appropriate* instrument. Thurstone told us what the Holy Grail of measurement looks like; Rasch tells us how to find it.

IRT models were developed from the time-honored statistical perspective of data-fitting and model building. Parameters were added if they reduced the *unexplained* variance *significantly*. Items are weighted to minimize the difference between the model predictions and the observed item responses. We should not be concerned if the item weights differ depending on the sample that was used to obtain them or even if the items are ordered differently for different people and different groups. That is the world as it was given to us and we have done our best to faithfully reproduce the observed item responses.

For Rasch, reproducing the observed item responses is not measurement. The construct is paramount, defined operationally by the items. Any empirical weights will change the definition from the one provided by the designers of the instrument. If different selections of data redefine the construct, through the empirical weights, we no longer have a firm grasp on what we are trying to measure. If different subdivisions of the sample give different orderings of the items, then we do not know what *more* or *less* means. A measurement model should be sounding alarms in this situation, not adapting of it.

If the instrument does meet Rasch's requirements, then we know which way is up and can focus on studying relationships, monitoring progress, establishing effectiveness, and perhaps determining cause and effect. We can be less preoccupied by the mathematics of IRT analysis. Our focus can be the validity and utility of the measures rather the stability and efficiency of our algorithms.

Rasch has laid out a roadmap for building instruments, defining constructs, and making measurements. The process should be theory-driven and data-verified. When we build instruments for measuring based on a solid substantive theory, the Rasch model provides a rigorous structure for collecting and evaluating evidence related to that theory. If the items, tasks, or elements comprising the instrument do not behave as expected, we have learned something, perhaps about how to write better items, perhaps about the limits and limitations of our theory.

Building instruments according to Rasch's principles not only makes life simpler for psychometricians, it is a powerful tool in developing theories and doing science.

If you want to do measurement, you have to do Rasch.

For its adherents, Rasch measurement is axiomatic: self-evident because this is the way it must be or it's not measurement:

> *The calibration of the agents must be independent of the objects used and the measurement of the objects must be independent of the agents used, over a useful range.*

This is not an unchecked assertion, but a rational criterion by which one can evaluate data in the context of a theory. From the model follow consequences. If the observed data are consistent with the anticipated consequences, we have met the conditions of fundamental measurement and can treat the results with the same respect we have for measures of, say, mass, heat, distance, or duration, which, like reading fluency or art appreciation, are not real things but aspects of things.

What is measurement?

Volumes have been written on the meaning of measurement (Klein, 1975; Fisher, 1992; Tavernor, 2007). While we tend to think the physical sciences had it easy, weights and measures were not standardized until the end of the eighteenth century; local rulers religiously defended their own systems and the debate about whether measures were to be based on human body parts, the one-second pendulum at sea-level near Paris, the circumference of the Earth around the equator, or the circumference of the Earth around the poles figured in the French Revolution (Travernor, 2007).

As a jumping off point, we will take measurement to mean the *process of quantifying an aspect of an object in a way that is general, reproducible, and amenable to analysis*. *General* means there is a broad class of objects and of agents over which comparisons are valid and interesting. *Reproducible* means competent observers using appropriate instruments, perhaps of their own devising, will obtain statistically equivalent measures. *Amenable to analysis* means standard statistics (e.g., means, variances, differences) will suffice.

The process of measurement begins long before any data are collected. The starting point is a notion about an aspect of a class of things we want to understand better. It might be, for example, an ordinary idea like hardness, the quality of being firm or solid. This description, while completely reasonable and perhaps meaningful, does not say much about how one might measure it. Attempts to quantify the idea of hardness led to more functional definitions: e.g., the degree to which the surface of a material may be scratched, indented, abraded, or machined.

This notion of hardness first produced the Mohs scale, which relies on ten materials (Table 1) that are used to determine what scratches what. A material (e.g. garnet) that scratches quartz and is scratched by topaz has a Mohs scale value between 7 and 8. The Mohs' scale is useful for distinguishing diamonds from glass but not precise enough to differentiate grades of steel. These results are certainly reproducible and general but are nothing like an interval scale. The distance between 9 and 10 is orders of magnitude larger than the distance between 1 and 2. These are (ordered) categories, not measures.

Scale of Hardness		
Substance	Mohs	Shore Units
Talc	1	1
Gypsum	2	2
Calcite	3	9
Fluorite	4	21
Apatite	5	48
Orthoclase	6	72
Quartz	7	100
Topaz	8	200
Corundum	9	400
Diamond	10	1500

Table 1: The Ten Substances of Mohs Scale of Hardness

Later efforts (e.g., Brinell, Rockwell, Vickers) attempted to standardize the agent used to create the scratch. There are several Brinell scales and more than a dozen Rockwell scales with different loads and shapes of hammers, intended for materials with different properties and in different ranges of hardness.

All these techniques have the drawback that the values reported are specific to the load, shape, and duration. Brinell values are typically reported as, e.g., HB 10/1500/30, which means a Brinell hardness using a 10 mm diameter sphere under 1500 kilograms load for 30 seconds. The values from any of these methods and scales could be compared only to those obtained by the same method under the same conditions, i.e., took the same form of the test.

The *scleroscope* involved dropping a diamond-tipped hammer inside a glass tube from a fixed height and measured the height of the rebound on an arbitrary scale, which defined 100 *Shore* units as the average rebound from pure hardened high-carbon steel. Later, the scleroscope was refined to measure the energy *loss* from the impact rather than height of the rebound. The percent of energy loss was more general than the height of rebound and it can be related (linked) to the values determined by the indentation techniques as well as the Mohs scale (see Table 1.) This is an attempt to free the measurement from the specifics of the situation by applying some basic Newtonian physics to control or eliminate effects due to the mechanics of the device. The result was a scale that can be used as a measure.

The issue of unidimensionality does not even come up in this discussion. The topic is hardness. It was not confused with density, brittleness, luster, tensile strength, color, or any other aspect that these materials might share and might be interesting. However, if someone encounters a substance that scratches topaz (*Mohs 8*) and is scratched by quartz (*Mohs 7*), everyone would be back to the drawing board.

This process of development confronts many of the same issues that were expressed in Thurstone's quest for *absolute scaling* and Guttman's *scalogram analysis*.

Guttman

Louis Guttman (1916–1987) in 1944 proposed a method of scaling for psychological attributes like attitude and preference (Guttman, 1950). An ideal Guttman scale, like the Mohs hardness scale, has no random or error component. When the items are ordered by scale location, a person will endorse every item up to the total score and will refuse to endorse any item at a higher location. All of a person's responses can be reconstructed exactly from the total score. Where possible to achieve, this scale ensures a unidimensional instrument. The significance of total score as the summary of performance was a powerful insight.

Thurstone

Much of the seminal work on scale construction, anticipating fundamental measurement, was provided in the 1920's by L. L. Thurstone (1887–1955). He proposed several scaling methods that are still useful, including the methods of *equal-appearing intervals*, of *successive intervals*, and of *paired comparisons*. However, Thurstone's most telling insight may have been when he stated a requirement for fundamental measurement:

> *Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement. (Thurstone, 1928, p. 547.)*

Thurstone also suggested that the measurement of an object should be independent of the instrument to the extent that it should be possible to omit some items without disturbing the measurement (Thurstone, 1926, p. 446). These statements seem to herald Rasch's work thirty years later.

Rasch

Georg Rasch (1901–1980) was a Danish mathematician who studied statistics with Sir Ronald Fisher and with Ragnar Frisch, a Norwegian Nobel laureate in economics. From Frisch, Rasch learned *confluence analysis*, similar to Thurstone's factor analysis; from Fisher, he learned maximum likelihood estimation and *sufficiency*. Sufficiency in particular had a dramatic impact. In Rasch's words:

> *What is left over when a sufficient estimate has been extracted from the data is independent of the trait in question and may therefore be used for a control of the model that does not depend on how the actual estimates happen to reproduce the original data. This is the cornerstone of the probabilistic models that generate specific objectivity.* (Quoted by Wright, 1980, p. xii.)

Rasch's interest was how best to extract all information from the data *relevant* to the construct of interest and not how to reproduce the data most precisely.

In 1951, Rasch was asked to analyze data related to the effect of extra instruction for poor readers. The students had been given oral reading tests over a number of years; scores that were recorded included the number of words read in a fixed length of time. Texts of increasing difficulty were used on successive occasions; hence, there were no connections between texts or between occasions. There seemed no sensible way to compare the scores from one text on one occasion to the scores from another text on a different occasion (Rasch, 1977, p. 63).

The Vanishing Person Parameter

Rasch's background in mathematics and statistics and his interest in sufficiency suggested a simple expression and process: the probability that a person reads a specific number of words from a text follows a Poisson distribution with one parameter for the person and one parameter for the item:

$$1. \qquad p\{a_{vi}\} = e^{-\beta_v \varepsilon_i} \frac{\beta_v^{a_{vi}} \varepsilon_i^{a_{vi}}}{a_{vi}!}$$

where a_{vi} is the number of words read, β_v is the ability of person v to read quickly, and ε_i is the easiness of text i.

Rasch (1960, pp. 13-33) then derived:

- $p(a_{vi}, a_{vj})$, joint probability of scores a_{vi} and a_{vj} for person v on texts i and j,
- $p(a_{vi} + a_{vj})$, probability for the total of the two scores $(a_{vi} + a_{vj})$, and
- $p(a_{vi}, a_{vj} \mid a_{vi} + a_{vj})$, probability of the two scores conditional on the total.

The conditional probability of the two scores given the total was the key for Rasch. It came out to be:

- $$p(a_{vi}, a_{vi} \mid a_{vi} + a_{vj}) = \binom{a_{vi} + a_{vj}}{a_{vi}, a_{vj}} \frac{\varepsilon_i^{a_{vi}} \varepsilon_j^{a_{vj}}}{(\varepsilon_i + \varepsilon_j)^{a_{vi} + a_{vj}}}$$

Using Fisher's maximum likelihood, Rasch derived the estimator, which is astonishingly simple, for the relationship between the text parameters:

$$2. \qquad \hat{\varepsilon}_i - \hat{\varepsilon}_j = \ln(a_{vi}/a_{vj}).$$

The relationship is estimated[i] by the ratio of the counts regardless of what person is used. With this result, Rasch was able to collect data from totally new samples that connected all the texts and to use those estimates to evaluate the progress of the original sample.

When Rasch described this work to Frisch in a casual conversation, the Nobel economist remarked repeatedly, *the person parameter has completely vanished*. Rasch repeatedly responded, *yes, it has*, and continued to explain what they had concluded about remedial reading instruction. It took Rasch several days to appreciate what had struck Frisch immediately: **separating the two sets of parameters suggested an important new class of models with simple sufficient statistics.** (Rasch, 1977, p. 66) Expression (2) has all the properties of fundamental measurement Thurstone was seeking.

Specific Objectivity, Sufficiency, and Separation

Rasch chose the term *specific objectivity* to characterize his new models: *objective* because they allow comparisons between items without reference to the people and comparisons between people without reference to the items[ii]; *specific* to distinguish it from all other uses of *objective* but also to emphasize that this property is not demonstrated once and for all for all potential situations.

Separation, sufficiency, and specific objectivity are the mathematical, statistical, and measurement faces of Rasch models.

Phrases like *instruments that conform to my principles* and *tests built using my methods* are sprinkled throughout Rasch's writings. These take it for granted that we know *what his principles are* and *what his methods were*.

The Rasch Principle

The principle is specific objectivity. If it holds, then any reasonable selection of people will provide the same estimate of the comparison between any two relevant items and any relevant selection of items will provide the same comparison between any two appropriate people. And again, the essential role of sufficiency is to allow consistent estimators of the model parameters, and after the sufficient statistics have been extracted, all that remains should be noise. If it is, we have measurement; if not, we have a bigger research study than we thought.

The Rasch Method

Rasch's method was to

(1) design agents to provoke valid responses,
(2) extract the sufficient estimators, and
(3) examine the remains for any structure.

Often we rush through the first by having an inadequate theoretical basis, forget about the third because we are afraid of the answer, and shortchange the second by not exploiting the sufficient statistics. If we get through these steps and there is no structure, we can proceed to making and analyzing measurements. If there is structure, relating to, say, subgroups, individuals, item types or item content, we need to revisit the theory, reconsider the observations and the instrument, perhaps revising or discarding items, or rethink the range of individuals for whom the instrument is appropriate. Reconsider anything and everything except the Model[iii].

The problem at hand in the oral reading example[iv] was to determine the effect of extra instruction for poor readers. If we can measure, in the strictest sense, reading proficiency, measurements could be made before the intervention, after the intervention, and perhaps several points along the way. Then the analysis is no different, in principle, than if we were investigating the optimal blend of feed for finishing hogs or concentration of platinum for reforming gasoline.

In order to obtain evidence about reading proficiency, a parent or teacher might listen to the student read, commenting on errors and flow. There are many other possibilities for evidence that might be collected, perhaps having students retell the story in their own words, or respond to multiple choice items about main ideas, vocabulary from context, literary devices, sequence of events, use of imagery, topic sentences, etc.

There very well might be people for whom reading speed is not a reasonable indicator of proficiency. For advanced readers, for students with vision or hearing impairments, or students reading in a second language, reading faster might not imply reading better and pronunciation errors may not imply lack of understanding. Or perhaps it works in Danish but not Mandarin.

Once we are satisfied that we have chosen appropriate indicators of the construct for our class of objects, we could begin to quantify the activity by counting the words read rather than just listening and commenting. In order to standardize, we could fix the amount of time allowed[v]. Students reading from the same text for the same length of time could be ranked by the observed counts of words read. These counts are not measurements: without further refinement, the counts can not be compared to others based on other texts or other time intervals.

These deliberations are basic instrument development and hardly invented by Rasch. They are recounted here to emphasize the crucial role of the instrument and the rationale that supports it. One can not expect to take a poorly thought-out instrument and salvage it through psychometric manipulations or incantations.

Controlling the model to establish that specific objectivity holds is the center piece of Rasch's method. It is generally appropriate, when estimating the model parameters, to use all the data one can get one's hands on, because larger samples mean smaller standard errors. However, for control, the total collection is partitioned and the results compared every way that is a potential threat[vi]: high vs. low performers, males vs. females, fourth grade vs. fifth grade, or computer vs. paper-and-pencil administration. Having done all these with satisfactory results, we still do not know if the relationship is independent of visual acuity, mobility, computer experience, ethnicity, type of school or community, language spoken at home, eligibility for free or reduced lunch, region, mother's occupation, father's education, ad infinitum. Objectivity is *specific* to the threats eliminated.

