Popular Measurement

Journal of the Institute for Objective Measurement

Volume 4
2002

A Message from the Publisher

It is with great pleasure that the Institute for Objective Measurement (IOM) provides you with its first online issue of *Popular Measurement*. This periodical is targeted to the community at large. Its articles translate scientific knowledge into serviceable public information that can be used to improve civilization.

The IOM would like to thank its new editor, Dr. Matthew Enos, for his diligent efforts in bringing the new project into reality. We hope that you enjoy the publication.

Valerie Been Lober, Ph.D.
Executive Manager
Institute for Objective Measurement

From the Editor

Thanks, Donna

When Ben Wright's colleague and former student Donna Surges Tatum suggested a new magazine, one that would be aimed at a broader audience than the typical academic journal, Ben Wright and the IOM Board eagerly endorsed her project. The idea of a magazine that would reach deep into the measurement community of scholars, practitioners, and students had a special appeal for Ben Wright, who was enlarging the scope and applicability of his own writing and teaching. We remember his excitement in class when he introduced new students to *Popular Measurement*.

Donna plunged into the challenge of producing a journal unlike any other. She brought to the task her own unique background in effective communication. She knows the printing industry, is an accomplished college speech teacher, directs research and psychological services for the American Society for Clinical Pathology, and wrote her doctoral thesis on the complexities of judging speech effectiveness.

The new magazine both delighted and enlightened its readers. Donna insisted on articles of immediacy and importance, as well as production values of attractiveness and striking design quality. The result was a journal that was both a visual feast and an intellectual delight. Articles ranged from provocative theoretical observations to practical applications of measurement and charming profiles of leading lights in the Rasch movement. The temptation was to read each new issue right through without putting it down.

Donna's first issue convinced us of the importance of her original inspiration, and by now *Popular Measurement* has established its claim on our interest and loyalty. With the new online format, the Institute for Objective Measurement hopes to broaden the readership of *Popular Measurement*. In this endeavor, the new editor can only aspire to approach the standards Donna established. As we move into a new phase of *Popular Measurement*, the inspiration and support of Donna Tatum will be a great asset.

Going Online

The present editor grew up in a newspaper family and loves writing, editing, and publishing. In our youth, we set and redistributed pieces of cold type by hand and heated and poured molten lead into asbestos forms for illustrations and pictures. To us, typefaces, leading, and hot type were tangible entities, not figures of speech.

But what are we to make of this new world of online publishing, where everything is electronic and so little seems solid? It will be a challenge and an adventure. It may take some getting used to, but we are determined to make it succeed. The IOM Board of Directors is proud to be moving into the future of journal publication. In this adventure, your assistance (and sometimes patience) will be appreciated.

Where Credit is Due

This first online issue of *Popular Measurement* was necessarily a collaborative effort. The articles had already been gathered by Donna Tatum. Members of the IOM Board provided counsel and guidance. Valerie Lober and Mary Lunz organized the peer review process. Michael Linacre shared his own extensive publishing experience and technical expertise at every turn.

Future Issues

Our tentative plans call for two issues of *Popular Measurement* a year, perhaps Spring and Fall. Since most contributors and readers are tied to the academic year, we think this schedule may be most useful. Online publishing lends itself well to a shorter journal appearing more often.

In addition to the typical research and theoretical articles, we would like to include reflections that help illuminate the Rasch world. One such feature might be “First Encounters with Rasch,” in which both veterans and newcomers would share their initial impressions of and experiences with Rasch measurement. Another might be “Twenty-Five Words or Less…,” in which writers would attempt to capture the essence of the Rasch model, as they understand it, in a page or so. The lively interchange of ideas on the ACER Rasch website suggests a wealth of creative ideas and expression we hope to tap. See “From the Classroom” (page 34) for another feature we hope to continue and expand. We welcome your suggestions for other new features.

Of course the key ingredient to the continuation of *Popular Measurement* as a vital journal is the desire of measurement professionals and students to share their ideas, research, and theories with others. Please consider writing for *Popular Measurement*.

Feedback and New Articles

Please send us your comments, questions, and other feedback. We are especially eager to receive prospective articles for publication in Volume 5, Number 1, scheduled for Spring 2003. Send feedback and new articles to Matthew Enos, Editor, *Popular Measurement* at:

~~InstObjMeas@worldnet.att.net~~

or

www.rasch.org

Institute for Objective Measurement (IOM)

155 North Harbor Drive #1002
Chicago, IL 60601

Contents

Residuals: Trash or Treasure? 1
Larry H. Ludlow

Identifying Problematic Examiners in an Oral Examination 8
Jessica Heineman-Pieper and Mary E. Lunz

Food for Thought from Carnot 13
William P. Fisher

Quality Control in Testing 15
Mark H. Stone

Rule-based Aptitude Measurement: Artistic Judgment 24
Nikolaus Bezruczko and Ambra Borgognoni Vimercati

An Introductory Text on Rasch Measurement 31
Trevor G. Bond

From the Classroom 34
Matthew Enos

2002 Membership of the Institute for Objective Measurement 37
Valerie Been Lober

Residuals: Trash or Treasure?

Larry H. Ludlow, Ph.D.

Boston College

Introduction

One of the significant challenges to the successful teaching of statistics is that of giving the topic some relevant historical perspective. Such an effort is worthwhile because many students need a context within which they can understand why we are studying a given technique and how that technique came about. In fact, many of our traditional statistical procedures can be associated with specific interesting individuals or time periods. For example, the history of probability estimation (as a study of gambling behavior) and the development of correlation and regression (Galton and his study of individual differences) are familiar interesting examples. We can also paint a broader picture when we consider the development of schools of philosophy and their attendant methods of observation and analysis (a good overview is provided by Butterfield, 1957). The purpose of the present paper is to illustrate that progress in understanding the world around us can be grasped in terms of efforts to explain unexpected events under some existing theory and mathematical model — unexpected events that ultimately led to re-formulations of both theories and models. In particular, this paper will argue that current psychometric models of item response data not only have an interesting history in their own right but may also be considered the result of progressive efforts focused on explaining unexpected observations, i.e., item responses.

Unexpected observations are noticeable because they are discrepancies in behaviors expected under existing models. Technically, the difference in what is predicted by a model and what is actually observed is termed a residual. While some researchers seem to dismiss the residual as a nuisance and a distraction (i.e., garbage to be ignored), other researchers consider the residual a key to progress in the development of theories and models (i.e., a gold nugget). Evolution in scientific model building, from this perspective, may be characterized as an effort to reduce residual variation.

Science and Error

The Scientific Revolution, for example, brought about an appreciation that unaccounted phenomena could be important, occasionally more important than the original research. An unexpected result, experimental discrepancy, residue, or "residual phenomena" came to be defined as whatever remains outstanding and unaccounted for after "subducting the effect of all known causes, as well as the nature of the case permits, either by deductive reasoning or by appeal to experience" (Herschel, 1851, art. 158). Herschel underlines the role played by residuals: "Almost all the greatest discoveries in astronomy have resulted from the consideration of what we have elsewhere termed residual phenomena…" (Herschel, 1871, art. 856). In effect, the "expected unexpected" became sought after (Brannigan, 1981, p. 159).

A hallmark of the Scientific Revolution was the emphasis placed on testing one's theories experimentally. As the practice of experimentation became commonplace, the physical sciences began to make rapid progress in the formulation of laws of nature. Myths and Aristotelian logic are examples of philosophical paradigms that were no longer taken as a substitute for experimentation. Early experiments, however, led to qualitative theories, e.g., the phlogiston theory of combustion (Roberts, 1989). Qualitative theories left qualitative residuals. Consequently, there was no way to assess quantitatively the surprise or importance associated with discrepancies from expected outcomes.

It was eventually recognized that qualitative statements allowed too much leeway in a stated relationship. As experimental techniques improved, qualitative laws were either transformed into quantitative laws (e.g., Boyle's Law replaced the "spring of the air") or discarded as wrong (phlogiston theory replaced by oxygen theory). The scientific attitude shifted: if the purpose of a law was to explain and predict observations, then the law had to be quantified. "Indeed, it is a character of all the higher laws of Nature to assume the form of precise quantitative statement" (Herschel, 1851, art. 116).

To arrive at quantified laws meeting Herschel's description requires precise and accurate observations. Kepler, for example, was familiar with, but could not use, the inaccurate data compiled by Ptolemy — his law of elliptic orbits required the accurate measures of Brahe. In fact, Ptolemy's data confounded the efforts of astronomers for generations because there were so many unusual values that increasingly complex equations had to be specified in order to account for them (a situation not unlike multi-parameter item response theory models). Francis Bacon was one of the early scientists to stress the importance of accurate measurement: "Truth emerges more readily from error than from confusion" (Bacon in Kuhn, 1970, p. 18). Lord Kelvin, too, attributed the major discoveries of science to "accurate measurement and long-continued labors in the minute sifting of numerical results" (Kelvin in Conant, 1954, p. 121).

As measuring became more precise, discrepancies between observed and expected phenomena attracted more attention. The inter-dependency of experimentation and quantification led to awareness that discrepancies from predictions, while expected to be of some magnitude, ought to be reasonably small and not systematic. A means of determining what might constitute "reasonable agreement" between data and theory became possible through replicated experimentation. Replications provide statistical estimates of standard errors and provide practical rules for defining "good fit," "outliers," and "random error." Replications establish precision that, if the precision is acceptable for the task, can expose inaccuracy because residuals can be used to expose lack of congruence between theory and data. The magnitude and direction of the residual then provides information useful for supporting, rejecting, or modifying the theory.

This quantitative test of fit of observations to an explanatory model is a primary distinction between myth and science. There are no residuals from a myth. The myth is simply revised to account for new phenomenon. Residuals appear *only* against the background of a scientific paradigm. "It is the presence of the research as a specifically motivated course of action that makes the event accidental or fortuitous in the first place" (Brannigan, 1981, p. 73). Paradigms change, but if a theory can lead to an assertion that a specific condition should produce an expected outcome, then residuals will occur.

The Problem of Residuals

Since residuals always occur in empirical research, the problem becomes "what to do with them?" Every researcher notices and then passes by residual variation. There is neither enough time nor effort available to analyze every residual. Even "large" residuals usually disappear under scrutiny, i.e., observational, instrumental, recording, and computational mistakes routinely occur. When such mistakes are corrected, the residual usually becomes negligible. But significant residuals, relative to some measure of "reasonable agreement," do occur and, more important, recur. The cause of a residual may be trivial but it may also lead to an important discovery.

Unfortunately, it is seldom obvious whether a residual is the key to an old puzzle or the clue to a new direction of inquiry. But, as Pasteur said, "In the fields of observation, chance favors only the prepared mind" (Pasteur in Kuhn, 1961, p. 49). Recognition occurs only when the scientist "*knowing with precision* what he should expect, is able to recognize that something has gone wrong" (Kuhn, 1962, p. 65). The crucial step is taken when the observation, "Something has gone wrong…," is followed up by analysis.