Any of the possible subdivisions of the data can be investigated, using likelihood ratio tests (Fischer & Molenaar, 1995), between group χ^2 (Wright & Panchapakesan, 1969; Wright & Stone, 1978), residual analyses (Mead, 1976), mean squares, weighted mean squares (Smith & Smith, 2004), or any number of other statistics[vii]. Rasch often did it graphically. His work is filled with plots comparing the performance of groups of examinees, demonstrating the degree to which specific objectivity held, and identifying the instances where it did not.

Returning to oral reading, the counts of words read (like a_{vi} and a_{vj}) and the probabilities (like $p\{a_{vi}, a_{vj}\}$) are **person-dependent**. Raw scores in any form depend heavily on the people who produced them. However, from expression (2), any effects of the people are eliminated from the conditional probability $p(a_{vi}, a_{vj} \mid a_{vi} + a_{vj})$ and from the ratio (a_{vi}/a_{vj}). The ratio should be the same, within statistical limits, whether it came from a very fast reader, a very slow reader, from a male or female, fourth grader or fifth grader, etc. Comparing these ratios based on different disaggregations of the group tested is what Rasch meant when he said "*the relationship should be found in several sets of data which differ materially in some relevant respects*." (Rasch, 1960, p. 9)

The process that we have gone through, albeit rather crudely, to obtain useful evidence is:

- What aspect of the people are we trying to understand?
- What evidence might we collect that would relate to the status of the people in this aspect?
- What groups of people do we intend to measure?
- Can we reasonably expect this evidence to be valid for comparing all members of these groups?

These are not strictly psychometric questions although the psychometrician will probably have plenty of advice.

Rasch's method to take these observations from the lowly level of counts to the lofty level of measures includes:

- Find a mathematical form that
    - is consistent with the process that generated the counts, and
    - has separable sets of parameters.
- Eliminate *nuisance* parameters from the estimation equations using the model's sufficient statistics.
- Estimate the parameters.
- Check that the sufficient statistics really are sufficient by:
    - Dividing up the total group, and
    - Confirming that the relationships among the parameter estimates hold for all important subgroups.
- If not, revisit the theory, revise the items, reconsider the instrument.

Computers will do the calculations and there is a growing family of mathematical forms to choose from. We will discuss how to do the arithmetic in Part III and will describe some of the mathematical forms in Part II. The psychometrics are easy; constructing appropriate instruments is the hard part.

The ultimate goal of this exercise is to develop measurement scales that are as well-defined and as useful in the classroom and clinic as the one shown in Figure 1 is in the kitchen. Thurstone and Guttman, and others, defined what measurement must be. Rasch provides the mechanism to achieve it and the structure to know when we have. It depends on rigorous development of the agents (items) from a substantive theory and a careful verification of performance based on data. When accomplished, we will have measuring instruments that we can place along side thermometers, rulers, and scleroscopes, no apologies needed.

Part I Notes

[i] We keep saying the *relationship is estimated*, not the *parameters are estimated*. There are an infinite number of values that would satisfy the relationship. We will eventually impose some convenient constraint to resolve this ambiguity.

[ii] Wright (1968) introduced the often-misunderstood terms *item-free* and *sample-free* to denote the two aspects of specific objectivity.

[iii] It is acceptable to reconsider if we have the correct form of the Rasch Model.

[iv] Any statistician, Rasch undoubtedly included, would prefer to be brought into the process earlier. Contacting the statistician after several years of data have been collected will likely cause grumbling.

[v] We might have used the amount of time taken to read a fixed number of words. It may be that counting the seconds is equally valid as counting the words, or maybe not.

[vi] A partition is threatening if we suspect the relationships among agents or among objects might not hold and it matters if they don't.

[vii] No single fit statistic is either necessary or sufficient. *David Andrich*

Part II: The Family of Measurement Models

Rasch (1960) defined an important class of models. These models are characterized by a property Rasch labeled *specific objectivity*, which depends on *separable* parameters and simple *sufficient* statistics. This property allows the models to achieve a state that Thurstone, perhaps overly optimistic, described as *absolute measurement*. The utility of this property in practice is apparent in the remarkably simple estimation equation for item difficulties, which does not involve the parameters for the sample of people used to obtain the data.

Models can be very useful, highly predictive, and completely wrong[i]. The geocentric model of the solar system predicts nicely the days and seasons of our lives, which was adequate for gathering nuts and berries, following migrations of wooly mammoths, and setting school calendars, but not for landing probes on other planets. Newton's laws of motion describe beautifully everything humans see but do not describe the movement of satellites accurately enough for GPS devices nor explain *why* the moon stays in orbit. The stochastic processes posited by Rasch models do not *explain* the student behavior that produced the item responses, but the models can provide structure for using and not misusing the information the responses contain.

In educational measurement, there may be causal explanations (instruction, perhaps?) for what students can and cannot do. For the stochastic measurement model, high ability examinees pass easy items and low ability examinees miss hard items. That's hardly breakthrough thinking for the psychometrician nor useful information for the teacher. But it does form the basis for *measuring* the students and the items, for defining *easy* and *hard*, for establishing what *high* and *low* ability are, and, at least equally important, recognizing when something is odd.

Acceptance into the Rasch family requires a model with sets of parameters that interact in a simple way, permitting their effects to be separated. Simplicity is not so easy, but there are a number of models that qualify. This section provides an informal introduction to some of the most common, unidimensional members.

The appropriate form of the model to use is determined by the nature of the observation and the process that generates it, not by the computer software available. Is the process dichotomous, polytomous, or Poisson counts? Are the observations independent? Is the observation the result of an interaction between the object of measurement (e.g., person) and the agent of measurement (e.g., item) or are there other participants (e.g., judges)? The first and third questions are easy enough to answer; the second may be harder. The models discussed here all expect independence.

Poisson Counts

The Poisson form was the first member of the Rasch family; Rasch used it in the 1950's to analyze oral reading after observing the number of words read and the number of errors made. It is often presented in introductory probability courses as the distribution of rare events. A standard example is the number of defects in a bolt of cloth. The probability of finding a defect at any given spot is small; all spots are equally likely candidates for a defect; and there is no upper limit on the number of spots or the number of defects. Similarly for oral reading, the probability of misreading any word is small and all words are equally likely to be misread. Perhaps more realistically, because all the probabilities are very small, there is no real difference among them.

The basic model is:

$$3. \qquad p\{a_{vi}\} = e^{-\lambda_{vi}} \frac{\lambda_{vi}^{a_{vi}}}{a_{vi}!}$$

where a_{vi} is the count observed, and, in our case, $\lambda_{vi} = \beta_v \varepsilon_i$, where, in Rasch's original study, β_v is the proficiency of the person at reading aloud, and ε_i is the *easiness* of the passage to be read.

This expression yields the estimation equation for the relationship of two passages:

$$4. \qquad \varepsilon_i - \varepsilon_j = \ln(a_{vi}/a_{vj}).$$

As a Rasch model, the Poisson has been successfully applied to counts of errors made in oral reading (Rasch, 1960), of errors of various types in written essays (Andrich, 1973), of number of words read in a given time, of time taken to complete a task, of points scored in various games, and variety of cases where the score was a count of events with no definite upper limit. Generally, the number of *trials* (e.g., words that might be read or points that might be scored) is large compared to the number of *events* (e.g., words actually read, mistakes made, or points scored). Some of the basic equations were presented in expressions (1) and (2) of Part I; all the details are in Chapter I of Rasch (1960).

Dichotomous Rasch Model

The most familiar Rasch model, the one most people mean when they say *The Rasch Model*, is for dichotomous data, i.e., questions that are scored right or wrong, yes or no, agree or disagree, checked or not checked, present or absent, hit or miss, 1 or 0. This model may be compared to shooting an arrow at a simple bull's-eye.

The model asserts the probability of *hitting* the target is determined solely by the archer's bowmanship B_v and how hard the target is to hit, its difficulty Δ_i. The probability of a *hit* is then simply:

$$5. \qquad p_{vi}(hit \mid B_v, \Delta_i) = \frac{B_v}{B_v + \Delta_i}.$$

The probability of a *miss* is (1 – p):

$$6. \qquad p(miss_{vi} \mid B_v, \Delta_i) = \frac{\Delta_i}{B_v + \Delta_i}.$$

Obviously, we cannot grab any two numbers out of the air and shove them into expressions (5) and (6) and expect to come away with a sensible probability of anything. The ratio B_v / Δ_i defines the odds of the person winning over the item and, incidentally, B_v and Δ_i could be viewed as the odds versus a standard agent or object respectively. Ultimately, we would like to have good estimators for both B_v and Δ_i. That is where we are going now.

If we consider a two-target contest, there are only four possible outcomes for person v: hit both targets, miss both targets, hit the first and miss the second, and miss the first and hit the second. Probabilities of these four outcomes, assuming the two shots are **independent**, are shown in Table 2. The upper right cell is the probability p_{10} of scoring one out of two by hitting the first target only; the lower left is the probability p_{01} of scoring exactly one by hitting the second target only.

		Target 1: *Miss*	Target 1: *Hit*
Target 2	*Miss*	$p_{00} = \frac{\Delta_1\Delta_2}{(B_\nu+\Delta_1)(B_\nu+\Delta_2)}$	$p_{10} = \frac{B_\nu\Delta_2}{(B_\nu+\Delta_1)(B_\nu+\Delta_2)}$
	Hit	$p_{01} = \frac{\Delta_1 B_\nu}{(B_\nu+\Delta_1)(B_\nu+\Delta_2)}$	$p_{11} = \frac{B_\nu^2}{(B_\nu+\Delta_1)(B_\nu+\Delta_2)}$

Table 2: The Four Outcomes from a Two-Target Test

There are two ways of scoring 1 and the probability of scoring exactly *1* unconditionally is the sum of the separate probabilities, $p(r=1) = p_{10} + p_{01}$. Finally, expression (7) is the probability of hitting target 1 and missing target 2, given a total score of 1:

$$7. \qquad p(10 \mid r=1) = \frac{p_{10}}{p(r=1)} = \frac{p_{10}}{p_{10}+p_{01}} = \frac{\Delta_2}{\Delta_1+\Delta_2} \ .$$

If we know that you hit one of the targets, this is the probability that it was target 1 that you hit. As with expression (2) for the Poisson, the person parameter has completely vanished.

Connecting the model to the data

The counts needed to estimate the probabilities in Table 2 can be obtained by administering the two-item test to any pack of archers we find in the forest and counting the number of people who land in each of the four cells. The magic, elegance, power of Rasch's solution is that it does not matter whom we pick to take the test or how the person parameter is distributed in the sample. The relationship between the items is the same regardless of whom or what the people are.

Changing from Δ_i to D_i to indicate estimates rather than parameters, the relative target parameters can then be estimated with:

$$8. \qquad \hat{p}(10 \mid r=1) = \frac{n_{10}}{n_{10}+n_{01}} = \frac{D_2}{D_1+D_2} \ ,$$

where n_{10} is the number of people hitting the first but missing the second and n_{01} is the number missing the first but hitting the second.

Then, with a little rearranging:

$$9. \qquad \frac{D_2}{D_1} = \frac{n_{10}}{n_{01}} .$$

A variety of conditions could be imposed to resolve the ambiguity in (9), e.g., D_1=5, or D_2=23, or D_1*D_2=1. All would yield equally correct and usable solutions .

Expression (5) is not the form most people are accustomed to seeing for *The Rasch Model*. It is more common and more useful, but slightly messier, to express the parameters in *logits*: $\beta = ln(B)$ and $\delta = ln(\Delta)$. Logits are the units most amenable to analysis; relationships appear as differences rather than ratios. Expression (5) then becomes the familiar:

$$10. \quad p(x_{vi}=1 \mid \beta_v, \delta_i) = \frac{e^{\beta_v}}{e^{\delta_i} + e^{\beta_v}} = \frac{e^{\beta_v - \delta_i}}{1 + e^{\beta_v - \delta_i}} \ ,$$

And expression (9) for the relationship between two items becomes the difference:

$$11. \quad d_2 - d_1 = \ln(n_{10}) - \ln(n_{01}).$$

Again, we need a constraint to resolve the ambiguity; the most common one is $d_1 + d_2 = 0$, leading to estimates for the item parameters of a two-item test:

$$12. \quad d_1 = -d_2 = \frac{\ln(n_{01} / n_{10})}{2} \ .$$

When administering the test to obtain the data needed to estimate the cells of Table 2, any convenient sample of people can be used. For purposes of expressions (9) and (12), which compare item 1 to item 2, it does not matter at all how the people are distributed[iii]. All that matters is that they took the test and that their responses conform to the model.

Expression (7) is another simple demonstration of specific objectivity, for a test with two dichotomously scored items. What was required to arrive here was, first, an appropriate collection of items and, second, a mathematical model of the situation (expression 5 or 10) that described the data and that involved sets of parameters that could be separated. Separating the parameters led to a simple sufficient statistic for the person parameter, and made it possible to eliminate all influence of the sample from the estimation equation for the item parameters.

Polytomous Rasch Models

For many testing situations, simple zero-one scoring is not enough and it is not appropriate to assume Poisson-type counts. Polytomous Rasch models allow scored responses from zero to a maximum of some small integer m. The integer scores must be ordered in the obvious way so that responding in category k implies more of the attribute than responding in category $k-1$. While the scores must be consecutive integers, there is no requirement that the categories be equally spaced. To continue the archery metaphor, we now have a number of concentric circles rather than just a single bull's-eye with more points given for hitting within smaller circles. The case of $m = 1$ is the dichotomous model and $m \to \infty$ is the Poisson, both of which can be derived as special cases of almost any of the models that follow.

Rating scale

The rating scale model (Andrich, 1978; Wright & Masters, 1982) characterizes the person's responses as a simple function of the person's status (e.g., attitude, preference, condition, ability), and the item's strength, with several levels up or down for the response categories. The response category formats and parameters are assumed to be identical for every item. Because of this requirement,

the model is used more frequently for attitude, preference, or evaluation questionnaires than achievement testing. One common format is a series of statements that the respondent is asked to react to on, say, a four-point scale from "*strongly disagree*" to "*strongly agree*".
If we are considering, for example, the statement:

"The Rasch model is the very definition of measurement"

and the response format is:

Strongly Disagree	*Disagree*	*Agree*	*Strongly Agree*

and we intend to respond in either category "*agree*" or category "*strongly agree*", the probability of choosing "*strongly agree*" over "*agree*" is:

$$13. \qquad p^*(k \mid \beta_v, \delta_i, \tau_k) = \frac{p(k)}{p(k-1)+p(k)} = \frac{e^{\beta_v - (\delta_i + \tau_k)}}{1 + e^{\beta_v - (\delta_i + \tau_k)}},$$

where response category k is "*strongly agree*" and $p(k)$ is the unconditional probability of responding in category k, which we have not yet revealed. Because, at this point, we are considering only two categories, expression (13) is identical to the dichotomous case with the item difficulty δ_i replaced by $\delta_i + \tau_K$. The categories other than k and $k\text{-}1$ do not enter into the equation.