Individual skill and insight and relevant instruments and concepts are the conditions necessary for recognizing consequential unexpected results. Examples abound in the history of science where one researcher has encountered a persistent residual phenomenon, but not pursued it, and another has come along and "discovered" it. Cavendish, for example, saw and measured a residual gas when nitrogen was removed from air. But it was Ramsay and Raleigh a century later who "discovered" argon.

When a significant residual persists it usually passes from the category of "novelty" to "anomaly." An anomaly is defined as "A recognition that nature has somehow violated the paradigm-induced expectations that govern normal science" (Kuhn, 1962, p. 52). A persistent peculiarity may be considered a novelty and ignored, an anomaly to be pursued and explained, or an anomaly to be noticed but left for future generations to explain. Many of the laws of science are the result of explanations of violations of expectation which could no longer be ignored (Ashall, 1994; Kantorovich, 1993).

Research may be undertaken to delimit the boundaries of a persistent residual anomaly. What are the circumstances under which the residual happened? Can it be reproduced? Will it always happen if the circumstances recur? Sometimes, as in chemistry, more refined experimental techniques eliminate the discrepancy. Some persistent residuals have been left as known anomalies. Newton's theoretical value for the speed of sound, for example, was a scientific anomaly for 50 years until refined measurements attained by Delaroche and Berard were used by Laplace to reduce the discrepancy from 20 per cent to 2.5 per cent (Kuhn, 1977, p. 196). Other residuals have been resolved by the discovery of a new phenomenon. The discovery of Neptune accounted for the inexplicable orbit of Uranus. The prediction of Neptune is a beautiful example of a new theory (Newtonian physics) that was more explanatory than it's predecessor (Kepler's laws).

Models and Expectations

Whatever the ultimate explanation, residuals can only be understood relative to the mathematical and theoretical model from which they result. Models provide the potential for expecting certain conditions to occur; they provide a background against which surprises stand out. But once a significant residual is identified and replicated — where does one search for the cause? Does one question the theory, data, instruments, or computations? It is not always obvious where adjustments should be made. The residuals from the Ptolemaic mathematical model led Tycho Brahe to obtain more accurate measurements of the universe but he still retained Ptolemy's mathematical model (based on circular motion) and geocentric model for the relation of the heavenly bodies to earth. The same residuals led Copernicus to reject the geocentric model in favor of the heliocentric model but retain Ptolemy's mathematics. Kepler, in turn, retained the heliocentric model but rejected circular orbits in favor of elliptical ones.

In education the phenomena of interest are variables such as arithmetic ability, academic motivation, inductive reasoning skill, etc. A variable is not measured by a single question (item) on a test. A variable is measured by a set (sample) of questions written to cover (replicate) a single topic. Test items are selected to cover a range from a lesser to a greater degree of the intended variable, e.g., knowledge, ability, motivation. When a set of content-homogeneous items is used to define a variable operationally it is called a "scale." Administering the scale to students results in scale value estimates for students ("measures") and scale value estimates for items ("calibrations"). The comparison of person measures with item calibrations leads to estimates of performance expected when a person takes any item. The difference between the observed and expected performance is then a residual.

In educational measurement, replication is accomplished through the administration of a homogeneous set of items — a measurement instrument. This is not only because a set of homogeneous observations estimates more precisely the performance level of a student than does a single item, but also because the intended replications provide evidence for consistency when it is obtained and they expose inconsistency, the unexpected residual, when that occurs. A response pattern inconsistent with a modeled expectation of performance leads to doubts about the relevance of the measurement. But since observed response patterns are never perfectly consistent with their expectations, some subjectivity is unavoidable in the determination of how inconsistent a pattern must be before the measurement should be judged inaccurate.

Many questions arise as a test of fit is carried out. How well have the expectations of the model been met, i.e., were all statistical assumptions met? How accurate is the instrument, e.g., how many items were there and how were they scored? Does the instrument yield useful information about the people, e.g., was the instrument measuring a construct appropriate for the people? Did the instrument work for the task at hand, e.g., to what extent did it measure a wide range of variability in the construct? Have relevant extraneous variables been controlled, e.g., were testing conditions standardized and were there individuals or groups of people with extraordinary characteristics? Each question is a check that the observed relation between model and data is of the form where the simple difference between the observed response given by a person and the predicted response under the statistical model is a residual. If the data have been collected carefully and the model is useful for explaining them, then observed and modeled (or predicted) values should be similar and the residual should vary in magnitude and pattern like a random variable.

How Important is the Residual?

The problem is to determine whether a residual is negligible and occurred as expected or whether the residual suggests something more like:

Observed value - Modeled value = Random error + Significant outlier + Systematic relation.

In an educational application a significant outlier may result from a student with a relatively low estimate of ability who nevertheless succeeds surprisingly on one or more difficult items. Similarly, a systematic relation might be traced to a classroom of relatively capable students who surprisingly miss a set of linked items on a scale. In these examples the important concern is whether or not the measurement process has yielded quantitative estimates of student ability that are useful. When residuals exhibit variation uncharacteristic of that expected for random error, the measures (if they can still be called "measures") are determined by influences not expected or accounted for in the model. And the question now becomes: "Do we maintain the measurement model as it is or do we alter the model to account for variation peculiar to the specific testing circumstances?"

My answer to this question relies on a distinction between measurement theory and statistical theory. Measurement theory, in my opinion, started with the early efforts of the German psychophysicists Weber and Fechner to formulate a quantitative relationship between mind and matter. Their efforts succeeded in establishing that physical stimuli and subsequent reactions can be formulated as quantitative laws. From Galton and Pearson we came to understand that statistical techniques can be applied to human behavior to reveal individual differences and that those individual differences lead to statistical discrepancies that can be modeled as random error. From Binet we began to quantitatively measure and statistically test for cognitive differences. From Thurstone we discovered we can quantitatively measure and statistically test for affective differences. From the deterministic models of Guttman we saw that variables may be constructed as linear hierarchies within which performance at one level of functioning presumes successful performance at lower levels of functioning. And, finally, through the probabilistic models of Rasch we understand that we can model the expected response a person gives to an item as the simple difference between the level of ability possessed by the person and the level of ability required by the item to be successfully responded to. A residual worth noting and thinking about, in the Rasch model, is either an unexpected success or unexpected failure on a given item.

Statistical theory, on the other hand, prospered through Galton and Pearson who established that human variability in one domain (variable) may be understood as a function of variability in one or more other variables. From this simple concept of co-relation came regression, or the opportunity to incorporate as many variables as one chose to use to explain behavior. At the root of early and present statistical theory, however, is the recognition that as one accounts for more and more variation in some outcome variable, one is also reducing the unaccounted-for-variation (error). Thus as each additional variable is added to the statistical model, residual variation is reduced. In general, this situation is desirable. The problem is that in any given analysis, particularly an exploratory one where the effort truly is to maximize the proportion of predictable variance (which is the same as minimizing error variation), the risk of "capitalizing on chance" greatly increases until one could, theoretically, reduce residual variation to nil but have a model with no likely generalizability. Consequently, statistical theory took a positive step forward with the introduction of structural equation models.

In these models the investigator specifies the variables to be employed based on theoretical reasoning, the relationships between the variables (and error components) are specified, and the data are tested to see the extent to which they fit the proposed structure. Residuals from this model are then useful for pointing out areas of the theory to be strengthened, modified, or discarded.

Trash or Treasure?

To a great extent, then, Rasch models are structural equation models to be confirmed through the data. If the data do not fit the model, then the data irregularities may be investigated through the residual patterns. Here, then, we have a measurement model (Rasch) where the meaning of a residual remains constant across all similar circumstances. Two and three parameter logistic (2/3

PL) item response theory models, in contrast, are exploratory models to be fit and modified as the local testing conditions suggest. Residuals, in those models, represent noise to be eliminated. That is, unlike the one parameter Rasch model, two and three parameter item response theory models estimate item parameters that subsume and mask variation that would otherwise be noticed and interpreted as problematic and undesirable for measurement.

The analysis of measurement-model residuals is important for many audiences. The classroom teacher, tester and evaluator need to know how useful their measurements are. The progress of a student or class and the effectiveness of a curriculum is typically measured by some form of performance level score. This score is usually assumed to measure the variable of interest and not something else irrelevant to the task. If the scale is not working as intended, then the measures may be meaningless. Fortunately, analyses of residuals can make the use of meaningless measures avoidable.

The following examples, from my own work, illustrate the types of problems that were revealed through the analysis of residual variation. I have found: (1) *start-up effects,* attributed to subjects being initially confused about the responses expected of them; (2) *speed effects,* where limited time led to identifiable guessing behavior; (3) *interviewer effects,* where interviewers had different impressions about how to score interviewee responses; (4) *instrument effects,* where scoring sheets were miskeyed, items had multiple correct answers, or items had no correct answers; (5) *dimensionality effects,* where items clearly addressed two or more constructs (dimensions); (6) *teacher effects,* where teachers used their own unique interpretation of how to administer a performance assessment to their students; (7) *classroom effects,* where students were not taught a specific curricular component taught on a statewide assessment; and (8) *special characteristic effects,* where students with unique cognitive (low verbal ability), affective (highly anxious), or physical characteristics (wheelchair user) did much worse or better than was expected of others with their same estimated ability level. In each case, the problem was identified and a change was made in the instrument or the testing conditions — the model itself remained the same.

In conclusion, then, I encourage everyone who engages in the development or application of measurement instruments to aggressively dig through the residual variation. What you find may be more important than what you set out to investigate in the first place.

References

Ashall, F. (1994). *Remarkable discoveries!* New York: Cambridge University Press.

Brannigan, A. (1981). *The social basis of scientific discoveries*. Cambridge: Cambridge University Press.

Butterfield, H. (1957). *The origins of modern science 1300-1800*. New York; The Free Press.

Conant, J. B. (1954). *Science and common sense.* New Haven: Yale University Press.

Herschel, Sir J. F. W., & Bart, K. H. (1851). *Preliminary discourse on the study of natural philosophy* (new ed.). London: Longman, Brown, Green & Longmans, Paternoster Row.

Herschel, Sir J. F. W., & Bart, K. H. (1871). *Outlines of astronomy* (11th ed.). London: Longmans, Green, and Co.

Kantorovich, A. (1993). *Scientific discovery: Logic and tinkering*. Albany, NY: State University of New York Press.

Kuhn, T. S. (1962). Historical structures of scientific discoveries. *Science, 136*, 760-764.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: The University of Chicago Press.

Kuhn, T. S. (1977). *The essential tension*. Chicago: The University of Chicago Press.

Roberts, R. M. (1989). *Serendipity: Accidental discoveries in science*. New York: Wiley.

Notes

1. Key words: residual, scientific models, theory building, psychometrics.

2. Correspondence should be addressed to Larry H. Ludlow, Boston College, Lynch School of Education, 140 Commonwealth Avenue, Chestnut Hill, MA, 02467.

Identifying Problematic Examiners in an Oral Examination

Jessica Heineman-Pieper, M.A. and Mary E. Lunz, Ph.D.

Measurement Research Associates, Inc.

Introduction

Oral examinations are most valid and reliable when the analyses identify and account for stable differences in how examiners assess and rate examinees (Heineman-Pieper & Lunz, 2000). This can be achieved using the multi-facet model. When oral examinations are analyzed using the multi-facet model, the multi-facet analysis calculates and adjusts for each examiner's unique level of 'severity' in rating examinees. As a result, examiners need no longer strive for the impossible goal of perfect agreement on candidate raw scores. Instead, examiners need only agree on the basic meaning of the rating scale categories and implement that understanding consistently within their own grading behavior. The multi-facet model provides fit statistics to indicate how well an examiner's rating behavior fits the expectations of the model. These fit statistics are frequently used to indicate how well an examiner understands and implements the rating scale categories. These statistics will also help determine which examiners will be allowed to participate in subsequent examinations.

Examiner fit statistics can signal when examiners are inconsistent, or when they fail to distinguish among candidates of different abilities. However, only the latter examiner behavior, *non-discrimination*, can be identified unambiguously from model outputs. In contrast, examiner *inconsistency* cannot be definitively diagnosed without careful review of specific raw data that may or may not have been collected. This paper demonstrates: (1) how non-discrimination can accurately be diagnosed using facets outputs; (2) why examiner inconsistency cannot be accurately diagnosed exclusively from standard model outputs, and (3) what additional information can accurately identify the quality of apparently inconsistent examiners.

Data

These data are from an oral certification examination. Candidates were rated on several skills across several practice areas by two independent but co-present examiners. Accordingly, the data were structured to allow a four facet analysis with the following facets: candidate, examiner, skill, and practice area.

Methods

The data were analyzed using the multi-facet model (FACETS), which adjusts the raw ratings that candidates earn to account for variations in examiner severity, skill difficulty, and difficulty of the practice area in which the ratings were earned. The multi-facet model generates measures of candidate ability, examiner severity, etc., in logits. Logit scores reflect the log odds that a candidate will receive a rating of a particular value, when rated by an examiner of a particular severity, in a practice area of a particular difficulty, on a skill of a particular difficulty.

In addition to constructing measures, the multi-facet model generates quality control statistics to indicate where the data do not provide a good fit to the model. The most important of these are the mean squared fit statistics (infit and outfit). The mean squared fit statistics construct a ratio that compares observed to expected responses. When an examiner's observed and expected responses perfectly coincide, the fit statistics for that examiner will be 1.0. When the examiner's observed responses are inconsistent with the expectations of the model, the mean squared fit statistics will be "high." Numerically, high has been found to correspond to approximately 1.5 or greater. When the variation in observed responses is less than would be expected by the model, the mean squared fit statistics will be low (generally, approximately 0.5 or lower).

When an examiner's fit statistics are too high or too low, the examiner is usually singled out as a potentially incompetent examiner. In the case of low fit statistics, this conclusion can indeed be verified using additional model outputs (specifically, the pattern of ratings). However, in the case of high fit statistics, it is far more difficult (and sometimes impossible) to distinguish a capricious examiner from an examiner faced with several slightly inconsistent candidates. A case example will show why (see Results section).

Results

1. When low fit statistics reveal that examiners are not differentiating among candidates of different abilities

Occasionally, examiners have extremely low fit statistics (<0.5). When this occurs, the most likely reason is that the examiner is not distinguishing among relevantly different levels of candidate ability. This hypothesis can be confirmed by examining the distribution of the examiner's ratings among the various rating scale categories. If the overwhelming majority of the examiner's ratings are concentrated in a single rating category, the examiner is not adequately distinguishing among candidate ability (see Table 1). If this pattern is pervasive among examiners, then the problem may lie with the definition of the skills and rating scale categories, which may not allow differentiation at the level necessary for the tested population. However, if this pattern is unusual among examiners, then the problem can be attributed to the examiner's understanding and implementation of those definitions. Table 1 shows a non-discriminating rating pattern and a well-distributed rating pattern. Figures 1 and 2 portray these contrasting patterns graphically.

2. Why high fit statistics are ambiguous with regard to examiner quality

To see why high fit statistics are ambiguous with regard to examiner quality, we turn to a case example. Examiner A was of thoroughly average severity (0.00 logits), and had a very high outfit statistic (1.9). This examiner gave some unexpectedly low ratings to some relatively able candidates. Overall, Examiner A gave 44 highly misfitting ratings (standard residual >=3.0), out of a total of 2329 ratings (2% of ratings). Based on these indicators, Examiner A appears inconsistent.

However, closer inspection of the rating patterns from Examiner A reveals that many of these unexpectedly low ratings were *corroborated* by Examiner B. In other words, Examiner B independently gave the same candidate the same low rating on the same skill and practice area — a response that similarly appeared as a misfit. Thus, the lack of fit may not reflect any problems in Examiner A's rating behavior. Examiner A accurately discriminated and documented fluctuations in candidate ability. Of the 44 misfitting responses by Examiner A, half (22) were corroborated by the examiner's various partners.

Still, this may still be an inflated figure. Skill #6 was so easy that the model expects everyone to score well. When an examiner of average or lower severity gives a middling to excellent candidate a low mark on an easy skill, misfits may result even when the low mark is well deserved. In these instances, the misfit can actually indicate laudable examiner behavior, namely, that the examiner was able to report accurately the levels of performance on a skill where other examiners were non-discriminating. When we further eliminate misfitting responses for Examiner A that occurred on Skill #6, only 10 (0.4%) misfitting responses remain.

Conclusions

As these data and analyses reveal, extreme caution should be exercised before concluding that a misfitting examiner is internally inconsistent and should be excused. Whereas non-discriminating examiners can be accurately identified from model statistics alone, inconsistent examiners can only be identified definitively when additional information can be brought to bear. Specifically, if the examination structure enables examiner ratings to be compared with the ratings of a partner, evidence should be weighed for signs that candidate performances significantly differed from expectation. These data are necessary to determine when significantly elevated fit statistics indicate genuine instances of examiner inconsistency, and when they are produced artifactually from unusual candidate performance patterns.

In contrast, standard outputs can definitively identify examiners who fail to distinguish among candidates of widely varying ability. This pattern of non-discrimination is possible whenever an examiner has extremely low fit statistics (<0.5). The pattern can be definitively confirmed whenever the examiner's ratings overwhelmingly favor a middle category (Table 1). When this pattern of non-discrimination occurs pervasively, the responsibility for the problem may lie with vague or overly inclusive skill definitions and rating scale category definitions. Contrarily, when this pattern occurs in a minority of examiners, the problem may lie in examiners' individual interpretations and implementations of the rating scale. When non-discrimination is a problem only for a minority of examiners, those examiners should be instructed in the types of distinctions that must be recognized among candidates. Sometimes examiners do not adequately discriminate among candidates because the examiners are insecure about their rating abilities. At other times, examiner non-discrimination may reflect a problem in scaling; the examiner may have in mind a much wider ability range than is manifest in the actual population being tested. In either case, if training does not fix the problem, the examiner can justifiably be excused from future examinations.

Table 1. Examiner Use of the Rating Scale*

Rating Category	Percent of Ratings at each level	Mean Candidate Ability at each level
Well-Distributed Rating Pattern: Examiner Fit = 1.00 (Ex. 387)		
Unsatisfactory	4%	-0.76
Marginal	27%	-0.18
Satisfactory	53%	0.74
Excellent	15%	2.52
Non-Discriminating Rating Pattern: Examiner Fit = 0.4 (Ex. 431)		
Unsatisfactory	1%	-1.64
Marginal	2%	-.76
Satisfactory	90%	1.17
Excellent	7%	5.20

* For each examiner, the sample of candidates examined has a distribution of candidate abilities that is consistent with the overall candidate pool. Accordingly, rating patterns do not result from sampling errors.

Figure 1. Probability of Ratings at Each Level (0-3) Graphed against Candidate Ability: Well Distributed Rating Pattern

```
Probability Curves
      -4.0             -2.0              0.0              2.0              4.0
       ++----------------+----------------+----------------+----------------++
     1 |                                                                     |
       |                                                                     |
       |                                                                    3|
       |000                                                              333 |
       |   000                                                        333    |
     P |      00                                                    33       |
     r |        00                                                33         |
     o |          00                                            33           |
     b |            00         111111                          3             |
     a |              0    1111      111        22222222     33              |
     b |               0011             11   222        222 3                |
     i |               110                1*2             3*2                |
     l |             11   00             22 11           3   22              |
     i |           11       00         22     1        33      22            |
     t |         11           0      22        11    33          22          |
     y |      111              00  22            11 3              222       |
       |    11                   **               3*1                 22     |
       |1111                  222  000         333   111                2222 |
       |                   222        000   333         111                 2|
       |             222222           333***000            111111            |
     0 |*************33333333333333333         000000000000000000************|
       ++----------------+----------------+----------------+----------------++
      -4.0             -2.0              0.0              2.0              4.0
```

Figure 2. Probability of Ratings at Each Level (0-3) Graphed against Candidate Ability: Non-discriminating Rating Pattern

```
Probability Curves
      -6.0     -4.0     -2.0      0.0      2.0      4.0      6.0      8.0
       ++--------+--------+--------+--------+--------+--------+--------++
     1 |0                            2222222                            |
       | 00000                  22 222       222222                      |
       |      00              22                  22                  33|
       |        0           22                      22              33  |
       |         0         2                          2            3    |
     P |          0                                    22        33     |
     r |           0      2                              2      3       |
     o |                 2                                2    3        |
     b |            0                                      2  3         |
     a |             0  2                                   23          |
     b |               2                                                |
     i |              0                                     32          |
     l |              2                                    3  2         |
     i |               0                                  3    2        |
     t |             2                                   3      2       |
     y |                0                              33        22     |
       |            2111 0                            3            2    |
       |          1*1   11*                         33              22  |
       |       11*2        *11                    33                  22|
       |  1111122           00*11111        333333                      |
     0 |*******3333333333333333*****************************************|
       ++--------+--------+--------+--------+--------+--------+--------++
      -6.0     -4.0     -2.0      0.0      2.0      4.0      6.0      8.0
```

Food for Thought from Carnot

William P. Fisher, Jr., Ph.D.