The distinction between p and p^* is that p is the probability of responding "*strongly agree*" but p^* is the probability of "*strongly agree*", given the response is either "*agree*" or "*strongly agree*". We have, in effect, already dismissed the less positive responses from our consideration.

Apply a little algebra to expression (13) and we have a recursive expression for $p(k)$:

$$14. \qquad p(k) = e^{\beta_v - (\delta_i + \tau_k)} p(k-1)$$

or equivalently and sometimes more conveniently, the log odds of k versus $k\text{-}1$ (i.e., logit):

$$15. \qquad \ln\left\{\frac{p(k)}{p(k-1)}\right\} = \{\beta_v - (\delta_i + \tau_k)\}.$$

While we now have an expression for $p(k)$, we need a starting point. It is convenient, and no more arbitrary than any other value, to define the logit for category *0* as *0*, and then the probabilities can be developed as in Table 3[iv].

k	*Logit*	*Numerator*	Pr*obability*
0	0	1	$1/\gamma$
1	$\alpha_1 = \beta_v - (\delta_i + \tau_1)$	e^{α_1}	e^{α_1}/γ
2	$\alpha_2 = \beta_v - (\delta_i + \tau_2)$	$e^{\alpha_1+\alpha_2}$	$e^{\alpha_1+\alpha_2}/\gamma$
3	$\alpha_3 = \beta_v - (\delta_i + \tau_3)$	$e^{\alpha_1+\alpha_2+\alpha_3}$	$e^{\alpha_1+\alpha_2+\alpha_3}/\gamma$
4	$\alpha_4 = \beta_v - (\delta_i + \tau_4)$	$e^{\alpha_1+\alpha_2+\alpha_3+\alpha_4}$	$e^{\alpha_1+\alpha_2+\alpha_3+\alpha_4}/\gamma$

Table 3: Response Category Probabilities for a Rating Scale Model

For the sake of completeness and the compulsively mathematical, the relationships of Table 3 can be captured in the standard expression of the rating scale model for the probability that person ν taking item i will respond in category k, given the person parameter β_ν, the item parameter δ_i, and m category parameters τ_j:

$$16. \qquad p\{k \,|\beta_\nu, \delta_i, (\tau_{j=1,m})\} = \frac{e^{\sum_{j=1}^{k}(\beta_\nu - \delta_i + \tau_j)}}{1 + \sum_{x=1}^{m} e^{\sum_{j=1}^{x}(\beta_\nu - \delta_i - \tau_j)}} = \frac{e^{k(\beta_\nu - \delta_i) - \sum_{j=1}^{k}\tau_j}}{1 + \sum_{x=1}^{m} e^{x(\beta_\nu - \delta_i) - \sum_{j=1}^{x}\tau_j}} \,.$$

The summation in the exponent represents the summing of logits in Table 3; the summation in the denominator is the summing of the numerators, the numerator of $k = 0$ being one.

Figure 2 shows the category characteristic curves for a four-category item with nicely spaced categories. The category parameters used to create the plot are (-3, -1, 1, 3). These parameters appear in the figure as the intersections between adjacent categories. The curves for category *0* and category *1* cross at *-3*; the curves for category *1* and category *2* cross at *-1*; *etc*. Items never look like this in real life.

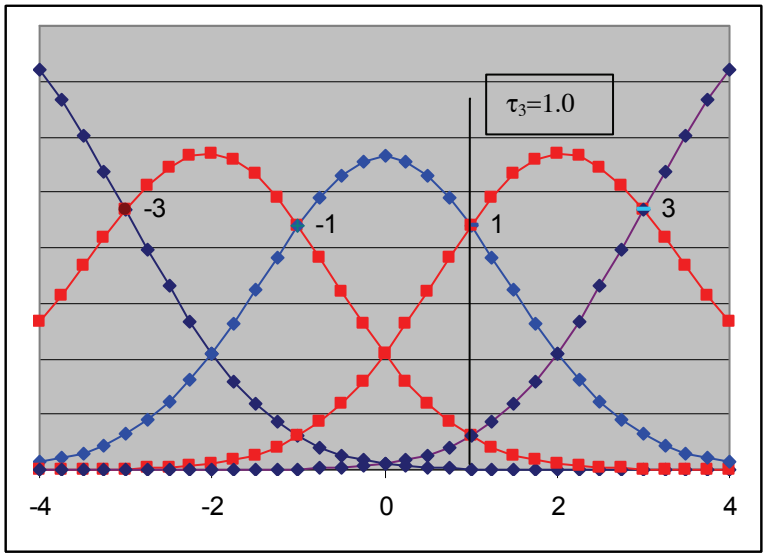

Figure 2: Rating Scale with Four Equally Spaced Thresholds: Parameters = (*-3*, *-1*, *1*, *3*)

A person in category k is not described appropriately by the category parameter. For this example, although τ_2 is *-1.0,* the most likely value for a person's location, given an observed category of 2, is a logit of *0.0*. Because of the symmetry of this example, this estimate happens to be half-way between adjacent category values.

Table 4 illustrates some of the calculations behind Figure 2; specifically, the calculations needed for a point on each curve where $\beta_v - \delta_i = 1$. Column *1* is the category score k. Column *2* is the category parameter τ_K; as with the dichotomous case, there is one fewer parameter than categories. The third column, $\exp(1 - \tau_K)$, is the exponentiation at the point on the logit continuum where the person parameter exceeds the item parameter by one logit. The *Numerator* is the exponentiation times the previous numerator. The *Probability* is the *Numerator* divided by the sum of the numerators. This is a repeat of Table 3 with numbers instead of symbols.

k	τ_k	exp(1- τ_k)	Numerator	Probability
0			1	0.00
1	-3	54.60	54.60	0.06
2	-1	7.39	403.43	0.44
3	1	1.00	403.43	0.44
4	3	0.14	54.60	0.06
			917.05	1.00

Table 4: Step Probabilities for $\beta - \delta = 1$ and $\tau = (-3, -1, 1, 3)$

Partial credit

The *Partial Credit* model (Masters, 1980; Wright & Masters, 1982) looks almost identical to the rating scale model. When looking at a single item, the models are indistinguishable. There is nothing about Figure 2 that says rating scale, not partial credit. Tables 3 and 4 can be used here just as well if an i is added to the subscript of each τ_j. If we continue to belabor the archery metaphor,

in addition to circles of different sizes, different targets may use different numbers and patterns for the concentric circles.

For the mathematically inclined, the partial credit model for the probability of person v responding in category k on item i, given the person parameter β_v and the m_i item parameters $\delta_{ij} = \delta_i + \tau_{ij}$, may be written as:

$$17. \quad p\{k \mid \beta_v, (\delta_i, j=1, m_i)\} = \frac{e^{\sum_{j=1}^{k}(\beta_v - \delta_{ij})}}{1 + \sum_{x=1}^{m_i} e^{\sum_{j=1}^{x}(\beta_v - \delta_{ij})}} = \frac{e^{k\beta_v - \sum_{j=1}^{k}\delta_{ij}}}{1 + \sum_{x=1}^{m_i} e^{x\beta_v - \sum_{j=1}^{x}\delta_{ij}}}$$

The distinction between the rating scale model (16) and the partial credit model (17) is that the category parameters, τ_j, have been subsumed under the item parameters δ_{ij}. The practical implication of this change is that the response categories can differ across items; they can be different formats or have different numbers of categories. For attitude or preference questionnaires, this may mean that different response categories are used for each statement (e.g., *agree-disagree* versus *never-always*; *four-point scales* versus *five-point*). For achievement testing, it may mean zero points are given for completely wrong answers, m_i points are given for completely right answers, and integer scores between zero and m_i are given for partially correct answers according to the item rubric, with the maximum points m_i and the scoring rubric for the partial credit specific to each item.

It is a matter of style or context whether the partial credit model is written in terms of δ_{ij} or $\delta_i + \tau_{ij}$. The first form implies the item is best described by the m_i threshold values; $\beta_v - \delta_{ij}$ looks like a generalization of the dichotomous case. The second form implies the item can be described by a single location with the thresholds given as offsets around that; $\beta_v - (\delta_i + \tau_{ij})$ looks like a generalization of the rating scale model.

From the logic of partial credit scoring, the category parameters are typically perhaps inappropriately referred to as *steps*. This is the point on the continuum where the person has completed one *step* in the problem solution, receives credit for that work, and begins work on the next *step*. As with the rating scale formulation, this is the point on the scale at which the two adjacent categories are equally likely.

Ordered categories, disordered thresholds

The categories, whether rating scale or partial credit, are always ordered: *0* always implies less than *1*; *1* implies less than *2*; *2* implies less than *3*… The concentric circle for k is always inside (smaller thus more difficult) than the circle for k-*1*. The transition points, might or might not be ordered. Perhaps the circle for k-*1* is so close in diameter to k that it is almost impossible to be inside k-*1* without being inside k.

In Figure 2 above, everything was ordered nicely; Figure 3, below, illustrates another four-point item but with disordered values of (*-3*, *0*, *-1*, *3*). Category *2* becomes more likely than *1* at a logit value of *0* but category *3* became more likely than category *2* at a logit value of *-1*.

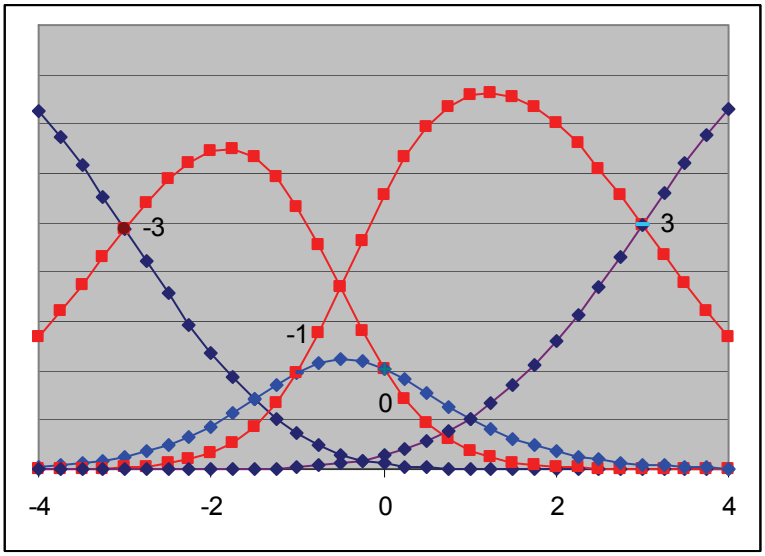

Figure 3: Rating Scale or Partial Credit with Four Categories and Disordered Thresholds: Parameters = (*-3*, *0*, *-1*, *3*)

There is no point on the continuum for which category *2* is the most likely response. The person who is most likely to be in category *2* has a logit location of -0.5; however, a person at this location is more likely to be in either category *1* or *3*. A person who is strong enough to leave category *1* is unlikely to stop at *2* but is expected to go immediately to *3*. In spite of this confusion of parameters, the category curves are still in the natural order: being in *2* implies more than being in *1* and less than being in *3*.

There is some controversy about how disordered thresholds should be interpreted[v]. Masters, arguing from the mathematics, contends that nothing in the model is violated and it simply reflects an under-used category, which can be informative, if less than optimal. Andrich, relying on his experience with rating scales, feels the disordering is a serious violation of Rasch's principles and indicates a problem with the data that must be explored at the very least, the item should be revised or discarded[vi]. For Masters, the concentric circles are very close to the same size; for Andrich, they might not be concentric or circles.

Many faceted Rasch models

The discussion thus far has considered the measurement problem with an *object* of measurement (e.g., candidate, student, patient, archer, rock) and an *agent* of measurement (e.g., test, questionnaire, checklist, bull's-eye, rock). Anyone, even a computer, can consult the scoring key and unambiguously assign a score to any response. Many situations are not so mechanical and there is some judgment involved in assigning the score. No matter how well-trained and conscientious, judges will sometimes differ. Part of this variation may be random, but part may also be due to real differences in the temperament or harshness of the judges. Either form of *judge* effect can be modeled and estimated just as well as the objects and agents (Linacre & Wright, 2004).

For the archery example, rather than employing near-sighted judges, we could introduce another type of facet, say, the distance between the archer and the target, which is not inherent to either the archer or the target, but that would affect the difficulty of the task and likelihood of success, in

addition to the archer's skill and the target's size. Distance could affect the likelihood that anyone will hit a target and should be separable. This would facilitate comparison among archers who shot at targets presented at different distances.

The probability for the dichotomous model with three facets *looks* very familiar but don't let it fool you:

18. $p(pass \mid \beta_v, \delta_i, \phi_j) = \frac{e^{\beta_v - \delta_i - \phi_j}}{1 + e^{\beta_v - \delta_i - \phi_j}}$, where β_n is the skill of person v, δ_i is the difficulty of item i and ϕ_j is the harshness of judge j.

Expression (18) only looks like the rating scale model (cf. expression 11) on the right side; it is <u>not</u> the same. The data from a rating scale comprise a two-way table: N people by L items. The entry in the table is the person's score, 0 to m, on the item. The rating scale model is the probability of observing a score of k from person v on item i.

The dichotomous facets model uses a three-way table[vii]: N people by L items by J judges; the table entry is *pass* or *fail*. Expression (19) is the probability of observing a score of *pass* for person v on item i as determined by judge j.

For simplicity in contrasting models, it is useful to express the model in terms of the logit, the log odds favoring a score of x over x-1; in the dichotomous case just presented, the log odds of a score of 1 over a score of 0 is:

19. $\ln(p_{vij1} / p_{vij0}) = \beta_v - \delta_i - \phi_j$

Most assessments involving judges use more elaborate scoring than just 1 or 0. Typically, the judges are applying a rubric that results in polytomous scores. If there is a common rubric that applies to all items, the facets model is an extension of the rating scale model, with the logit expressed as:

20. $\ln(p_{vijk} / p_{vij(k-1)}) = \beta_v - \delta_i - \phi_j - \tau_k$, which adds a parameter for each point in the rubric.