LSU Health Sciences Center

The Food

“In fact, it is no bad summing up of Carnot’s work to say that, as the Greeks gave us the abstract ideas (point, line, etc.) with which to think of space, and the 17th century those (mass, acceleration, etc.) with which to think of mechanics, so Carnot gave us those needed in thinking of heat engines. In each case the ideas are so pervasive that we use them even to state that they never apply exactly to visible objects.

“Carnot’s ‘unit of thought’ was the well-known perfectly frictionless, perfectly insulated engine, which gains and loses all its heat at two standard temperatures T and *t*, and imparts motion to nothing except the crankshaft; in particular, not to the particles of the steam. It is therefore ‘reversible,’ that is, capable, on reversal, of transferring all the heat back from sink or condenser to source. The expansions and contractions in it are all either isothermal or adiabatic [involving no loss or gain of heat], and we can reason only about a complete cycle of operations, that is, one which returns the working substance to its original state in *every* respect.

“With such an engine it can be shown to follow that the work done per unit of heat transferred (‘efficiency’) is independent of all details, such as the nature of the working substance, and is in fact simply equal to (T - *t*) / T; otherwise we can get an unlimited amount of work from it without recourse to the source.”

“The ‘Second Law’ [of thermodynamics] was now precisely stated as the impossibility of getting an unlimited amount of heat or work out of a Carnot engine (and, *a fortiori*, out of any other less efficient engine). Clausius (1850) and Thomson (1851) gave equivalent statements of the law ....

“Thomson had been much concerned at the dependence of ‘temperature’ on the properties of a particular gas or liquid; and it was because he saw in Carnot’s work a method of defining an ‘absolute’ (that is, a work) scale (1848) that he welcomed it. To give efficiency not unity (T - *t*) / T, T must be finite. Thus the suggestion, implicit in Charles’ law, of an absolute zero at about -273° C. was confirmed.” [all quotations from Pledge 1939, p. 144]

The Thought

So, in the same way that Plato’s redefinition of the elements of geometry (seeing lines as indivisible planes and points as indivisible lines) dramatically increased the productivity of geometry, and in the same way that Galileo’s thought experiment concerning a perfectly frictionless plane for balls to roll on became the basis for Newtonian mechanics, Carnot’s perfectly frictionless, perfectly insulated engine became the basis for advances in thermodynamics and in temperature measurement.

We thus see in the development of each of these sciences the same criteria and motivations that lead to Rasch’s models, especially the focus on an idealization of the variable as something that can stand on its own independent from the particular details of the specific lines, points, planes, gases, or liquids involved.

What will it take to bring researchers in the human sciences to recognize and accept the validity, utility, and opportunity opened up by these criteria? Is it a fear of reductionism? Is it a math phobia? Is it simply the inertia of existing rewards and motivations that support the status quo?

Or is it the lack of a context that rewards the mathematical coordination of different experiments into a common framework, that assumes total incommensurability as the norm, as seems to be traditional in the human sciences?

I'm betting on the latter and aim to educate, agitate, and lobby for a new measurement culture that values metrological networks and a realization of quantity that follows through on Thurstone's sense of it as the language in which the community of science thinks together. This is what every additional Rasch instrument calibration points to.

When we get to the point at which several instruments intended to measure each of the variables of interest have been calibrated, the commonalties and differences in the calibrations will cry out for explanation, and these explanations will lead to better theories, which will lead to better instruments, which will lead to better data, etc. (Ackermann 1985; Galison 1999). This process will then, in all likelihood, given the historical development of the other sciences, lead to the derivation of conventions for data quality and reference standard metrics.

The French revolutionaries thought they could institute the metric system inside of six months, but it took 50 years, and even now, 150 additional years later, global implementation is still incomplete. Though the efficient thing to do would be to take the bull by the horns and deliberately set out to create rational quantitative measurement in the human sciences, the process will inevitably be fraught with politics, emotions, and the protection of vested interests. We probably won't live to see the day when a metrology system for even a single psychosocial variable is implemented on a broad scale. Probably still less likely will we be around to appreciate the new breeds of research results that will be produced by communities of investigators able to think together in common mathematical languages for the first time. We can, however, help prepare the ground, sow the seeds, and cultivate the plants from which this fruit will grow. And each new calibrated scale brings that day closer.

References

Ackermann, J. R. (1985). Data, instruments, and theory: A dialectical approach to understanding science. Princeton, New Jersey: Princeton University Press.

Galison, P. (1999). Trading zone: Coordinating action and belief. In M. Biagioli (Ed.), The science studies reader (pp. 137-160). New York: Routledge.

Pledge, H. T. (1939). Science since 1500: A short history of mathematics, physics, chemistry, biology. London: His Majesty's Stationery Office [reprinted in 1940 by the Board of Education, Science Museum].

Notes

1. LSU Health Sciences Center, Public Health & Preventive Medicine, 1600 Canal Street, New Orleans, LA 70112 (504-568-8083, 504-568-6905).

Quality Control in Testing

Mark H. Stone, Ph.D.

Adler School of Professional Psychology

Introduction

Quality assurance in testing has been approached in two very general ways. The first has been to assure that test materials are only available to those with the appropriate level of education and training. Most publishers of assessment tools and tests use criteria to assure that only appropriately trained persons have access to their instruments. The second has been to recommend test development procedures following the *Standards for Educational and Psychological Testing* (AERA, 1999). Both approaches are necessary, but not sufficient to assure that quality can be maintained in testing. Neither one provides assurance that test variables are adequately built and maintained. Neither method meets quality assurance standards. To do so requires specific attention to quality control (Stone, 2000). Interestingly, the index to the *Standards* does not include quality or quality control as headings, but statistical quality control has long been employed to assure the highest standards in manufacturing goods.

The earliest and most systematic exposition of quality control was given by Walter Shewhart of Bell Laboratories (1931, 1986). His efforts have been propagated through the lectures and writings of W. Edwards Deming also well-known for his work in quality control. The problem of quality control in testing has been frustrated by several fundamental conceptual issues. The first is addressed by Deming in his introduction to the reprint of Shewhart's *Statistical Method* (1986). Deming writes:

> There is no true value of anything. There is, instead, a figure that is produced by application of a master or ideal method of counting or measurement. This figure may be accepted as a standard until the method of measurement is supplanted by experts in the subject matter with some other method and some other figure. (p. ii)

Deming goes on to point out that all values and constants are in error because they are conditioned by the methods of their determination. "Every observation, numerical or otherwise, is subject to variation" (1986, p. ii). However, there is useful information in variation. The issue is not just error, but control over error. The second issue raised is the need for a method for establishing quality control. Error must be brought under control if the resulting values are to have any practical use. Shewhart's model for statistical control over error requires answers to these five questions:

1. How are the observations to be made?
2. How are the samples to be drawn?
3. What is the criterion for control?
4. What action will be taken as a consequence?
5. What quantity of data is required?

He arranged these questions into a dynamic model (see next page):

Figure 1. Shewhart's Dynamic Model

SPECIFICATION -----> PRODUCTION ------> INSPECTION

(Their arrangement is best seen as a circle — or even better, as a continuing spiral)

Accuracy and Precision

Quality control in testing requires addressing accuracy and precision. The concepts of accuracy and precision in testing and measurement are called validity and reliability. The first and most important matter is to determine what any concept means. Bridgman (1928) specified what has come to be known as an "operational definition." Such a definition serves as the mechanism for understanding what a concept means. He indicated that, "The concept is synonymous with the corresponding set of operations" (p. 5). Concepts are defined, explicitly or implicitly, by a methodology. A concept equals the method that describes it and vice versa. Concepts without methods are nonsense and the bantering about of concepts without considering methods is irresponsible and not scientific. Therefore, concepts such as accuracy and precision, validity and reliability cannot be separated from the methods of their determination. To be specific, we cannot speak of validity or reliability, but only of some method of determining validity or reliability specified about some occasion.

A specific example of the confusion over concepts can be observed when reading reports of a test's "validity." The determination of validity is situational and not extensive. The validity of the test is conditioned by a point in time and the setting in which it took place, i.e., the methodology and sample producing the value(s). It is an inferential leap to assume that what occurred in one circumstance has any application to another circumstance and it is even less likely to expect it apply to every other instance or to all other circumstances. Concepts such as validity and reliability require more careful inquiry. Specifically, a determination of validity or reliability needs to be operationally decomposed into two important aspects: the contribution from the items, and contribution from the persons. Typically, and all too often, studies of test validity and reliability fail to provide any coefficient resulting from use of the sample.

For determining reliability, the KR20 is often calculated for items, but almost never for persons. Hoyt (1941) recognized both approaches, saying that "extended examination of the 'among items' variance would make it possible to decide on the heterogeneity of the respective difficulties of the items while a more extended examination of the 'among students' variance would make it possible to answer certain pertinent questions regarding the individual differences among students" (p. 41). His good advice is almost never followed. Jackson (1939), Hoyt (1941), Alexander (1947), and Guilford (1954) have all proposed an analysis-of-variance approach to estimate reliability. The advantage of this strategy is that "test reliability" can be decomposed into the variance due to examinees, the variance due to items and the remainder or error variance. This more complete analysis is in keeping with a quality control process in testing.

Wright and Stone (1999) have demonstrated that these matters can be even better accomplished using Rasch measurement techniques which are explained in *Best Test Design* (Wright &

Stone, 1979, pp. 151-166) and all of the analyses discussed below can be produced using WINSTEPS (Linacre, 2000). The shortcomings of using raw scores are remedied when a Rasch measurement analysis is made of the same data and reliability is calculated from Rasch values. In addition, Rasch measurement provides the standard errors for every person and item. These individual errors can be squared and summed to produce a correct average error variance for the sample or any subset of persons and for the items or any subset of items. When these results are substituted for those in the traditional KR20 formula, the result is a new formula, equivalent in interpretation, but giving a better estimate of reliability than any other value produced by using raw scores. Deming's adage of progressive improvement by better methods in quality control is clearly demonstrated through applying these methods. Shewhart (1986) also spoke of predication as an important aspect of quality control. "Every meaningful interpretation involves a prediction" (p. 92) and "Knowledge in this sense is a process or a method of predicting an ideal" (p. 104). The element of prediction makes scientific results useful. In the application of a test, it is the characteristics of the new sample to which we intend to apply the test, rather than simply the description of a previous sample, that is our focus. We want to know how the test will work with the new samples who are about to take it, not old history. We want a relevant reliability coefficient which applies to the people we intend to test, not one that only describes the people who were previously tested. But we can actually predict the reliability for a new sample if we postulate the mean and variance for that sample. One can use these statistics and the Rasch targeting formula to calculate the reliability of the test in its new application. (See Wright & Stone 1979, 129-140.)

Deming, as quoted above, indicated that new methods can supplant old ones when they provide better methods and values. The Rasch separation index is such a method for producing a more useful value. Correlation-based reliability coefficients are nonlinear. The increase in reliability from .5 to .6 is not twice the improvement in reliability from .9 to .95. In fact, the increase from .9 to .95 is actually about twice the improvement in precision of the other. The Rasch Separation Index (G) is the ratio of the unbiased estimate of the sample standard deviation to the root mean square measurement error of the sample. It is in a ratio scale in the metric of the root mean square measurement error of the test for the sample postulated. The Separation Index quantifies "reliability" in a more direct way with a clear interpretation.

Separation G = SDT/SET

SDT = The expected SD of the target sample

SET = The test standard error of measurement for such a sample, which is almost always well approximated by SET = 2.5 / □ L

SET can also be estimated as SET = □(C/L) where L is the number of items in the test and C is a targeting coefficient (see Wright & Stone, 1979, pp. 135-136). A figure given below expedites applying this procedure (see pp. 22-23 for remaining figures).

The *Standards* (AERA, 1999) in Section 13.14 recommend that "score reports should be accompanied by a clear statement of the degree of measurement error associated with each score" (p. 149). Rasch measurement analysis routinely provides standard errors for every possible test measure along the variable that fully meets this recommendation. If reliability, as defined by the *Standards*, is the degree to which test scores are free from errors of measurement, then it follows that every ability measure should be accompanied by a standard error as an index of the degree to which this criterion is met for that measure. Not to do so is to ignore the *Standards*.

The Rasch measurement standard errors satisfy this recommendation by providing individual errors of measurement for every observable measure. Where a collective index of reliability is desired, the Rasch Separation Index is even more useful than the traditional indices of reliability. Figure 2 describes the Rasch analysis of a response matrix and Figure 3 describes the computation of the Rasch person separation index. The targeting coefficient C varies between 4 and 9 depending on the range of items difficulties in the intended test and the target sample's expected average percent correct on that test. Figure 4 gives some values of C for typical item difficulty ranges and typical target sample mean percents correct. However, it is not the algebraic and statistical similarity of the KR20 and the Separation Index C that is of major importance. Instead it is the decomposition of these single indices into their constituent parts leading to a more detailed and more useful management of information. Quality control is now operating.

With Rasch measurement analysis, we are able to obtain the standard error of calibration for each individual item as well as the standard error of measurement for each person ability. With traditional methods, a single standard error of measurement is provided and only for measures at the group mean of person ability. The standard error specific to each item or person statistic is far more useful than any single sample or test average.

The location of each item and person on a line representing the variable together with their standard errors provides definition and utility to the test variable. The definition of the variable is specified by the location of the items. The utility of a test variable for measuring persons is quantified by the standard error that accompanies each person measure.

A variable can be thought of as a straight line. To measure successfully we must be able to locate both items and persons along this line. A simple example is given in Figure 5. Items are located by the number of persons getting a specific items correct. Persons are located by how many items they were able to answer correctly. Items to the left side of the line are easier than those to the right while persons to the left have less ability than others to the right.

It is necessary to locate persons and items along the line of the test variable with sufficient precision to "see" between them. Items and persons must be separated along this line for useful measurement to be possible. Separation that is too wide usually signifies gaps among item difficulties and person abilities. Separation that is too narrow, however, signifies redundancy for test items and not enough differentiation among person abilities to distinguish between them. Items must be sufficiently well separated in difficulty to identify the direction and meaning of the test variable. To be useful, a selection of items, i.e., a test, must separate relevant persons by their performance. The item locations are the operational definition of the variable of interest while the person locations are the application of the variable to measurement. Such an approach meets Bridgman's requirements for an operational definition.

Conclusion

Item and person separation statistics in Rasch measurement provide analytic and quality control tools by which to evaluate the successful development of a variable and by which to monitor its continuing utility. Successful item calibration and person measurement produces a map of the test variable (Stone, Wright, & Stenner, 1999). The resulting map is no less a ruler than the ones constructed to measure length. The map indicates the extent of content, criterion, and construct validity for the test variable. Empirical calibration of items and measures of persons should correspond to the original intent of item and person placement. Changes must be made when correspondence is not achieved. Rasch measurement provides the quality control necessary in testing.

There should be continuous dialogue between the plan for the test, the items calibrations, and person measures. Test variables are never created once and for all. Continuous quality control is required in order to keep the map coherent and up-to-date. Support for reliability and validity does not rest in coefficients, but in substantiating demonstration of relevance and stable indices for items and measures. Such procedures assure quality control in maintaining the test variable and assuring its relevance.

References

American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA. Author.

Alexander, H. (1947). The estimation of reliability when several traits are available. *Psychometrika, 12,* 79-99.

Bridgman, P. (1928). *The logic of modern physics*. New York: Macmillan.

Guilford, J. (1954). *Psychometric methods*. New York: McGraw-Hill.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6,* 153-160.

Jackson, R. (1939). Reliability of mental tests. *British Journal of Psychology, 29,* 269-287.

Linacre M. (2000). *WINSTEPS*. Chicago: MESA.

Shewhart, W. (1931). *Economic control of quality of manufactured products*. New York: Van Nostrand.

Shewhart, W. (1986). *Statistical method from the viewpoint of quality control*. New York: Dover.

Stone, M. (2000). *Establishing quality control in testing*. Second International Congress on Licensure, Certification and Credentialing of Psychologists, Oslo, Norway, July 20, 2000.

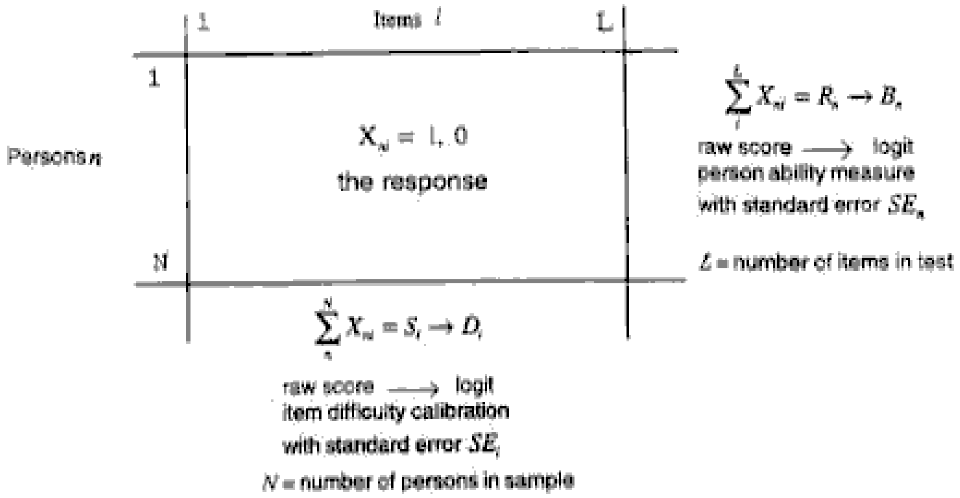
Stone, M., Wright, B., & Stenner, J. (1999). Mapping variables. *Journal of Outcome Measurement, 3, (4),* 308-322.

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

Wright, B., & Stone, M. (1999). *Measurement essentials*. Wilmington, DE: Wide Range, Inc.

Figure 2

Rasch Analysis of Response Data


1
Items i
L
1
Persons n
$X_{ni} = 1, 0$
the response
N
$\sum_{i}^{L} X_{ni} = R_n \rightarrow B_n$
raw score ⟶ logit
person ability measure
with standard error SE_n
L = number of items in test
$\sum_{n}^{N} X_{ni} = S_i \rightarrow D_i$
raw score ⟶ logit
item difficulty calibration
with standard error SE_i
N = number of persons in sample

Items: Logit item difficulty calibration:

$$D_i = \sum_{n}^{N} B_n / N_n + \left[1 + \left(SDB^2 / 2.9\right)\right]^{1/2} \log\left[\left(N - S_i\right) / S_i\right]$$

$$SDB^2 = \sum_{n}^{N}\left(B_n^2 / N\right) - \left(\sum_{n}^{N} B_n / N\right)^2$$

Logit item difficulty calibration error variance:

$$SE_i^2 = \left[1 + \left(SDB^2 / 2.9\right)\right]\left[N / S_i \left(N - S_i\right)\right]$$

Persons: Logit person ability measure:

$$B_n = \sum_{i}^{L} D_i / L_i + \left[1 + \left(SDD^2 / 2.9\right)\right]^{1/2} \log\left[R_n / \left(L - R_n\right)\right]$$

$$SDD^2 = \sum_{i}^{L}\left(D_i^2 / L\right) - \left(\sum_{i}^{L} D_i / L\right)^2$$

Logit person ability measure error variance:

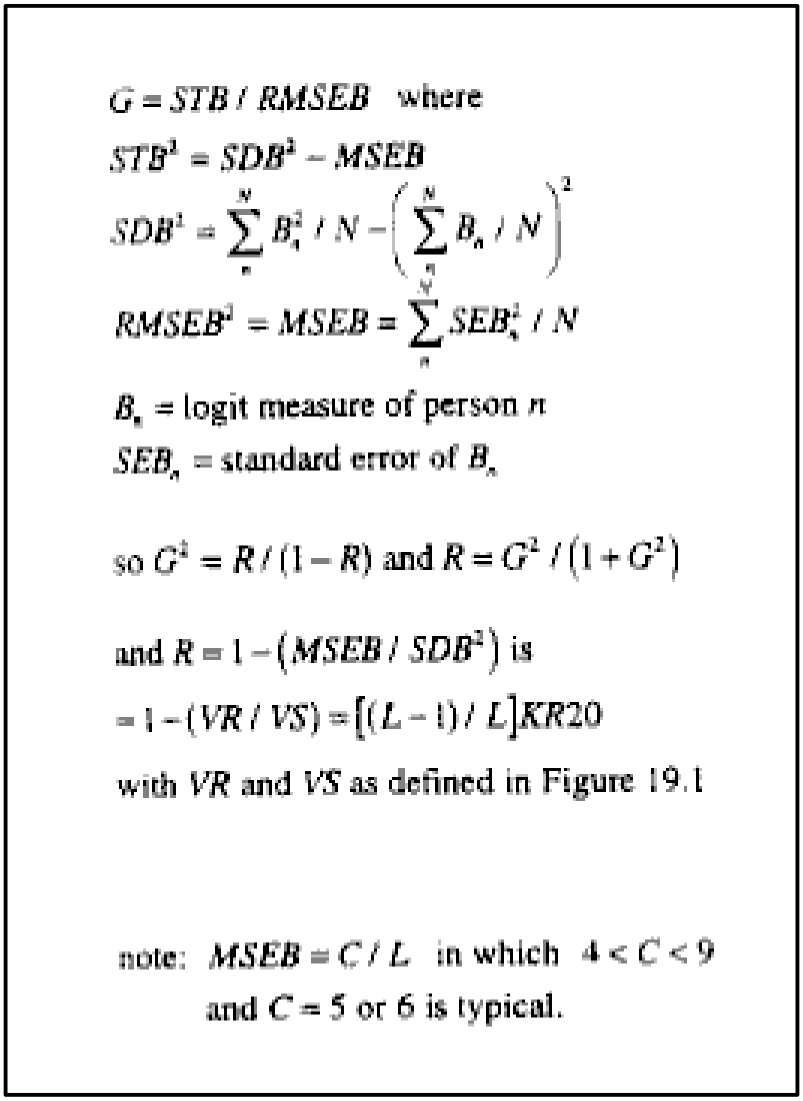
$$SE_n^2 = \left[1 + \left(SDD^2 / 2.9\right)\right]\left[L / R_n\left(L - R_n\right)\right]$$