Additional complexity can be introduced by allowing different rubrics for different items. Expression (22) is a partial credit with judges:

21. $\ln(p_{vijk} / p_{vij(k-1)}) = \beta_v - \delta_i - \phi_j - \tau_{ik}$.

This allows different rubrics for each item but still requires the judges to apply the rubrics in the same way. Additional forms of the facet model could be envisioned that allow the judges to interpret the rubric personally, τ_{jk}, or personally for each item, τ_{ijk}. Collecting data that adequately connect all components of the model can become an issue, which we will return to.

Linear logistic test model (LLTM)

There is another branch to the Rasch family tree that is better known in Europe than the US and may subsume everything that's gone before (Fischer, 1973, 1995a; Fischer & Molenaar, 1995). The task of hitting any target with an arrow has an inherent difficulty that could be estimated readily with a suitable field trial and data analysis. It may, however, be useful to think about the difficulty of

each task, not as a vague amalgam of unspecified characteristics, but as a specific linear combination of more basic components that describe what makes a target difficult to hit. Important components might include, for example, size of the bull's-eye, distance from the archer, and elevation of the target.

This is thinking about targets differently than we were with the multi-faceted model. Then the target was a piece of cardboard with circles drawn on it. All other aspects of the task were looked at as separate issues that were treated as other facets, or standardized and randomized away, or ignored. Now, the target is the entire task including the circles, the distance, and the elevation; we might also include things not under our control entirely like wind and lighting.

If the basic components and their linear combinations can be specified, the difficulty of any target can be decomposed into the basic operations:

$$22. \qquad \delta_i = \sum_{j=1}^{p} w_{ij}\eta_j,$$

where η_j is a parameter associated with operation j, and w_{ij} is a coefficient that indicates the role of operation j for item i. The weights w_{ij} are established *a priori* and are not estimated.

The decomposition can be imposed on any of the Rasch models discussed above, but for the dichotomous case, the probability of observing a response of *1* is:

$$23. \qquad p(x_{vi}=1 \mid \beta_v, \delta_i) = \frac{e^{\beta_v - \delta_i}}{1+e^{\beta_v - \delta_i}} = \frac{e^{\beta_v - \sum_{k=1}^{p} w_{ik}\eta_k}}{1+e^{\beta_v - \sum_{k=1}^{p} w_{ik}\eta_k}}$$

While, at first blush, this may not seem to make things easier, typically there are far fewer components than items so it can be a very parsimonious description of the instrument.

At the risk of stretching the archery analogy too far, we'll propose three two-level components for the difficulty of a target: 122 cm. or 60 cm. diameter targets, 30 m. or 90 m. distance, and level or downhill range. The eight distinct targets can then be described with four parameters as:

$$24. \qquad [\delta_i] = \left[\sum_{j=0}^{3} w_{ij}\eta_j\right] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \eta_0 = (base = 122cm.,30m.,level) \\ \eta_1 = (size = 60m) \\ \eta_2 = (distance = 90m) \\ \eta_3 = downhill \end{bmatrix}$$

The linear logistic test model (LLTM) was formalized by Gerhardt Fischer (1973, 1995a). The idea originated from consideration of the *cognitive operations* required to solve math test problems. Scheiblechner (1972) decomposed the items into seven operations (negation, disjunction, conjunction, …). The item's difficulty was then expressed as a linear combination of these basic operations.

This design matrix does not simply indicate the presence of a condition, as in expression (24), but indicates the number of times each operation is required in the problem's solution. For example, a possible decomposition of a specific set of items might be:

$$25. \qquad [\delta_i] = \left[\sum_{j=0}^{3} w_{ij}\eta_j \right] = \begin{bmatrix} 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 2 & 1 & 0 & 0 & 0 \\ \dots & & & & & & \end{bmatrix} \begin{bmatrix} \eta_1 = (negation) \\ \eta_2 = (disjunction) \\ \eta_3 = (conjunction) \\ \eta_4 = \\ \eta_5 = \\ \eta_6 = \\ \eta_7 = \end{bmatrix}$$

Typically, the number of possible operations, p, is much smaller than the number of items so LLTM can provide a very parsimonious description of the items. It can be the basis of a fuller understanding of what makes an item difficult and suggests the possibility of generating new items at precise levels of difficulty. A likelihood ratio test can readily determine if the decomposition into basic operations is an adequate description.

LLTM has also been applied to good advantage in the measurement of change. While somewhat counter-intuitive, Fischer (1995b) formulated this problem as a change in the item difficulties rather than a change in person abilities. If the same *L* items are used as the pre- and post-test[viii], each person is considered to have taken *2*L* items. The reparameterization for LLTM uses *L+1* parameters: one for each item and one for the pre-post effect.

The appeal of this solution is that it is a Rasch model; specific objectivity obtains and sufficient statistics exist for the nuisance parameters. The effect of the intervention between the pre- and post-tests is estimated consistently with no interference from the sample of students involved. When the estimation is done with conditional maximum likelihood (Mair & Hatzinger, 2007), likelihood ratio tests are naturally available for a variety of interesting hypotheses.

Summary II

Actually using these simple models is more demanding than doing the math. It requires building instruments with items that are all *equally valid*, *albeit imperfect*, *instances* of the idea being measured. Hence, except for consideration of their relative difficulties, the items are completely interchangeable. This level of uniformity requires careful control of the item development process, the item culling process and the item administration process. It does not mean that the items must or even should be homogeneous in their approach, content, or format.

The point of all these deliberations is measurement; there is something about a person, or other object, that we wish to quantify, measure, and analyze. This means developing and validating an instrument to provoke the person into providing us with relevant observations. What we have discussed in Part II is an assortment of models that can transform observations of various types into measures. Part III will finally discuss actually doing the arithmetic.

Part II Notes

[i]All models are wrong. Some are useful. *G.E.P.Box* Models must be used but must never be believed. *Martin Bradbury Wilk*

[i]A little less frivolously, this may be expressed as a normalization $\sum_{i=1}^{L} a_i \delta_i = c$, where $\sum_{i=1}^{L} a_i = 1$. Any transformation to a more user-friendly scale score metric is usually done at a later step.

[iii]The distribution does matter in one very practical way. The people who answer both items right (lower right cell of Table 2) or both wrong (upper left) provide no information about which item is harder. Consequently, if the test is badly off target for the sample, a larger sample is needed to achieve the same precision.

[iv]The γ term in Table 3 is a normalizing constant to make the probabilities sum to one. It is nothing more or less than the sum of the numerators. If we were being more rigorous, a form of this constant would be introduced in expressions (12) and (13).

[v]This summary of their positions is based on a number of conversation, both private and public, with Drs. Geoff Masters and David Andrich. See Andrich (2004) for a fuller discussion of his view.

[vi]Or treated with a different model, perhaps unfolding (Andrich, 1995) or multidimensional (Briggs & Wilson, 2004); neither topic is discussed here.

[vii]More facets are possible but we won't take them on.

[viii]Incomplete designs and designs with more than two time points are also possibilities but will not be considered here. See Fischer (1995b).

[ix]*Nuisance parameters* mean the abilities in this discussion.

Part III: Rasch Methods and Arithmetic

Doing the Math

Specific objectivity is the common thread that runs through everything and the basic equation that *the observations equal the expectations* is a recurring theme. The approach we are taking relies naturally on the so-called *joint maximum likelihood* estimation (UCON) of Wright and Panchapakesan (1969) and focuses on mean squares and person-item residuals. Everything here could be approached using the fully-conditioned estimation methods and likelihood ratios tests just as well (Fischer & Molenaar, 1995; Andersen, 1973). Our experience has been, if the data have something to say, the two approaches deliver the same message.

The minor payoff from applying a Rasch model, compared to what one needs to go through with IRT, is that the arithmetic involved borders on the trivial. This section will show some basic calculations for estimating ability and difficulty parameters (calibration), evaluating the consequences of specific objectivity (control), and building a measurement scale (linking, equating, scaling). There are alternative approaches, some more efficient or more powerful, but this is the basic idea and it's a start.

Defining the Variable

The major payoff is establishing the variable to be measured and developing instruments to measure it that conform to Rasch's principle of specific objectivity. The practical definition of the variable is the tasks we use to provoke the person into providing evidence. Items that are hard to get right, tasks that are difficult to perform, or statements that are rarely agreed to will define the high end of the scale; easy items, simple tasks, or popular statements will define the low end. The order must be consistent with what would be expected from the theory that guided the design of the instrument in the first place. Topaz is always harder than quartz regardless of how either is measured. If not, the items may be inappropriate or the theory wrong[i]. The structure that the model provides should guide the content experts through the analysis, with a little help from their friends.

Table 5 shows the results of a hypothetical archery competition. The eight targets are described in the center panel. It is convenient to set the difficulty of the base target (i.e., largest bull's-eye, shortest distance and level range) to zero. The scale is a completely arbitrary choice; we could multiply by 9/5 and add 32, or whatever, if that were more convenient. The most difficult target was the smallest bull's-eye, longest distance, and downhill range. Any other outcome would have raised serious questions about the validity of the competition or the data.

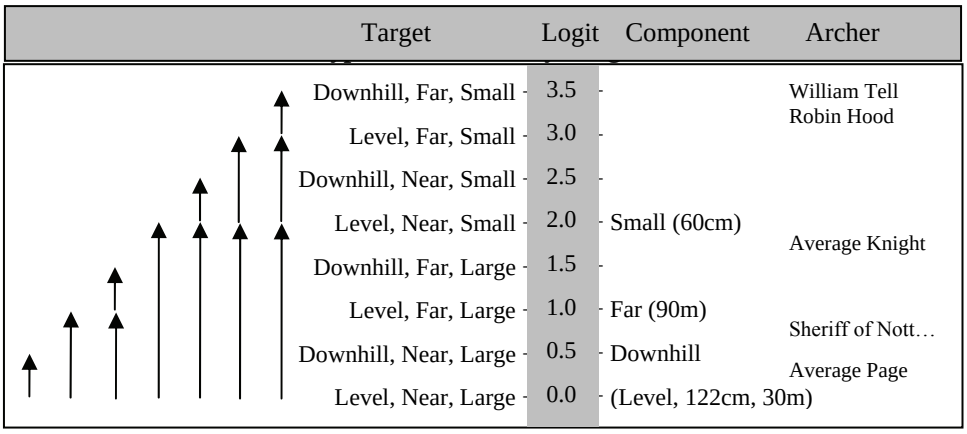

Table 5: A Definition of Bowmanship

The relative difficulties of the basic components of target difficulty are just to the right of the numeric logit scale: shooting downhill added 0.5 logits to the base difficulty; moving the target from 30 m. to 90 m. added 1.0 logits; and reducing the diameter of the bull's-eye from 122 cm to 60 cm added 2.0 logits.

The role of specific objectivity in this discussion is subtle but crucial. We have arranged the targets according to our estimated scale locations and are now debating among ourselves if the scale locations are consistent with what we believe we know about bowmanship. We are talking about the scale locations of the targets, period, not about the scale locations of the targets for knights or pages, for long bows or crossbows, for William Tell or Robin Hood.

While it may be interesting to measure and compare the bowmanship of any and all of these variations and we may use different selections of targets for each, those potential applications do not change the manner in which we define the variable *bowmanship*. The knights and the pages may differ dramatically in their ability to hit targets and in the probabilities that they hit any given target, but the targets must maintain the same relationships, within statistical limits, or we do not know as much about bowmanship as we thought.

The symmetry of the model allows us to express the measures of the archers in the same metric as the targets. Thus, after a competition that might have used different targets for different archers, we would still know who won, we would know how much better Robin Hood is than the Sheriff, and we would know what each is expected to do and not do. We could place both on the bowmanship continuum and make defendable statements about what kinds of targets they could hit.

Calibration

Estimating Logit Abilities:

This is almost beginning at the end and working backward but we will start with estimating the logit abilities (Wright & Stone, 1979).

The process assumes the logit difficulties are known (i.e., good estimates of the logit difficulty parameters are available); no additional data are needed. The ability estimate b_r associated with the raw score r is the value that satisfies the basic equation:

$$26. \qquad r = \sum_{i=1}^{L} E(x_{ri}) \ ,$$

where r is a raw score from 1 to $M-1$, M is the maximum number of points possible, L is the total number of items, $E(x_{ri})$ is the expected score on item i for a person with the ability associated with r, and the summation is over the items the person took. Because total raw score is the sufficient statistic for estimating ability, everyone who took the same items and got the same raw score gets the same estimated ability b_r. Hence the estimate can be indexed by the raw score, instead of the person.

That's all there is to computing logit abilities; the rest is detail. In the dichotomous case,

$$27. \qquad E(x_{ri}) = p_{ri} = \frac{e^{b_r{}^t - d_i}}{1 + e^{b_r{}^t - d_i}} \ ,$$

where d_i is the difficulty estimate for item i, assumed known at this point, and $b_r{}^t$ is the current and tentative estimate of ability associated with raw score r.

Equation (26) simply says the expected total score Σp_{ri} is equal to the observed total score r; if they aren't equal enough, the ability estimate needs adjusting. If the expected score is low, the estimated ability is increased; if the expected score is too high, the estimate is decreased. The ability estimate is adjusted by your favorite numeric method until equation (26) is satisfied. Wright & Panchapakesan (1969) applied Newton's method to give the iterations:

28. $$b_r^{t+1} = b_r^t + \frac{r - \sum_{i=1}^{L} p_{ri}}{\sum_{i=1}^{L} p_{ri}(1-p_{ri})},$$

An effective starting value for this process is:

29. $$b_r^0 = \ln\left(\frac{r}{M-r}\right) - \bar{d}$$

where $\bar{d} = \frac{\sum_{i=1}^{L} d_i}{L}$ is the center of the item difficulties, which is often zero.

Equation 29 makes it obvious, but it also follows more subtly and more profoundly from equation (26), that perfect scores, both $r=0$ and $r=M$, are problems. There is no ability low enough to ever satisfy equation (26) when r is 0, nor high enough when r is M. In the real world, it is generally necessary to manufacture something to report for examinees with these scores. One tactic is to solve the equations for non-integer scores arbitrarily close to the perfect scores, say, within 0.25. Whether the target should be off by 0.25, or 0.1, or 0.33, or some other value is completely arbitrary; the smaller the value, the more extreme the solutions will be. It is more a policy decision than psychometric issue.