$2.9^{1/2} = 1.7$ = scaling factor between logistic ogive and normal ogive

(See PROX estimation formulas, Wright and Stone, 1979, pp. 21-22)

Figure 3

Rasch Person Separation Index



$G = STB / RMSEB$ where

$STB^2 = SDB^2 - MSEB$

$SDB^2 = \sum_n^N B_n^2 / N - \left(\sum_n^N B_n / N \right)^2$

$RMSEB^2 = MSEB = \sum_n^N SEB_n^2 / N$

B_n = logit measure of person n

SEB_n = standard error of B_n

so $G^2 = R / (1 - R)$ and $R = G^2 / \left(1 + G^2\right)$

and $R = 1 - \left(MSEB / SDB^2\right)$ is

$= 1 - (VR / VS) = \left[(L - 1) / L\right] KR20$

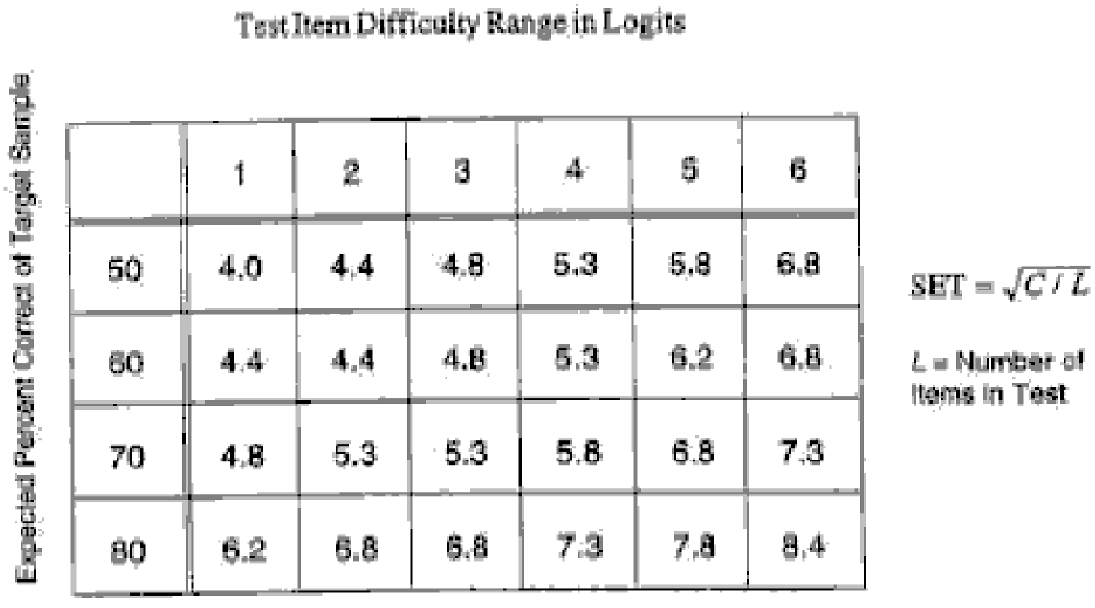
with VR and VS as defined in Figure 19.1

note: $MSEB = C / L$ in which $4 < C < 9$
and $C = 5$ or 6 is typical.

(See Wright and Stone, 1979, pp. 134-136)

Figure 4

Values of the Targeting Coefficient C



Test Item Difficulty Range in Logits

Expected Percent Correct of Target Sample

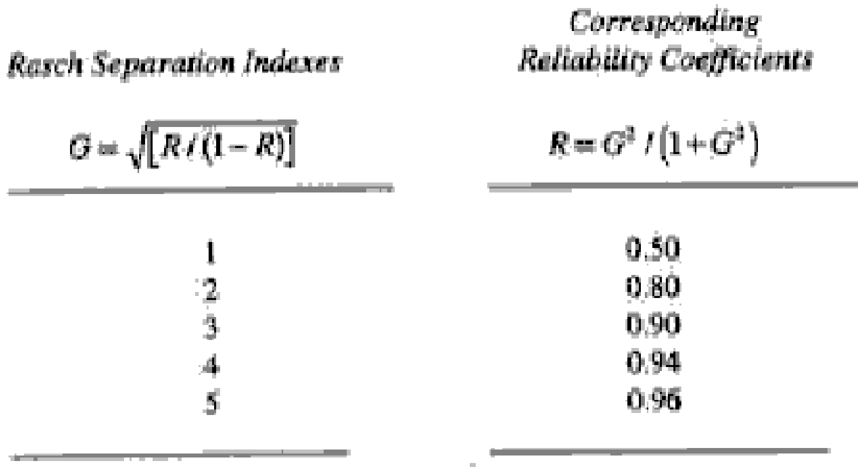
	1	2	3	4	5	6
50	4.0	4.4	4.8	5.3	5.8	6.8
60	4.4	4.4	4.8	5.3	6.2	6.8
70	4.8	5.3	5.3	5.8	6.8	7.3
80	6.2	6.8	6.8	7.3	7.8	8.4

SET = $\sqrt{C / L}$

L = Number of Items in Test

(See Wright and Stone, 1979, p. 214)

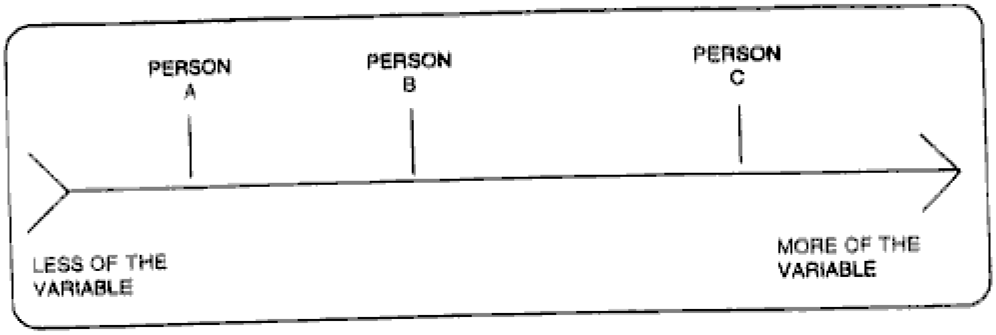
If an expected reliability is also desired, it can be obtained from: $R = G^2 / \left(1 + G^2\right)$.



Rasch Separation Indexes $G = \sqrt{\left[R / (1 - R)\right]}$	*Corresponding Reliability Coefficients* $R = G^2 / \left(1 + G^2\right)$
1	0.50
2	0.80
3	0.90
4	0.94
5	0.96

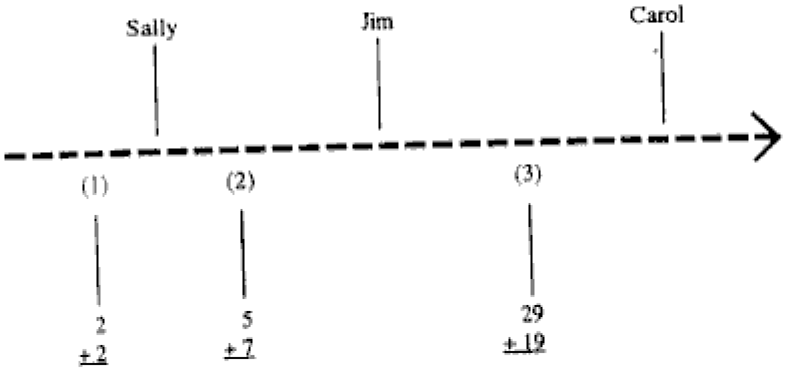
Figure 5

Positions of persons A, B, C on the line of the variable



Once the variable is constructed by the line of items, we can proceed to position students on this same line. Their probable positions can be specified initially by our best guess as to their ability to correctly answer the items which define the variable. The line of our variable shows both the positions of items and the positions of students. Eventually the positions of students will become more explicit and more empirical as we observe what items they correctly answer.

Consider this picture:



Sally's position on the variable is indicated by an expected correct response to Item 1 but expected incorrect responses to Items 2 and 3. Her differing responses to Items 1 and 2 locate her on the variable between two items that describe her ability in arithmetic computations. She can add 2 and 2 but not 5 and 7.

Jim's position is between Items 2 and 3 because we expect him to answer Items 1 and 2 correctly but not item 3. In Jim's case we have somewhat less precision in determining his arithmetic ability because f the lack of items between Items 2 and 3. If we had additional items in this region, we could obtain a more accurate indication of Jim's position on the variable as defined by his responses to these additional items.

Rule-based Aptitude Measurement: Artistic Judgment

Nikolaus Bezruczko, Ph.D. and Ambra Borgognoni Vimercati

Measurement and Evaluation Consulting

Introduction

In the 1920s educators and psychologists became aware of artistic judgment's importance for improving vocational selection and career development. They emphasized testing artistic judgment to select persons best suited for professions and occupations ranging from dentistry and architecture to window display and hair dressing (McAdory, 1929). (But what about plastic surgery, interior decorating, and film making? The instrumental influence seems almost limitless!) Unfortunately, valid and effective tests proved elusive for most of the twentieth century.

In retrospect, a major limitation was artistically dismal visual images serving as test items. Poor items in fact lead to some horrible disasters. The Meier Art Tests (1940), for example, presented old-world, masterpiece reproductions with an explicitly adulterated counterpart. Persons with high artistic judgment presumably would select the masterpieces. The artistic community (artists, museums, galleries, and so on) reacted with such shock and disbelief at this obvious violation of artistic dignity that they mocked and discredited all further attempts to test artistic judgment. To avoid embarrassment, commercial publishers abandoned artistic judgment testing, which to this day is hardly emphasized despite its obvious social benefits. The few contemporary tests purporting to measure artistic judgment have inconsistent, vague, or unknown reliability and validity.

Empirical Challenges

Twentieth century social researchers did substantially advance in understanding how visual images appeal to preference. Birkhoff, a noted mathematician, proposed that two image characteristics, complexity and uniformity, are functionally related to visual preference (1932). He systematically manipulated polygon sides to illustrate these characteristics but left empirical verification to others. Later, Hans Eysenck (1940, 1941) picked up this line by factor analyzing visual preferences for polygons. He found "T" and "K", described T as a general Taste factor, and K as a bipolar factor that distinguishes between artists and nonartists. Unfortunately, Eysenck's contributions were only based on raw score correlations and weak true score methods which ultimately would undermine his effort to test artistic judgment aptitude. Because visual artists logically appear at the extreme high end of the artistic judgment aptitude distribution where raw score distortions are greatest, a test derived from Eysenck's T, Visual Aesthetic Sensitivity Test (VAST) (Götz, 1981), was constantly plagued with validity and reliability problems. [Linear group differences in distribution tails only need to be 50% as large as their raw score differences to be statistically significant (Wright & Masters, 1982; see also Wright, 1999).]

In the 1960s, Berlyne experimentally showed that image complexity follows a curvilinear preference function (1971, 1974). Visual preference and image complexity monotonicly increase until complexity reaches a maximum, then preference steadily declines. Unfortunately, an outcome of his research was incredible confusion concerning complexity's influence on artistic judgment. Because his studies did not include artists, Berlyne and most other social researchers incorrectly assumed that preference for complexity is indicative of higher artistic judgment. Contemporary social researchers are typically astounded to learn that Berlyne's complexity function is *inversely* related to artistic judgment. In fact, complexity has profoundly different effects on artists and nonartists. Not grounded on objective measuring methods, this confusion concerning complexity and artistic judgment continues in contemporary research.

In addition to using "bad" art, twentieth century artistic judgment research suffered from:

- sample dependent, norm-based statistics
- nonlinear raw scores
- vaguely defined constructs
- deficient validity studies

Contemporary Advances

In the early 1980s, Johnson O'Connor Research Foundation (JOCRF), the oldest and largest aptitude testing organization in the United States, received a special gift from Christian A. Johnson Endeavor Foundation to develop new tests for its aptitude battery. Sensitive to artistic judgment's influence on occupations, JOCRF undertook to develop a reliable and valid artistic judgment test. JOCRF recruited the first author to research prior efforts, then to design an objective model for testing artistic judgment. Over the course of eight years, the outcome of this effort was remarkably successful. Exploiting the discriminant implications of Eysenck's K factor, he adapted a stochastic sampling model proposed by Attneave (1959) to construct rule-based images that varied only in complexity and redundancy. Moreover, these images emulated the contemporary art style known as Minimalism. According to Eysenck's K factor, dichotomously scoring image preferences that systematically vary in complexity should distinguish between artists and nonartists. Thirty-five items were bound into booklets, the Visual Design Test (VDT; 1987), which JOCRF extensively studied for psychometric properties in their testing offices (1989, 1990). VDT was finally validated with professional artists and other studies examined developmental implications among school children, and differences between artists and nonartists (Bezruczko & Schroeder, 1989, 1990, 1991, 1994, 1995; Schroeder & Bezruczko, 1990, 1998).

VDT Figurative Project

Despite this extraordinary success testing artistic judgment, JOCRF was reluctant to include VDT in its standard test battery. The extreme minimalism of VDT images was never accepted by the artistically staid foundation, and its staff frequently commented that VDT Abstract images "just don't look like art." In 1999 JOCRF asked Ambra Borgognoni Vimercati, an Italian fresco artist in Rome, to paint figurative images derived from rule-based VDT Abstract images. The purpose was to establish whether figurative paintings conforming to traditional, Western art standards and derived from an algorithm could be psychometrically reliable and valid. She painted 20 canvases in five styles: Renaissance, Baroque, Impressionism, Fauvism, and Surrealism, manipulating only their complexity. (In each style, she painted four variations of a theme.) VDT Figurative images, however, fundamentally differ from VDT Abstract because she defined artistic judgment as "coherent integration of multiple image elements in a theme" and systematically introduced unnecessary elements to increase complexity. Complexity in VDT Abstract is only manipulated by element frequency. This difference in operational definition motivates the present study.

Using the same item format as VDT Abstract, less-complex figurative paintings were paired with more-complex paintings and published as VDT Figurative (Bezruczko, Borgognoni Vimercati, & Calipari, 2000). JOCRF then administered these items in their testing offices in Chicago, New York, Boston, and Dallas to study their measurement properties. Figure 1 (see p. 28) shows VDT Abstract, Figurative, and their co-calibration to a common scale using WINSTEPS (Linacre & Wright, 2000). In this article, we present preliminary results from this study.

The sample was predominantly white, educated, and socioeconomicly upper middle class. Their arts background, however, was generally modest and none were professionally-trained artists. Co-calibration shows these dichotomously scored items are well targeted on 244 JOCRF examinees with generally acceptable fit values. [Infit and outfit meansquares are 1.00 and .99, respectively, for both persons and items. Person separation = 2.35 (reliability = .85) and item separation = 4.32 (reliability = .95).] (An issue concerning poor item fit for several hard items was accommodated by removing four persons with implausible preferences. We speculate high item difficulty elicited some random responses.) Residual analysis shows a systematic factor structure accounting for 12 percent of residual variation. (Factor 1 explains 5.04 of 42 residual variance units.) Residual variation revealed a second factor we call Uniformity that is consistent with prior factor analyses.

Figure 2 (see p. 29) shows that this variable is defined by a coherent order of figurative items. Impressionism, Surrealism, and Baroque styles define the lower portion while Renaissance and Fauvism define the upper portion. Because items are scored for their agreement with artist preference, the lower portion shows items on which it was *easy* to agree with artists, while items higher on this variable are much harder. Consequently, persons high on this continuum, though artistically untrained, tend to agree with preferences of professional artists, suggesting they have a natural "eye" for art. Approximately 12 percent fell in this category. The clustering of styles on this continuum reveals that Surrealism and Impressionism are substantially easier to pass (agree with professional artists), while Fauvism and Renaissance styles tend to be hardest.

Insights and Conclusions

This research departs from traditional artistic judgment studies because it examines visual preference within the quantitative rigor of an invariant linear framework. There it applies knowledge acquired from studies of synthetic, rule-based images to paint figurative art. A surprising result was that the influence of complexity on agreement with artists does not dramatically differ between abstract and figurative art. Co-calibration of VDT Abstract and Figurative items show them intermingled and supported by reasonable fit values.

Complexity, however, does appear to profoundly influence preference for representative and nonrepresentative figurative art. Complexity's influence on preference for representational figurative styles such as Renaissance, Baroque, and Fauvism is predictable — more complex is always preferred over less complex. However, when this nonartist sample was presented nonrepresentational figurative styles, such as Impressionism and Surrealism, they inexplicably abandoned more-complex images and preferred less-complex images! We speculate that items using nonrepresentational images "overload" viewers with uninterpretable content, leading to rejection of more-complex alternatives, and inadvertently increasing preference for less-complex, art-preferred images. Representational, figurative images, on the other hand, provide familiar, meaningful content, which allows complexity to stimulate viewers to higher preference but less agreement with professional artists.

These preliminary results are being extended by analyses that compare person measures for VDT Abstract and Figurative items separately to assess construct comparability (Bezruczko, in press). Further validity studies will verify consistency with professional artist preferences. Our implementation of an invariant measurement structure will make cross-cultural generality of these results relatively easy to examine. This structure will also be useful for assessing applicability of the underlying complexity principle, that is, association between complexity and nonartist preferences across other art styles. This capacity to replicate offers rare opportunities to extend knowledge about artistic judgment and better understand its role in professions and occupations. Finally, the linearity and invariance of this framework will be important to understanding the functional relationship between complexity and representationalism.

References

Attneave, F. (1959). Stochastic composition processes. *Journal of Aesthetics and Art Criticism, 17,* 503-510.

Berlyne, D. E. (1971). *Aesthetics and psychobiology*. New York: Appleton-Century-Crofts.

Berlyne, D. E. (1974). *Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation.* Washington, DC: Hemisphere Publishing Corporation.

Bezruczko, N., Borgognoni Vimercati, A., & Callipari, P. (2000). *Visual Designs Test: Figurative pilot edition.* Rome, Italy: Courage Press.

Bezruczko, N., & Schroeder, D. H. (1987). *Visual designs test: Abstract.* Chicago: Johnson O'Connor Research Foundation, Inc.

Bezruczko, N., & Schroeder, D. H. (1989). *Technical report 1989-2. Artistic Judgment Project I: Internal-structure analyses.* Chicago: Johnson O'Connor Research Foundation, Inc.

Bezruczko, N., & Schroeder, D. H. (1990). *Technical Report 1990-1. Artistic Judgment Project II: Construct validation.* Chicago: Johnson O'Connor Research Foundation, Inc.

Bezruczko, N., & Schroeder, D. H. (1991). *Technical Report 1991-1. Artistic Judgment Project III: Artist validation.* Chicago: Johnson O'Connor Research Foundation, Inc.

Bezruczko, N., & Schroeder, D. H. (1994). Differences in visual preferences and cognitive aptitudes of professional artists and nonartists. *Empirical Studies of the Arts, 12,* 19-39.

Bezruczko, N., & Schroeder, D. H. (1995). Development of visual preferences in art-trained and non-art-trained schoolchildren. *Genetic, Social, and General Psychology Monographs, 122,* 179-196.

Birkhoff, G. P. (1932). *Aesthetic measure.* Cambridge, MA: Harvard University Press.

Eysenck, H. J. (1940). The general factor in aesthetic judgments. *British Journal of Psychology, 3,* 94-102.

Eysenck, H. J. (1941). Type factors in aesthetic judgments. *British Journal of Psychology, 31,* 262-270.

Götz, K. O. (1981). *Visual aesthetic sensitivity test.* Dusseldorf: Concept Verlag.

Linacre, J. M. (2000). *WINSTEPS: Rasch analysis for all two-facet models, version 3.04.* Chicago: MESA Press.

McAdory, M. (1929). *The construction and validation of an art test.* New York: Bureau of Publications, Columbia University.

Meier, N. C. (1940). *Meier Art Tests: Part I. Art judgment.* Iowa City: Bureau of Educational Research and Service, University of Iowa.

Schroeder, D. H., & Bezruczko, N. (1990). *Visual preference dimensions that distinguish artists and nonartists.* Paper presented at Annual Convention of American Psychological Society, Dallas.

Schroeder, D. H., & Bezruczko, N. (1998). *Technical report 1998-1. Artistic judgment: Review of project and study of school children.* Chicago: Johnson O'Connor Research Foundation, Inc.