Another strategy, perhaps with slightly more psychometric motivation, produces almost the same results by assigning the logit ability to a raw score of zero that is the logit ability for a raw score of one minus the squared standard error of measurement:

30. $$b_0 = b_1 - s_1^2.$$

Analogously for a perfect score of M, the logit ability estimate is the estimate for a score of $M-1$ plus the squared standard error. The simple rationale for this tactic is that the difference between logit ability estimates for any adjacent scores looks approximately equal to the squared standard error of measurement. The more sophisticated rationale is that this is equivalent to using expression (28) to estimate the ability for zero (or M-1) and stopping after the first iteration.

Table 6 shows the arithmetic for a small test with 10 dichotomous items. It is typical for this process to stabilize in two or three iterations. The standard error for the logit ability is the inverse of the square root of the sum of p(1-p).

Logit Difficulties	Raw Score	Initial Logit	Round One	Round Two	Std Error
0.637	0			-3.49	1.74
-0.941	1	-2.197	-2.339	-2.347	1.071
-0.266	2	-1.386	-1.496	-1.499	0.814
0.382	3	-0.847	-0.922	-0.923	0.716
-0.455	4	-0.405	-0.444	-0.444	0.674
0.086	5	0	-0.001	-0.001	0.661
-0.881	6	0.405	0.441	0.442	0.674
0.000	7	0.847	0.92	0.921	0.717
0.297	8	1.386	1.496	1.499	0.815
1.141	9	2.197	2.341	2.349	1.073
	10			3.5	1.74
	Sum of p				
	1	1.138	1.007	1	
	2	2.175	2.005	2	
	3	3.149	3.002	3	
	4	4.085	4	4	
	5	5.003	5	5	
	6	5.92	6	6	
	7	6.854	6.998	7	
	8	7.826	7.995	8	
	9	8.86	8.993	9	
	Sum of p(1-p)				
	1	0.972	0.876	0.871	
	2	1.601	1.513	1.510	
	3	1.997	1.949	1.948	
	4	2.217	2.204	2.204	
	5	2.287	2.287	2.287	
	6	2.215	2.202	2.202	
	7	1.993	1.945	1.945	
	8	1.596	1.509	1.506	
	9	0.971	0.874	0.869	

Table 6: Calculations of Logit Abilities for a Test with 10 Dichotomous Calibrated Items

Extending the calculations to a polytomous case requires extending the definition of the expected score used in equation (26):

31. $$E(x_{ri}) = \sum_{k=1}^{m_i} k p_{rik} ,$$

where the case of $m_i = 1$ is the dichotomous case and, for example,

$$p_{rik} = \frac{e^{\sum_{j=1}^{k}(\beta r - \delta_{ij})}}{1 + \sum_{x=1}^{m_i} e^{\sum_{j=1}^{x}(\beta r - \delta_{ij})}}$$

is the estimated probability that a person with total score r will receive a score of k on partial credit item i.

Estimating Logit Item Difficulties:

The truly elegant estimators for logit difficulties that completely condition out the sufficient statistics (Rasch, 1960; Fischer & Molenaar, 1995; Andersen, 1973) were computationally laborious and time consuming on the computers of the 1960's. Wright and Panchapakesan (1969) proposed a less elegant, slightly biased, and much faster method that became the standard of Rasch analyses in the U.S. (Wright and Stone, 1979). It is the item difficulty analog to the person ability estimator of equation (26) above.

32. $$S_i = \sum_{v \in R_{min}}^{R_{max}} E(x_{vi}),$$

where S_i is the total number of points accumulated on item i by people who took the item and who have raw scores between a minimum score R_{min} and maximum score R_{max}. If everyone takes all the same items, it is generally faster to sum over raw scores rather than the people. Then, for the dichotomous case, if n_r is the number of people with score r:

33. $$S_i = \sum_{r=R_{min}}^{R_{max}} n_r E(x_{ri}) = \sum_{r=R_{min}}^{R_{max}} n_r p_{ri}.$$

This process uses two vectors of data: S_i, the item scores, and n_r, the number of people at each raw score, in addition to some estimate of logit abilities.

R_{min} and R_{max} could be one and one less than perfect respectively, but R_{min}, in particular, is often set a little higher to avoid the noise commonly found with people with very low raw scores. This again can be solved by almost any numeric method but Newton's method works well:

34. $$d_i^{t+1} = d_i^t - \frac{S_i - \sum n_r p_{ri}}{\sum n_r p_{ri}(1 - p_{ri})}$$

for the dichotomous case, with a starting value of:

35. $$d_i^o = \ln\left[\frac{N - S_i}{S_i}\right] - \bar{d},$$

where N is the total number of people in the R_{min} to R_{max} range.

Like zero and perfect scores, items that everyone gets right or everyone gets wrong can not be estimated but can be gotten around with strategies similar to the ones used for abilities.

The Wright-Panchapakesan method does not *condition* out the nuisance parameters but rather *estimates* them away. While computing the difficulty estimates, the abilities are assumed to be known. When the difficulties stabilize, attention is turned to the abilities. Then the difficulties are assumed known and revised abilities are computed using equation 3. The process continues alternating between estimating difficulties and estimating abilities until a reasonable convergence criterion is met.

The bias alluded to earlier is, on average, equal to M/(M-1), (Andersen, 1973). This is analogous to the bias in the maximum likelihood estimate of the variance and arises for a similar reason: assuming parameters are known when they are only estimated.

Tables 7 and 8 show the results of a small simulation with ten items and one hundred students to illustrate the calculations. Table 7 has the data: total item scores for the ten items and the counts of students at each raw score. Table 8 has the step-by-step summary for the item difficulty estimation cycle. The initial logits are the log odds for the reduced item scores centered on zero, e.g., $\ln[40/(95-40)]-\bar{d}$. The *reduced* item scores have simply eliminated the five students with a perfect score.

Raw Score	Count n_g	Item	Item Score S_i	Item Score Reduced
0	0	1	45	40
1	1	2	79	74
2	6	3	66	61
3	8	4	51	46
4	12	5	70	65
5	16	6	58	53
6	20	7	78	73
7	10	8	60	55
8	9	9	53	48
9	13	10	34	29
10	5			

Table 7: Vectors of Counts Needed To Estimate Difficulties

Table 8 shows two rounds of iterations needed to equate the observed items scores (reduced) to the expected item scores (sum of p). The ability estimates used to obtain the sum of p and pq were taken from Table 6. Additional cycles would mean recalculating the logit abilities using the revised difficulty estimates; the process continues until the difficulties are stable, which rarely takes more than two or three cycles with dichotomous items. The final steps in this activity are:

- adjusting the logit difficulties for bias by multiplying by $(M\text{-}1)/M$ and
- calculating a final round of logit abilities and standard errors using the final adjusted difficulties.

Initial Logits	Round One	Round Two	Sum of P			Sum of PQ		
0.637	0.783	0.784	42.71	40.02	40.00	18.62	18.32	18.32
-0.941	-1.151	-1.159	70.84	73.88	74.00	15.09	13.91	13.86
-0.266	-0.345	-0.346	59.58	60.99	61.00	18.02	17.77	17.76
0.382	0.462	0.462	47.51	46.00	46.00	18.88	18.83	18.83
-0.455	-0.575	-0.576	62.92	64.97	65.00	17.37	16.88	16.87
0.086	0.090	0.090	53.09	53.00	53.00	18.76	18.77	18.77
-0.881	-1.081	-1.088	69.92	72.89	73.00	15.41	14.31	14.27
0.000	-0.016	-0.017	54.69	55.00	55.00	18.64	18.61	18.61
0.297	0.356	0.356	49.10	48.00	48.00	18.89	18.89	18.89
1.141	1.410	1.420	33.64	29.15	29.00	17.22	16.10	16.06

Table 8: One Cycle of Item Difficulty Estimation

Standard Errors of Estimation

A reasonable value for the standard error of estimation for either logit difficulties or logit abilities is the reciprocal of the square root of $\Sigma p(1\text{-}p)$, i.e. for ability at a score of r:

$$36. \qquad s_r = \frac{1}{\sqrt{\sum_{i=1}^{L}\{E(x_{ri}^2) - [E(x_{ri})]^2\}}} = \frac{1}{\sqrt{\sum_{i=1}^{L} p_{ri}(1-p_{ri})}}$$

The relevant standard error of estimation for an item difficulty is computed with the analogous expression:

$$37. \qquad s_r = \frac{1}{\sqrt{\sum_{r=R_{min}}^{R_{max}} n_r\{E(x_{ri}^2) - [E(x_{ri})]^2\}}} = \frac{1}{\sqrt{\sum_{r=R_{min}}^{R_{max}} n_r p_{ri}(1-p_{ri})}}$$

Wright and Douglas (Wright & Stone, 1979) devised a rule-of-thumb estimate for the standard error, useful for designing tests and determining sample sizes. The maximum value of 0.25 for $p_{ri}(1\text{-}p_{ri})$ occurs when $p_{ri} = 0.5$. If every p_{ri} took this value, then $s_i = 2/\sqrt{N}$, which represents the best case with the item targeted perfectly for every person. A more conservative value might be $s_i = 3/\sqrt{N}$, which assumes p_{ri} about 0.9. Wright opted for a middling value of $s_i \approx 2.5/\sqrt{N}$.

A similar formulation applies to the ability estimates as well: $s_r \approx 2.5/\sqrt{M}$, which can be extended to give an approximate reliability with one more sweeping assumption. In many situations, the observed *within grade* standard deviation of ability estimates is approximately one logit. Then a reasonable approximation of reliability is:

$$38. \qquad \rho = \frac{\sigma_b^2 - \bar{\sigma}_r^2}{\sigma_b^2} \cong \frac{1 - 6/M}{1} = \frac{M-6}{M},$$

where M is the maximum total points possible and 6 is close enough to the square of 2.5. This can be turned around to give the test length needed for a given reliability $\hat{\rho}$:

39. $M = 6/(1-\hat{\rho})$.

All of this assumes reasonable items of equal quality.

Linking and Equating

Invoking the full power of Rasch models means defining a useful construct and developing a pool of calibrated items that measures it. Equation (26) for estimating logit abilities can be solved for any selection of calibrated items. An ability estimate from any subset of the calibrated items can be compared directly to estimates based on any other subsets taken from the same pool. In that sense, all possible forms and scores from the pool are *equated*. All this assumes that the pool conforms to Rasch's requirements; that is, composed entirely of equally valid and reliable items.

The terms, *link* and *equate*, are often used interchangeably. If there is a distinction within the context of Rasch measurement,

- *Linked* means two forms are connected through some common element, either overlapping items or a shared group of examinees.
- *Equated* means the (logit) scores from one form can legitimately be compared to (logit) scores from the other form.

Linked implies a connection exists between the forms; hence, they are capable of being equated. Equated implies the forms were linked and all the necessary work has been done[ii].

In principle, two forms can be linked with a single item (or a single person). Because it is the same item, any difference in the item's logit difficulty estimate, beyond random error, is due to a difference in the arbitrary origins of the forms. With an easy form, centered on zero, the link item could have a positive logit, implying it is harder than the average. With a difficult form, also centered on zero, the link item may have a negative logit, implying it is easier than the average. The goal of an equating analysis is to eliminate that arbitrary difference by adjusting the logit estimates on one form so that the link item has the same numeric value in both contexts.

Assuming our link item has a difficulty estimate of d_{iA} when calibrated with form A and an estimate of d_{iB} when calibrated with form B, then the form B estimate can be made to equal the form A estimate by subtracting d_{iB} and adding d_{iA}.

If we do it to one item on form B, we need to do it to every item to maintain their relative positions. The two forms are equated by adding $t = d_{iA} - d_{iB}$ to every logit on form B. The adjusted form B difficulty for the linking item is:

40. $d^*_{iB} = (d_{iB} + t) = d_{iB} + d_{iA} - d_{iB} = d_{iA}$.

In practice, it may not be a good idea to equate through a single item. With several link items, the translation constant t is simply the difference between the means of the link items from the two calibrations.

41. $$t = \bar{d}_{ALink} - \bar{d}_{BLink} = \frac{\sum_{i \epsilon Link} d_{iA}}{n_{Link}} - \frac{\sum_{i \epsilon Link} d_{iB}}{n_{Link}} = \frac{\sum_{i \epsilon Link} (d_{iA} - d_{iB})}{n_{Link}},$$

where n_{Link} is the number of link items.

The equating constant t is added to every item on form B as before. After adjustment, the mean of the link set will be identical in both contexts. The only hard part is remembering when to add and when to subtract.

The number of items that should be in the link, like any sample size, depends on how precisely you need to know the answer. Using the Wright-Douglas approximation, the squared standard error for t is:

42. $$se_t^2 = \frac{1}{n_{Link}} \left\{ \frac{6}{N_A} + \frac{6}{N_B} \right\},$$

where N_A is the number of students used in the form A calibration and N_B is the number used in the form B calibration. Turning the relationship around, and assuming equal sized calibration samples, provides the link length needed for a given standard error.

43. $$n_{Link} = \frac{12}{se_t^2 N}.$$

Inconveniently, limitations on the test length and on item exposure often have more to do with the size of the link then does the magnitude of an acceptable standard error needed to make the psychometricians happy. And most get very uncomfortable when n_{Link} is as low as 10 items, the impeccable logic of equation (43) notwithstanding.

Multiple link items, in addition to increasing the precision of the estimate of the equating constant, can be used for control of the process. Each item pair provides an estimate of the equating constant. Like statistics everywhere, they will not be identical. The problem is to recognize and eliminate outliers from the calculations of the means[iii]. There are a number of more or less heuristic techniques for dealing with the uncertainty.

Figure 4 shows the simplest possible link analysis. The data plotted are the logit difficulties from the bank and the logits obtained from a calibration of the current administration. The data should follow a slope one line with the intercepts representing the required translation constant. This example is a very clean link, with one outlier, and an x-intercept of 0.5, ignoring the outlier. Adding 0.5 to every current logit will shift the plot vertically so that the slope one line passes through the origin and the current administration has been equated to the bank[iv].

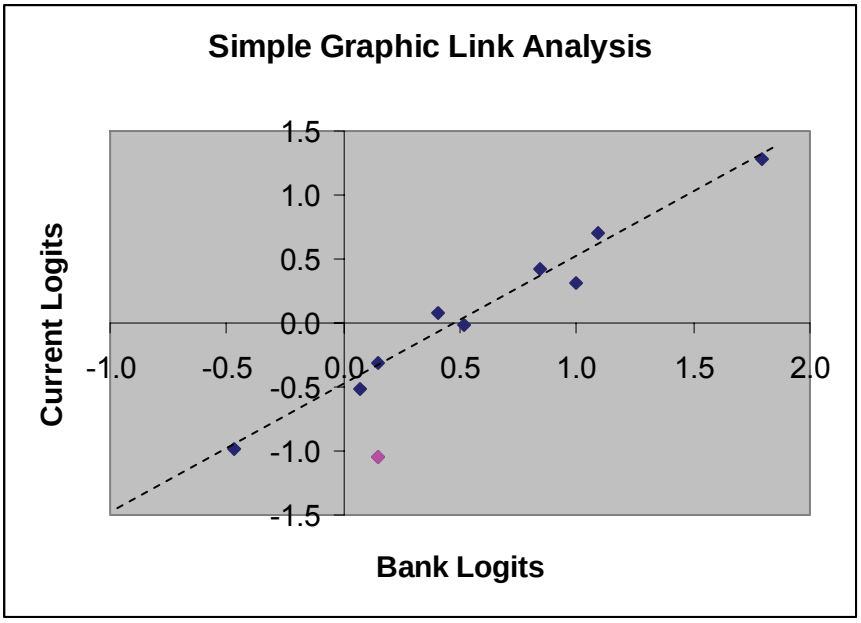

Figure 4: Sample Link Analysis

The same analysis and the same result can be achieved with the appearance of more rigor if we use tables and numbers. The simplest non-graphical method is to choose a criterion logit value and reject any item from the link if its estimate $t_i = d_{iA} - d^*_{iB}$ is larger than the criterion in absolute value. The most commonly mentioned criterion is 0.3 logits. This is easy to apply but is criticized because it ignores the standard errors of estimation; it applies the same criterion regardless of how well we know the item's difficulty.

A little more stringent strategy performs a Student's t-test on each item using the standard errors from expression (37). Items are rejected if the *t-statistic* is larger than the psychometrician's tolerance level. This is criticized because it doesn't ignore the standard errors; it is more tolerant of discrepancies when they come from poorly estimated items.

Our final strategy, which incorporates the weaknesses of both, uses a simple robust estimate of the standard error of the differences and applies it uniformly across the items. A *robust-z* is computed as:

$$44. \qquad z_i = \frac{t_i - Q_2}{0.74(Q_3 - Q_1)} ,$$

where Q_1, Q_2, and Q_3 are the first, second, and third quartiles of the distribution of the t_i. An item is rejected if its t_i is greater than 1.645 in absolute value. This approach is criticized because it will almost always find outliers; if the items are very consistent, that's a good thing but $Q_3 - Q_1$ will be very small. To provide a rational stopping rule, no items are dropped after the ratio of standard deviations of the two sets of difficulties is between 0.9 and 1.1 and the correlation between the two sets of difficulties is at least 0.95.

Dropping an item from the link does not imply that the item must be dropped from the test. It does imply some violation of specific objectivity because that item is not consistent across forms; the item still may have functioned acceptably in both situations, although differently. The first thing to check

is that the difficulties have been matched correctly. If so, it may be the item was unusually amenable to instruction or it may mean it interacted with popular culture (e.g., movies, commercials, music lyrics, current events) in some way others did not. Or it may mean there was a security breach. It is good for the psychometrician's peace of mind and often informative to identify the source of any disturbance.

Table 9 contains a summary of these analyses for the same data used in Figure 4. The Pool and Current logits are given. The *Difference* is the *Pool – Current*; the *Adjusted* is the *Current + average Difference*; and the *Discrepancy* is the difference between the *Pool* and *Adjusted* logits. The *Student's t-statistic* is *Discrepancy* divided by the standard error from the data of table 8[v]. The *Robust Z* is defined by equation (44).

First Round: All Items	Pool d_{iA}	Current d_{iB}	Difference $d_{iA}-d_{iB}$	Adjusted $d_{iB}+t$	Discrepancy $d_{iA}-(d_{iB}+t)$	Student's t-statistic	Robust Z
1	1.089	0.705	0.383	1.266	-0.177	-0.76	-1.23
2	0.149	-1.043	1.193	-0.483	**0.632**	**2.35**	**6.57**
3	0.148	-0.311	0.459	0.250	-0.102	-0.43	-0.50
4	0.844	0.415	0.429	0.976	-0.132	-0.57	-0.79
5	0.074	-0.519	0.592	0.042	0.032	0.13	0.78
6	0.402	0.081	0.320	0.642	-0.240	-1.04	-1.84
7	-0.472	-0.979	0.508	-0.419	-0.053	-0.20	-0.03
8	0.515	-0.015	0.529	0.546	-0.031	-0.13	0.18
9	0.998	0.320	0.678	0.881	0.118	0.51	1.61
10	1.792	1.278	0.514	1.838	-0.046	-0.19	0.03
Mean	0.554	-0.007	t = 0.561		Q2 = -0.05	-0.03	
Std Dev	0.644	0.732	0.244		Q3 = 0.02	0.90	
Ratio SDs	0.88				Q1 = -0.12		
Correlation	0.94						
Second Round: Drop Item 2							
1	1.089	0.705	0.383	1.196	-0.107	-0.46	-1.67
2	0.149	-1.043		-0.553			
3	0.148	-0.311	0.459	0.179	-0.031	-0.13	-0.65
4	0.844	0.415	0.429	0.906	-0.062	-0.27	-1.06
5	0.074	-0.519	0.592	-0.028	0.102	0.42	1.14
6	0.402	0.081	0.320	0.572	-0.170	-0.74	-2.52
7	-0.472	-0.979	0.508	-0.489	0.017	0.07	0.00
8	0.515	-0.015	0.529	0.476	0.039	0.17	0.29
9	0.998	0.320	0.678	0.810	0.188	0.82	2.29
10	1.792	1.278	0.514	1.768	0.024	0.10	0.71
Mean	0.599	0.108	**t = 0.490**		Q2 = 0.02	0.00	
Std Dev	0.667	0.674	0.108		Q3 = 0.04	0.44	
Ratio SDs	0.99				Q1 = -0.06		
Correlation	0.99						

Table 9: Sample Link Calculations

Using all items, the ratio of standard deviations is 0.88 and the correlation is 0.94. These imply something should be dropped. Only item two is eligible with a *robust z* larger than 1.645 (and a discrepancy larger than 0.3 logits and a *Student's t* twice as big as anything else.) Dropping this item changes both the SD ratio and the correlation, coincidently, to 0.99, which implies we are finished. The end result, no matter how we did it, was we dropped one item and added 0.49 to the current logits to place them on the pool logit scale. (The *Student's t* was the weakest test in this situation but the calibration was based on only 100 examinees.)

Because all the criteria for a satisfactory link (no discrepancy larger than 0.3, correlation greater than 0.95, ratio of standard deviations between 0.9 and 1.1) are met, there is no reason to have computed the *robust z* statistics for the second round. However, having done it, there are two that are larger than 1.645. This is the nature of this calculation; there will always be a most extreme value. If all the other criteria are met, we end the process, report the scores, and go to dinner.

Multiple Link Forms

If we can equate two forms, or one form to a bank, we can also equate multiple forms with multiple interconnections. We can use this same analysis to proceed one link at a time until eventually the entire network is equated. Any redundancies can be used to monitor and control the process. For example, linking form A to form C should give the same result as linking form A to form B to form C. Or, alternatively, *A to B to C to A* should bring us back to where we started and result in a zero shift, within statistical limits.

There is a straightforward least squares path to resolving any inconsistencies due to random noise. If we have k forms and there is a link t_{ij} between each pair (i,j), then summing all the values for form i,

$$45. \qquad T_i = \sum_{j=1}^{k} t_{ij} = \sum_{j=1}^{k} (t_i - t_j) = kt_i - \sum_{j=1}^{k} t_j$$

where T_i is the sum of all form i links, t_{ij} is the link for form i to form j, and t_i is the general equating constant for form i that we are after. Then, if all form-to-form links are present and if we let $\sum_{j=1}^{k} t_j$, the equating constant for form i is simply:

$$46. \qquad t_i = \frac{T_i}{k} .$$

If not all the links are present, the solution is a little more complicated involving at the worst a system of k simultaneous equations. In matrix form,

$$47. \qquad \mathbf{A}\underline{\boldsymbol{t}} = \underline{\boldsymbol{T}},$$

where $\underline{\boldsymbol{T}}$ is a kx1 vector of the row sums T_i from equation (45), and $\underline{\boldsymbol{t}}$ is the vector of equating constants we are after. If all the links are present, $\mathbf{A}$ is a diagonal matrix with the value k along on the diagonal and expression (46) is the solution to expression (47).

In general, however, $\boldsymbol{A}$ is symmetric with:

48. $a_{ij} = 0$ if the link (i,j) is present and 1 if not, for $i \neq j$, and

49. $a_{ii} = k - m_i$, where k is the number of forms and m_i is the number of missing links for form i.

This tactic of *completing the sum* can be used in a variety of situations involving paired comparisons.

To illustrate, assume we have five forms with the form-to-form equating constants shown in table 10. The values in the table are added to the *row form* to equate it to the *column form*. Equating A to B means adding -0.49 to form A logits; alternatively, equating B to A means adding 0.49 to form B logits. The first section of the table is completely filled so the equating constant for each form is the row mean.

	Completely Filled Table of Links						Equating Constant
	A	B	C	D	E	Sum	
A	0.0	-0.49	-1.10	-1.47	-1.94	-5.00	-1.00
B	0.49	0.0	-0.41	-0.88	-1.71	-2.50	-0.51
C	1.10	0.41	0.0	-0.42	-0.90	0.19	0.04
D	1.47	0.88	0.42	0.0	-0.64	2.13	0.43
E	1.94	1.71	0.90	0.64	0.0	5.18	1.04
	Missing Links						
A	0.0	-0.49				-0.49	-0.93
B	0.49	0.0	-0.41			0.08	-0.44
C		0.41	0.0	-0.42		-0.01	-0.03
D			0.42	0.0	-0.64	-0.22	0.39
E				0.64	0.0	0.64	1.03

Table 10: Resolution of Multiple Links

The second half of the table is missing six links: A-C, A-D, A-E, B-D, B-E and C-E. Finding the equating constants requires solving the matrix equation derived by expressions (48) and (49):

50. $$\begin{bmatrix} 2 & 0 & 1 & 1 & 1 \\ 0 & 3 & 0 & 1 & 1 \\ 1 & 0 & 3 & 0 & 1 \\ 1 & 1 & 0 & 3 & 0 \\ 1 & 1 & 1 & 0 & 2 \end{bmatrix} * \begin{bmatrix} t_A \\ t_B \\ t_C \\ t_D \\ t_E \end{bmatrix} = \begin{bmatrix} -.49 \\ 0.08 \\ -.01 \\ -.22 \\ 0.64 \end{bmatrix}$$

The two solutions give slightly different answers, but they are using rather different data. In either case, adding the linking constant to every logit for a form will shift it to the common origin, centered on the mean of all forms. The only hard part is knowing when to add and when to subtract.

Connectedness

Rasch's original problem with oral reading arose because the data could not answer the basic question, were differences in the reading scores because the student had changed or because the text was different? The data were not properly connected, or in the language of experimental design, the *occasion* effect was confounded with the *text* effect, or in the language of mathematics, the model was not adequately identified.

The specific objectivity of his new model allowed Rasch to construct the connections using entirely different samples. The design, shown in Table 11 (Rasch, 1960, p. 5), provided the data needed to estimate the differences between successive occasions and successive texts. This sort of design is now common practice with either overlapping subtests or overlapping samples or both.

	Text				
Grade	ORF	ORU	ORS	OR5	OR6
2	X				
3	X	X			
4			X	X	
5		X	X	X	
6			X	X	X
7				X	X

Table 11: Design of Remedial Reading Linking Study

With the facets model, the overlaps can be more difficult to identify and manage. The cleanest solution is to have every response scored by every judge but that is rarely feasible and never necessary. Table 12, adapted from Linacre & Wright (2004, p. 303), illustrates the problem and implies the solution. This example involves two blocks of four students, two judges, and two tasks.

	Tasks	X		Y	
	Judges	A	B	A	B
Students	101	*	*		
	102	*			
	103	*	*		
	104	*			
	201				*
	202			*	*
	203				*
	204			*	*

Table 12: Linacre-Wright Deficient Judging Plan

It is clearly possible to contrast Judge A with Judge B because there are several pairs of cells that involve the same student and the same task, but different judges. However, there is no comparison between Task *X* and Task *Y* that does not include the difference between student group *100* and student group *200*. The problem would be eliminated by administering Task *X* to some of the group *200* students or Task *Y* to some of the group *100* students. In principle, one additional observation could be enough but more overlap would give smaller standard errors and better balance would allow for more rational control of the model.

Scaling

After all the calibrating, linking and equating, nothing is left for *scaling* to mean but the linear transformation that converts logits into something more palatable (Smith, E., 2004):

51. $Scale\ Score = a + b*logit.$

Mathematically, logits are very convenient. However, they look a lot like standard normal deviates, involve positive and negative numbers, and require decimals. These are not particularly appealing for reporting. The scaling constants a and b are chosen to communicate as effectively as possible; there are no right or wrong choices but there are some useful guidelines:

- *Never report a negative scale score to anyone's parents*. They will react negatively.
- *Never send numbers with decimals to superintendents or school boards*. They will try to interpret the decimal part.
- *Never make a scale that looks like someone else's scale, especially* SAT, IQ, *or* percent correct. Someone will assume they are what they appear to be.

Beyond this, it's your choice.

Control of the Model[vi]

The first principle of Rasch analysis is specific objectivity: any appropriate sample of people will lead to statistically equivalent estimates of the item difficulties. If the results are not statistically equivalent, we have a problem or at least a concern or a limitation. Typically, this is put under the heading *Goodness of fit*; but that label was something of a foreign concept to Rasch; his phrase *control of the model*, however, was central. The amount of noise remaining in the data after removing the effect of the sufficient statistics was not of much concern, so long as it was noise. The intent is to achieve specific objectivity, not account for variance[vii].

From specific objectivity, it cannot matter whether we estimate the item parameters using the total group who took the test or any subgroup of the total. We will get statistically equivalent estimates from high scorers or low scorers, from Caucasians or African-Americans, from males or females, from this year's students or last year's, from fourth graders or fifth graders, from computer-administered or paper-and-pencil.

When we estimated the item difficulties, we solved the basic equation, the expected equals the observed, for d_i:

52. $$S_i = \sum_{r=R_{min}}^{R_{max}} n_r \frac{e^{b_r - d_i}}{1 + e^{b_r - d_i}} \ .$$

Then for any and all subgroups G, the equation:

53. $$S_{iG} = \sum_{v \epsilon G} \frac{e^{b_v - d_i}}{1 + e^{b_v - d_i}}$$

must also be solved[viii] within some level of statistical tolerance. The change in subscripts is intended to indicate the counts and sums include only the members of sub-group G. If equation (52) is not satisfied, specific objectivity does not extend to group G or some members of it.

This can be adapted to look like a mean square statistic (and replacing $\frac{e^{b_v - d_i}}{1 + e^{b_v - d_i}}$ with p_{vi}):

$$54. \qquad MS_{iG} = \frac{\left(S_{iG} - \sum_{v \epsilon G} p_{vi}\right)^2}{\sum_{v \epsilon G} p_{vi}(1 - p_{vi})} \ ,$$

This is very similar in appearance to the Wright-Panchapakesan iteration for estimating difficulties, expression (34), which suggests that the mean square is closely related to the amount the difficulty estimate needs to be changed to satisfy expression (52) for this group.

Between Score Group Mean Square

How one defines the groups G is important. If one suspects that the measurement may be threatened by left-handedness, attending a rural school, limited English proficiency, age, gender, or some other characteristics of the examinees, then those groups should be checked. Wright and Panchapakesan (1969) suggested a *between-score-group* mean square, which was easily implemented with limited computing capacity and is relevant to every instrument. Groups of examinees were defined on the basis of total score, providing a simple and direct check that the calibration was in fact independent of the ability distribution of the examinees.

$$55. \qquad MS_{iB} = \frac{1}{n_g} \sum_{G} \frac{\left(S_{iG} - \sum_{v \epsilon G} p_{vi}\right)^2}{\sum_{v \epsilon G} p_{vi}(1 - p_{vi})} \ .$$

In this expression, G now represents a range of adjacent raw scores and there is a total of n_g such groupings. Score groups were constructed to be as homogeneous in estimated ability (i.e., as few raw scores) as possible and to contain approximately equal numbers of examinees. In 1969, the number of groupings was severely limited by the capacity of the available computers. In today's world of computing, the score groups can be as narrow as the sample size allows.

The between group mean square was useful but did not do everything. It is sensitive to calibration differences related to ability (i.e., the shape of the item characteristic curve), which is fundamental to specific objectivity. It is not necessarily sensitive to the multitude of other characteristics that might distinguish examinees. It also was annoyingly affected by how many score groups were used and how they were defined.

Total Unweighted Mean Square (Outfit)

To avoid the arbitrary aspects of the between group mean square, a total mean square statistic was suggested (Mead, 1976) that took the group size to the absolute and completely objective minimum of one:

$$56. \qquad MS_{iT} = \frac{1}{N} \sum_{v=1,N} \frac{(x_{vi} - p_{vi})^2}{p_{vi}(1 - p_{vi})} \text{ , where } x_{vi} = (0,1) \text{ is the score of person } v \text{ on item } i.$$

The total mean square can be viewed as the extension of the between group mean square down to groups of size 1. However, it is no longer a simple test of the shape of the item characteristic curve. It will be affected any time a high scoring person misses an easy item or a low scorer passes a difficult item.

Basing this type of statistic on a dichotomous variable is problematic. The basic residual for person v on item i is:

57. $$y_{vi} = x_{vi} - p_{vi}.$$

If we *standardize* by dividing by the standard expression for the standard error and label it z_{vi}, the expression looks like a standard normal deviate:

58. $$z_{vi} = \frac{x_{vi} - p_{vi}}{\sqrt{p_{vi}(1-p_{vi})}},$$

Starting with a one or zero, the result is hardly a standard normal. Because only two things can happen, correct $(x_{vi} = 1)$ or incorrect $(x_{vi} = 0)$, the squared standardized residual will take one of two forms:

59. $$z^2_{vi} = \frac{1-p_{vi}}{p_{vi}} = \frac{e^{d_i}}{e^{b_v}}$$, if the response is correct, and $$z^2_{vi} = \frac{p_{vi}}{1-p_{vi}} = \frac{e^{b_v}}{e^{d_i}}$$ if incorrect.

In either case, z^2 is the odds against the observed response, which is remarkably handy and interpretable.

Total Weighted Mean Square (Infit)

The odds statistic can get very large in the best of circumstances. For example, with a rather small large-scale assessment, say 50 items and 20,000 examinees, one should not be surprised to see an event that has odds of one in a million.[ix] With the hope of dampening the level of alarm, Wright proposed a *weighted* mean square:

60. $$MS_{iW} = \frac{\sum_{v=1,N}(x_{vi} - p_{vi})^2}{\sum_{v=1,N} p_{vi}(1-p_{vi})} = \frac{\sum_{v=1,N} y^2_{vi}}{\sum_{v=1,N} w_{vi}} = \frac{\sum_{v=1,N} w_{vi} z^2_{vi}}{\sum_{v=1,N} w_{vi}}$$, where $w_{vi} = p_{vi}(1-p_{vi})$.

Wright has labeled the value from expression (59) *infit* and, to contrast with that, he has labeled expression (56), $MS_{iT} = \frac{1}{N}\sum_{v=1,N} z^2_{vi}$, *outfit*. The *outfit* and *infit* statistics are highly correlated with each other, with the between group mean square, and with the traditional point biserial correlation for the item. They are, however, not identical and many threats to specific objectivity are not readily detected by any of them.

Do-It-Yourself Mean Squares

Before we move on, the *between score group mean square* can also be written as an aggregation of the residuals:

$$61. \qquad MS_{iB} = \frac{1}{n_g}\sum_{G}\frac{\left(S_{iG}-\sum_{v\in G}p_{vi}\right)^2}{\sum_{v\in G}p_{vi}(1-p_{vi})} = \frac{1}{n_g}\sum_{G}\frac{\left(\sum_{v\in G}x_{vi}-\sum_{v\in G}p_{vi}\right)^2}{\sum_{v\in G}p_{vi}(1-p_{vi})} = \frac{1}{n_g}\sum_{G}\frac{\left(\sum_{v\in G}y_{vi}\right)^2}{\sum_{v\in G}w_{vi}} .$$

As before, each of the n_g instances of G represents a range of adjacent raw scores. The statistic will become large whenever the basic estimation equation $S_i = \sum_{r=R_{min}}^{R_{max}} n_r \frac{e^{b_r - d_i}}{1+e^{b_r-d_i}}$ is not satisfied within each restricted range of scores.

The more one knows about the construct to be measured, the theory behind it, and the examinees, the better able one is to design the appropriate analysis to control the measurement model in your neighborhood. We can reorganize the person-item residuals *ad infinitum*. If we are creative in the definition of the groups G, beyond just adjacent raw scores, expression (61) can be applied to almost anything.

Two obvious choices are *gender* and *ethnic group*. We can immediately extend this to define the groups as *gender by ethnic* or as *gender by ethnic by raw score*. This is letting the Rasch model do the grunt work of adjusting for differences in the ability distributions between groups in a DIF analysis. The appropriate group definition depends on what we suspect about, for example:

- Is there a gender difference? (Uniform gender DIF)
- Is it the same for all ethnic groups? (Gender by ethnic interaction)
- Is it the same for all ethnic groups at all levels of ability? (Non-uniform interaction)

There are just the obvious, rather generic choices.

Person Analysis Statistics

Given the symmetry of the Rasch model, we have been rather parochial in our discussion of model control. The entire preceding discourse could be framed in terms of a person just as well as an item by simply fiddling with the subscripts. The total unweighted mean square (*outfit*) for person v is:

$$62. \qquad MS_{vT} = \frac{1}{L}\sum_{i=1,L} z_{vi}^2 ,$$

And the *between item cluster* mean square is:

$$63. \qquad MS_{vC} = \frac{1}{n_c}\sum_{C}\frac{\left(\sum_{i\in C}y_{vi}\right)^2}{\sum_{i\in C}w} .$$

Item *clusters* have replaced person groups. Clusters can be defined anyway we find diagnostic, informative, or threatening. They could be defined, to list a few of the mundane, by content, passage, format, sequence, mode, type, exposure, Lexile, and, of course, difficulty.

Partitioning the Person-by-Item Residuals

The basic data behind all of this is a *person by item* matrix of residuals. It is the Rasch extension of Sato's *S-P chart* (Wu, 1998). We can operate on the columns to investigate items or on the rows to investigate people. We can also take it to a whole other level. The matrix can be partitioned into blocks that involve both rows and columns. We might ask, for example, if there are differences between boys and girls on items involving sports, cooking, horses, or auto mechanics.[x]

$$64. \qquad MS_{GC} = \frac{1}{n_{gc}} \sum_{CG} \frac{\left(\sum_{v \epsilon G} \sum_{i \epsilon C} y_{vi} \right)^2}{\sum_{v \epsilon G} \sum_{i \epsilon C} w_{vi}} .$$

Defining methods of aggregation is easy; interpreting the results, not so much. Understanding requires experts with real subject matter expertise, e.g., teachers, counselors, physicians, parents, who know more substantive things than do psychometricians.

Concluding Remark

The measurement model devised by Georg Rasch is an elegant statement of the conditions necessary for Thurstone's *fundamental measurement*. When the conditions are met, the results are *measures*, in exactly the same sense that the physical sciences have measures. Operationally for educational and psychological testing, the conditions boil down to items that are equally valid and reliable no matter whom we apply them to. This does not imply that the either the items or the people are homogeneous, (in fact, adequate control suggests that they should not be) but the interaction between the person and the item must ultimately be controlled by the one aspect of interest.

In the narrow sense, the measurement model is the technique for converting an observation into a measure. The arithmetic needed to accomplish this with Rasch is simple. In the broader sense, the measurement model is the process that begins with a vague concept of some aspect of a class of objects and ends with the valid, useful, general quantification of that aspect. Following Rasch's reasoning, this may also be simple but it's definitely not easy.

While there are efficient, more powerful alternatives for many of the methods discussed here, our basic point is that any interested party can follow the motivation and the principles involved when the maths don't get in the way. There is ample room and real value to investigating these and other methods, but for me, the urgency is to devote energy and resources to developing better methods to obtain good data rather than better methods of manipulating poor data. If we are not sure what we are measuring, does it matter how precisely we do it?

Returning to our very first point, validity trumps reliability.

Part III Notes

[i]A startling new discovery, like quartz scratching topaz, usually means that the data are miscoded.

[ii]In other contexts, the distinction is made that equated implies that the tests measure the same construct while *linked* implies the tests have been connected but they may not measure the same thing. Measures from equated tests are interchangeable; measures from *linked* tests may be useful as validations, as predictors or as mileposts but they are not interchangeable. In the context of Rasch measurement, we only consider tests that measure the same construct and so can be equated if they are linked. You can do what you want in the analysis phase.

[iii]It should also be noted that n_{Link} is the number of items we want in the link after the outliers are dropped.

[iv]This process is symmetrical. The *y*-intercept, -0.5, could be added to the bank logits to shift the plot horizontally and place them on the current scale if that made sense to anyone.

[v]For example, the standard error for item 1 is $1/\sqrt{18.32} = 0.234$, which is in the ballpark of $2.5/\sqrt{100}$. We are also acting as though the bank values are known without error.

[vi]To keep this exposition to a manageable length, we are restricting ourselves to statistics for the dichotomous model. See Ludlow (1983).

[vii]However, if the remaining noise doesn't look something like $\Sigma p(1-p)$, we would be a little concerned and Rasch would probably have thought specific objectivity had not be achieved.

[viii]Our rather cavalier assertion that because the estimation equations are solved within the total group, then, by dint of specific objectivity, they are solved within any relevant subgroup needs to be qualified somewhat. The equations were solved within the calibration sample using the biased estimates. Now that we have *de-biased* the estimates, the equations are not solved quite so well.

There seem to be four reactions to this news:

- The mean squares should be computed using *re-biased* estimates if we are to have any hope of understanding the null distributions.
- The unbiased estimates are the closest thing to truth we have and should be used for everything including fit analyses once we have them.
- The differences are too small to worry about given the problems building instruments that actually satisfy specific objectivity.
- We should be using fully conditional estimation methods and likelihood ratio tests anyway.

[ix]However, it still is curious that that student responded that way to that item. What was he thinking?

[x]These are included as illustrations of how the matrix might be partitioned; it is not a recommendation of how a test should be built. But if such divisions do exist in the data, you probably should check them.

References

Andersen, E. (1973) A goodness of fit test for the Rasch model. *Psychometika*, 38, 123-140.

Andersen, E. (1977) Sufficient statistics and latent trait models. *Psychometika*, 42, 69-81.

Andrich, D. (1973) *Latent trait psychometric theory in the measurement and evaluation of essay writing ability*, Doctoral dissertation, The University of Chicago.

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

Andrich, D. (1988) *Rasch Models for Measurement*. Newberry Park, CA: Sage Publications.

Andrich, D. (1995) Hyperbolic cosine latent trait models for unfolding direct responses and pairwise preferences. *Applied Psychological Measurement*, 19, 269-290.

Andrich, D. (2004) Understanding resistance to the data-model relationship in Rasch's paradigm: a reflection for the next generation. In E. Smith & R. Smith (Eds.) *Introduction to Rasch measurement. Theory, models and applications*. (pp. 25-47) Maple Grove, MN: JAM Press.

Briggs, J. & Wilson, M. (2004) An introduction to multidimensional measurement using Rasch models. In E. Smith & R. Smith (Eds.) *Introduction to Rasch measurement. Theory, models and applications*. (pp. 322-341) Maple Grove, MN: JAM Press.

Choppin, B. (1985) A fully conditional estimation procedure for Rasch Model parameters. *Evaluation in Education*, 9, 29-42.

Fischer, G. (1973). "The Linear Logistic Test Model as an Instrument in Educational Research." *Acta Psychologica*, 37, 359–374.

Fischer, G. (1976) Some probabilistic models for measuring change. In D. N. M. De Gruijter & L. J. T. Van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 97-110). New York: Wiley.

Fischer, G. (1995a) The linear logistic test model. In G. Fischer & I. Molenaar. (Eds.) *Rasch models – foundations, recent developments, and applications*. New York: Springer.

Fischer, G. (1995b) Linear logistic models for change. In G. Fischer & I. Molenaar. (Eds.) *Rasch models – foundations, recent developments, and applications*. New York: Springer.

Fischer, G. & Molenaar, I. (Eds.) (1995) *Rasch models – foundations, recent developments, and applications*. New York: Springer.

Fisher, R.A. (1947). *The design of experiments*. (4th) Edinburgh: Oliver and Boyd.

Fisher, W. (1992). Objectivity in measurement: a philosophical history of Rasch's separability theorem. In M. Wilson (Ed.) *Objective measurement: theory and practice. Vol. 1* (pp. 29-58). Norwood, NJ: Ablex.

Guttman, L. (1950). The basis for scalogram analysis. In Stouffer et al. *Measurement and Prediction*. The American Soldier Vol. IV. New York: Wiley.

Klein, H. (1975). *The world of measurements*. London: Allen and Unwin.

Linacre, J. (2004) Estimation methods for Rasch measures. In E. Smith & R. Smith (Eds.) *Introduction to Rasch measurement. Theory, models and applications*. (pp. 25-47) Maple Grove, MN: JAM Press.

Linacre, J. & Wright, B. (2004) Construction of measures from many-facet data. In E. Smith & R. Smith (Eds.) *Introduction to Rasch measurement. Theory, models and applications*. (pp. 296-321) Maple Grove, MN: JAM Press.

Ludlow, L. (1983) *The analysis of Rasch model residuals*. Doctoral dissertation. University of Chicago.

Mair, P. & Hatzinger, R., (2007) Extended Rasch modeling: the eRm package for the application of IRT models in R. Journal of Statistical Software. 20, 9. Retrieved July, 2007, from http://www.jstatsoft.org/v20/i09/v20i09.pdf

Masters, G. (1982) A Rasch model for partial credit scoring. *Psychometika*, 47, 149-174.

Mead, R. (1976) *Fit of data to the Rasch model though the analysis of residuals*. Doctoral dissertation. University of Chicago.

Rasch, G. (1960) *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedogogiske Institut. (reprinted 1980 with Foreword, Afterword, and References, Chicago: The University of Chicago Press)

Rasch, G. (1977) On specific objectivity: an attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy* 14: 58-94.

Scheiblechner, H. (1972). The learning and solving of complex reasoning items. *Zeitschrift fur Experimentelle und Angewandete Psychologie*, 3, 456-506.

Smith, E. (2004) Metric development and score reporting in Rasch measurement. In E. Smith & R. Smith (Eds.) *Introduction to Rasch measurement. Theory, models and applications*. (pp. 25-47) Maple Grove, MN: JAM Press.

Smith, E. & Smith, R. (2004) *Introduction to Rasch measurement. Theory, models and applications*. Maple Grove, MN: JAM Press.

Stone, M. (2004) Substantive scale construction. In E. Smith & R. Smith (Eds.) *Introduction to Rasch measurement. Theory, models and applications*. (pp.201-225) Maple Grove, MN: JAM Press.

Tavernor, R. (2007) *Smoot's ear: the measure of humanity*. New Haven: Yale University Press.

Thurstone, L. (1926) The scoring of individual performance. *Journal of Educational Psychology*, 17, 446-457.

Thurstone, L. (1928) Attitudes can be measured. *American Journal of Sociology* 33: 529-554.

Wright, B. (1968) Sample-free test calibration and person measurement. In *Proceedings of the 1967 invitational conference on testing problems*. (pp. 85-101) Princeton: ETS.

Wright, B. (1977) Solving measurement problems with the Rasch model. *Journal of Educational Measurement* 14, 2, pp. 97-116.

Wright, B. (1980) Foreword. In G. Rasch. *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

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

Wright, B. & Panchapakesan, N. (1969) A procedure of sample-free item analysis. *Educational and Psychological Measurement* 29, pp. 23-48.

Wright, B. & Stone, M. (1979) *Best test design*. Chicago: MESA Press.

Wu, Hsin-Yi. (1998) Software based on S-P chart analysis and its applications. *Proceedings of the National Science Council ROC(D)*. Vol. 8, No. 3, 1998. pp. 108-120. Downloaded September, 2007, from nr.stpi.org.tw/ejournal/ProceedingD/v8n3/108-120.pdf

Questions for Discussion

1. When should you consider using the Rasch model?
2. If "*validity trumps reliability*," do we need to worry about reliability at all?
3. Write the expressions for the three probabilities mentioned, but not derived, between equations (1) and (2), recalling that the joint probability of independent events is the product of the individual probabilities and that the sum of two Poissons is also a Poisson with the parameters summed and the counts summed.
4. What does it really mean to say the items are *equally valid*?
5. Aren't we being inconsistent when we let the data establish the scaling of the items but not the weighting?
6. Assuming the Sheriff of Nottingham has a logit value of 0.75 for bowmanship, what is the probability that he will hit a large, level, near target?
7. There is an abrupt change from probabilities to counts between equations (7) and (8). Show that summing over an arbitrary sample to obtain the counts does not interfere with the claim of specific objectivity.
8. How did we get from the probability in expression (18) to the log odds in expression (19)?
9. Table 4 contains the category probabilities for a rating scale or partial credit item with the step values of (-3, -1, 1, 3) used in Figure 2. Prepare a similar table for Figure 3 using the step values (-3, 0, -1, 3) at the logit $\beta - \delta = 0$.

k	τ_k	exp(0-τ_k)	Numerator	Probability
0			1	
1	-3			
2	0			
3	-1			
4	3			
Sum			98.49	1.00

10. What would happen to the probabilities in the table above if we made a different arbitrary choice for the numerator of category 0, say, five instead of one?
11. If the Rasch models are truly *sample-free*, why all the anxiety about connectedness?
12. Given the importance of unidimensionality, why might it be a good practice to use items that are *not homogeneous in their approach*, *content*, *or format*?
13. Since Rasch models are symmetric for items and people, why do we use data to estimate difficulties when we didn't need any to estimate abilities?
14. What did the phrase "completing the sum" mean and what sum did we complete to solve the equation (47) $At = T$ for the multiple links?
15. Why is no single fit statistic necessary? Sufficient?
16. If we always assume $\Sigma d_i = 0$, when would we ever need to recenter the item difficulties in equation (26)?

Discussion

1. The conventional IRT answer is that one should use the Rasch model when there aren't enough resources (i.e., sample size, time, expertise) to do better. A more generous view is that one uses the one-parameter model if the two- and three-parameter models have been tested against it and have not been found to *explain significantly* more variance.

   The Rasch answer is that one uses a Rasch model whenever one wants to do measurement. This implies that one has sufficient interest, expertise, and resources to define the variable and develop the instruments appropriately. These activities are not finished until the data conform to the model's requirements. The question of model choice is a philosophical not empirical decision.

2. Of course we do!

3. Joint probability of the counts a_{vi} and a_{vj} is the product of the individual Poisson functions:

$$p(a_{vi}, a_{vj}) = e^{-\beta_v \varepsilon_i} e^{-\beta_v \varepsilon_j} \frac{\beta_v^{a_{vi}} \varepsilon_i^{a_{vi}} \beta_v^{a_{vj}} \varepsilon_j^{a_{vj}}}{a_{vi}! a_{vi}!} .$$

   Probability of the sum of counts is a Poisson with $\lambda = \beta_v \varepsilon_i + \beta_v \varepsilon_j$ and $x = a_{vi} + a_{vj}$:

$$p(a_{vi} + a_{vj}) = e^{-(\beta_v \varepsilon_i + \beta_v \varepsilon_j)} \frac{\beta_v^{a_{vi} + a_{vj}} (\varepsilon_i + \varepsilon_j)^{a_{vi} + a_{vj}}}{(a_{vi} + a_{vj})!} .$$

   Conditional probability of the counts given the sum is the ratio of the two:

$$p(a_{vi}, a_{vj} \mid a_{vi} + a_{vj}) = \frac{p(a_{vi}, a_{vj})}{p(a_{vi} + a_{vj})} = \binom{a_{vi} + a_{vj}}{a_{vi}, a_{vj}} \frac{\varepsilon_i^{a_{vi}} \varepsilon_j^{a_{vj}}}{(\varepsilon_i + \varepsilon_j)^{a_{vi} + a_{vj}}} .$$

4. *Equally valid* means no item is any better instance of the aspect we are trying to measure than any other item. The items are truly interchangeable and we will not rely on the data to tell us how they should be weighted. Empirical weights will generally lower the estimated standard errors because they will account for more of the variance, perhaps idiosyncratic, in the observed data. The cost is this increased reliability is reduced validity. The definition of the construct has changed because the empirical weights are different than the developers of the instrument envisioned. Rasch analysis is willing to risk slightly higher standard errors in order to preserve the purest definition of the construct. IRT textbooks refer to this property as "equal item discriminations"; just another set of item parameters to be estimated.

5. Scaling relates directly to the one aspect of the agent and the object that we are trying to measure and which are parameterized in the model. Rasch models have sufficient statistics for managing the estimation of these parameters, allowing scaling of the agents without referencing the objects and vice versa. There are no sufficient statistics for the weights, and, hence, no such thing as *sample-freed* estimates of them.

6. With a logit for bowmanship of 0.75, the probability of hitting a level, large, near target (logit difficulty of 0.0) is:

$$pr(hit \mid \beta = 0.75,\ \delta = 0.0)\ \frac{e^{0.75}}{1+e^{0.75}} = 0.68$$

				Page	Sheriff	Knight	R Hood	W Tell
			Logit	0.25	0.75	1.75	3.25	3.50
Downhill	Small	Far	3.5	0.04	0.06	0.15	0.44	0.50
Level			3	0.06	0.10	0.22	0.56	0.62
Downhill	Large		2.5	0.10	0.15	0.32	0.68	0.73
Level			2	0.15	0.22	0.44	0.78	0.82
Downhill	Small	Near	1.5	0.22	0.32	0.56	0.85	0.88
Level			1	0.32	0.44	0.68	0.90	0.92
Downhill	Large		0.5	0.44	0.56	0.78	0.94	0.95
Level			0	0.56	0.68	0.85	0.96	0.97

7. Our expected count is the probabilities summed over whatever sample we have. The probabilities and the counts are both sample dependent. Specific objectivity says the item parameter estimators are not. Because the person and item parameters can be separated:

$$n_{10} = \sum_{v=1}^{N} p_{v10} = \sum_{v=1}^{N} \frac{B_v \Delta_2}{(B_v+\Delta_1)(B_v+\Delta_2)} = \Delta_2 \sum_{v=1}^{N} \frac{B_v}{(B_v+\Delta_1)(B_v+\Delta_2)}\ .$$

Similarly for n_{01}, so

$$n_{10}+n_{01} = (\Delta_1+\Delta_2)\sum_{v=1}^{N} \frac{B_v}{(B_v+\Delta_1)(B_v+\Delta_2)} \quad \text{and} \quad \frac{n_{10}}{n_{10}+n_{01}} = \frac{\Delta_2}{\Delta_1+\Delta_2}\ .$$

8. If the probability of pass is $p = e^x/(1+e^x)$, the probability of fail is $1-p = 1\ /\ (1+e^x)$. The odds of pass to fail is $p\ /\ (1-p)$ = e^x and the natural logarithm is x.

9. Because this table is evaluated at logit 0, which is the step parameter for category 2, the probabilities for categories 1 and 2 are equal (and, because of the disorder, category 3 is much more likely than either 1 or 2 for a person at this location).

k	τ_k	exp(0-τ_k)	Numerator	Probability
0			1	0.01
1	-3	20.09	20.09	0.20
2	0	1.00	20.09	0.20
3	-1	2.72	54.60	0.55
4	3	0.05	2.72	0.03

10. Nothing. The five, or anything else, would appear as a factor in each of the numerators and their sum. Because it occurs once in each of them, it cancels out when the division is done for the probabilities.

11. Wright's term *sample-free* may be misleading. The model is not magic. All parameters must still be identified. In order to compare any two components, there must be a way to arrange the data that includes a simple contrast between the two components. *Specific objectivity* means that the result of the comparison does not depend on what sample was used to form this contrast; it doesn't mean you can do anything without data.

12. Rasch's stipulation that "*the relationship should be found in several sets of data which differ materially in some relevant respects*" (Rasch, 1960, p. 9) applies to items as well as people. Using items of different types broadens the definition of the construct and strengthens the validity arguments. If we use stationary bull's-eyes exclusively to gather evidence about bowmanship, we can never know whether we are measuring bowmanship or proficiency at hitting stationary bull's-eyes. This may be good for winning competitions; not so good for winning wars or bringing home the king's venison.

13. We could treat ability and difficult in exactly parallel manners but that is less efficient. Because the number of people is usually much larger than the number of items, using the observed item scores and the frequencies saves time and effort when estimating difficulties. To estimate abilities, we needed to know the difficulty of every item on the test. Then we estimated the ability for every possible score on the test without worrying about whether or not anyone got that score. We could do the same thing for items. If we know the ability of everyone who took the test, we could estimate the difficulty for every possible item score from one to the number of people minus one with no additional information. It is also possible to use the analogous procedure presented to estimate difficulties to estimate abilities, but we frequently want the ability estimates for all scores rather they have happened yet or not.

14. In order to solve the equations, we need to impose one constraint and we would like that to be $\Sigma t_j = 0$. If there is a link t_{ij} connecting every pair of forms, then the row sum is:

$$T_i = \sum_{j=1}^{k} t_{ij} = \sum_{j=1}^{k} (t_i - t_j) = kt_i - \sum_{j=1}^{k} t_j \cdot$$

Σt_j can be set to zero and the solution is easy: $t_i = T_i/k$. If, however, the link between, say, forms 1 and 2 is missing, then the row sums are incomplete because row 1 is missing $(t_1 - t_2)$ and row 2 is missing $(t_2 - t_1)$.

$$T_i^* = \sum_{ij \neq 1,2} t_{ij} = kt_1 - \Sigma t_j - (t_1 - t_2) \cdot$$

Writing it in this form, we have kept the complete sum of the t_i that we want to set to zero but have a little left over. Simplifying a little, we have:

$$T_1^* = (k-1)t_1 + t_2 \cdot$$

which says we reduce the coefficient for the diagonal term by one and add one to the off-diagonal term. The equation for form 2 would take the same form. The equations for forms 1 and 2 now need to be solved simultaneously but how hard is that?

15. No single fit statistic is necessary because there are multiple, equally defensible ways one can approach the analysis and arrange the data. No single fit statistic is sufficient because there is an unimaginable number of ways that the data can depart from the model. No single statistic can be powerful against all alternative hypotheses.

16. Number one, we don't always assume that. Sometimes the assumption is that a subset, perhaps just one, of the items have some average difficulty. Second, after the items have been linked to another form or to a bank, the center of the items will have been shifted away from zero even if zero was the initial assumption. Third, we may want to tailor the testing to match the expected location of individual students. If students take different forms, the forms will, in general, have different centers.

Data Recognition Corporation
13490 Bass Lake Rd, Maple Grove, MN
(763) 268-2000
www.datarecognitioncorp.com