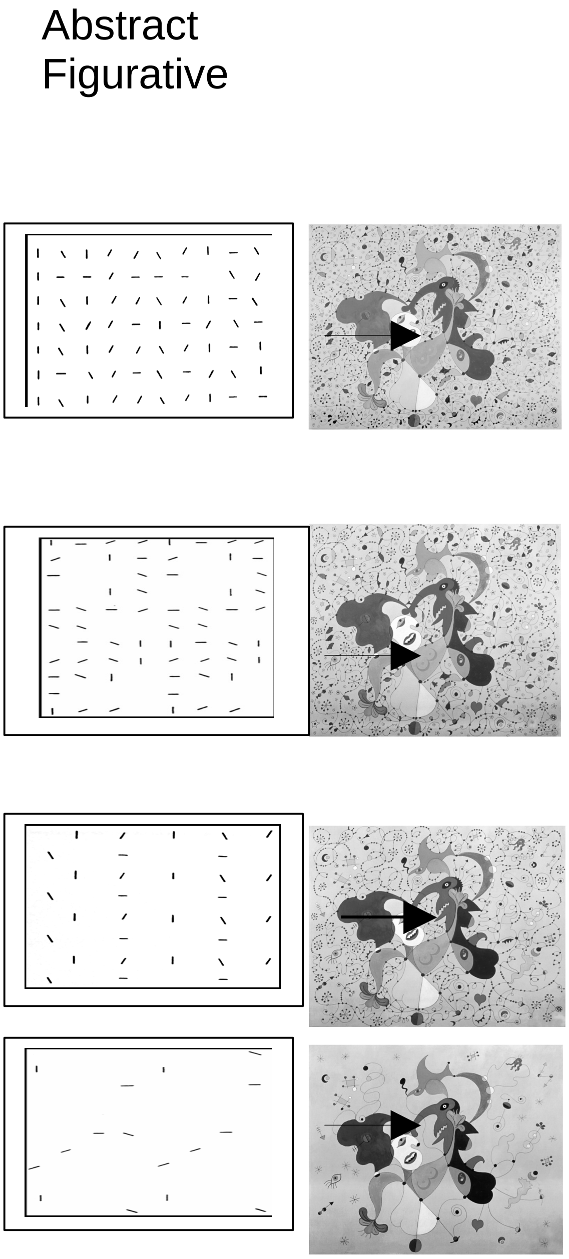
Wright, B. D. (1999). Fundamental measurement in psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement.* Mahwah, NJ: Lawrence Erlbaum.

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

Notes

1. Nikolaus Bezruczko, Ph.D., is a MESA graduate and co-founder of the Chicago Objective Measurement Table (COMET).
2. Ambra Borgognoni Vimercati is a professional artist specializing in fresco restoration based in Rome, Italy. Her current work emphasizes emic characteristics in visual art.
3. We gratefully acknowledge cooperation of Johnson O'Connor Research Foundation, and in particular Dr. David H. Schroeder, Research Manager, and Robert F. Kyle, Director of Research, in collecting data for this research. We are also indebted to Professor Ben Wright for permitting us to collect responses to mockups for VDT Figurative paintings from his students. Their responses were instrumental in producing effective figurative items. Sample VDT Abstract and Figurative items are at: www.artisticjudgement.com.

Figure 1. Co-calibration of VDT Abstract and Figurative



```
PERSONS MAP OF  ITEMS
         <more>| <rare>
               +
               |
            .  |
               |
               |
               |
               |
            .  |
              2+
            .  |
            #  |
              T|
               |
               |T
               |
            .  |  Abs2
           .# 1+  MAT13  MAT9
          .##  |  Abs20  MAT18
           .#  |  Abs4
          ### S|S Abs1 Abs12  Abs16  Abs21
           ##  |  Abs3   Abs5   MAT5   REN6
      .######  |
          .##  |
         ####  |  Abs10  Abs19  Abs6   Abs8
          ### 0+M Abs18 Abs27 B3 R16 S2 S7
     .#######  |  Abs17  Abs9 B20 B8 R12
    #########  |  B15 R1 S11 S17
         .### M|  Abs25
         .###  |  Abs33
          ###  |S
         .###  |
          ###  |
         .###-1+  MIR4
           ##  |  Abs14  Abs31  MIR19
          .##  |  Abs35  MIR10
          .## S|T
               |  MIR14
           ##  |
           .#  |
               |
            #-2+
            .  |
              T|
            .  |
               |
               |
          .##  |
               |
             -3+
               |
            .  |
               |
               |
               |
               |
            .  |
               +
         <less>|<frequ>
```

More artistic judgment

22 Abstract items
20 Figurative items
244 persons

Less artistic judgment

Note. Abstract images are constructed by an algorithm that randomly assigns elements to a design. Painted figurative images conform to complexity level in corresponding VDT Abstract. Both VDT Abstract and Figurative images were co-calibrated to define a common artistic judgment aptitude construct.

Figure 2. Figurative item profiles

```
                                                            Painting            Complexity
                                                              style               level
-4    -3    -2    -1     0      1      2      3      4
|------+------+------+------+------+------+------+------|
0                                   1                   1   Fauvism9            3rd
0                                   1                   1   Fauvism13           1st
|                                                       |
0                                  1                    1   Fauvism18           2nd
|                                                       |
|                                                       |
0                                1                      1   Fauvism5            ---
0                               1                       1   Renaissance6        ---
|                                                       |
|                                                       |
0                            1                          1   Renaissance16       1st
0                            1                          1   Surrealism7         2nd
0                            1                          1   Surrealism2         1st
0                            1                          1   Baroque3            3rd
0                           1                           1   Baroque8            ---
0                           1                           1   Baroque20           1st
0                           1                           1   Renaissance12       3rd
0                          1                            1   Renaissance1        2nd
0                          1                            1   Surrealism11        3rd
0                          1                            1   Baroque15           2nd
0                          1                            1   Surrealism17        ---
|                                                       |
|                                                       |
0                     1                                 1   Impressionism4      1st
|                                                       |
0                   1                                   1   Impressionism19     2nd
0                   1                                   1   Impressionism10     3rd
|                                                       |
0                 1                                     1   Impressionism14     ---
|------+------+------+------+------+------+------+------|
-4    -3    -2    -1     0      1      2      3      4

                 1  11 2 22 3 21 1 1
1     2    5   9 0  31 7152 2 38 1 57 12 4        1        PERSONS
              T     S     M      S     T
```

Note. All forced-choice items are keyed by artist preference. Nonrepresentational figurative styles define lower portion of this variable, indicating items on which it is easier to agree with artists. Upper portion is defined by representational figurative images that are very difficult. We speculate they identify persons commonly referred to as "natural artists."

An Introductory Text on Rasch Measurement

Trevor G. Bond, Ph.D.

James Cook University

Applying the Rasch Model: Fundamental Measurement in the Human Sciences
by Trevor G. Bond and Christine M. Fox

Birth of a New Book

I guess I just grew sick and tired of the same old request after almost every presentation I made at conferences involving developmental psychologists: “Trevor, could you just give me a simple ten minute explanation of what Rasch analysis is all about?” After a dozen or so inquiries of this nature, I thought there must be a shortcut. A couple of us Piagetians gathered and developed a little web site for developmentalists interested in Rasch analysis. When we tried to suggest introductory readings for our colleagues — many of whom had very little explicit training in statistics or quantitative research methodology — we realized that we could not produce the goods.

That’s not to deny in any way the importance of the classic Rasch texts — Wright and Stone, Andrich, Wright and Masters, and that of Georg Rasch himself. It highlighted, however, the absence of an introductory text — a text that would take an interested beginner to the point of being able to participate in the wider Rasch collegial community.

So really, I had the plan for the book in my mind for more than a couple of years. And the book would still be just that — plan in my mind — if I hadn’t had the good fortune to meet Christine Fox at a Midwest Objective Measurement Seminar (MOMS) meeting in Chicago. Consequently, I was able to read the “Uses of Rasch Modeling Counseling Psychology of Research” (Fox & Jones, 1998). I was impressed enough to suggest to Christine that we could collaborate in writing this book. I then sent her a copy of the book plan and the single chapter that I had already written in the hope of convincing publisher Erlbaum to undertake the publishing task.

Thanks to the marvels of e-mail communication, and a more than occasional phone call, we started the collaboration that has produced this volume. Long time Piagetian colleague Bill Gray organized an opportunity for me to visit the University of Toledo as a Distinguished Visiting Professor. James Cook University granted me a half year sabbatical in 2000 which resulted in the first draft of the complete text being sent to the publisher the day before I flew back to Australia!

One of the best aspects of visiting the University of Toledo was the opportunity to participate in teaching Christine’s course on Rasch measurement to graduate students in education and other related disciplines. We had decided to use key chapters from the draft of our book as the core material for the students, and took turns presenting those ideas to a lively bunch of committed and critical future researchers. This reinforced an idea that I had long cherished: that teaching students had always been the most important stimulus to my own learning.

While the book had started off as a sort of ‘how-to’ text for developmental psychologists, the editor and a number of our colleagues encouraged us to broaden the scope of our writing. More importantly, the emphasis on the fundamental scientific measurement role that Rasch analysis could play — as repeatedly argued on the Rasch list — gradually influenced the style of the book. It's true: for many researchers, the Rasch model provides a very practical solution to data analysis in the social sciences. This volume contends that Rasch measurement is the model of choice because it is the closest to realizing the Holy Grail of human science research: the objective fundamental measurement so long revered in the physical sciences.

Above all, this book should be seen as a tribute to the supportive collegial nature of the Rasch measurement community. A number of our colleagues actively contributed to the ideas the book now contains. William Fisher took on the unenviable, time-consuming, and rather thankless task of being our critical friend. It is rewarding to think that our text will be adopted for a number of introductory courses in measurement. Conversely, those who wish to work through the text independently can have access to the data files that we refer to in the text, and will find those files and further support at: www.jcu.edu.au/~edtgb.

Our aim has been to provide the means by which others could become fully participating members of our international community of the Rasch scholars. Joel Michell's continuing work on the inadequacy of current approaches to psychological measurement (1999) suggests that the time is ripe for Rasch models to have an unprecedented impact. We hope that this volume might help to capitalize on that impetus.

List of Contents

Introduction
Chapters
1. Why Measurement Is Fundamental
2. Important Principles of Measurement Made Explicit
3. Basic Principles of the Rasch Model
4. Building a Set of Items for Measurement
5. Test Equating: Can Two Tests Measure the Same Ability?
6. Measurement Using Likert Scales
7. The Partial Credit Rasch Model
8. Measuring Facets Beyond Ability and Difficulty
9. Revealing Stage-Based Development
10. Rasch Model Applied Across the Human Sciences
11. Rasch Modeling Applied to Rating Scale Design
12. The Question of Model Fit
13. A Synthetic Overview.
Appendix A: Technical Aspects of the Rasch Model
Appendix B: Rasch Modeling Resources
Glossary
References

References

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Erlbaum.

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

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept.* New York: Cambridge University Press.

Notes

1. Further information at: www.erlbaum.com.
2. *Applying the Rasch Model,* by Trevor Bond and Christine Fox, is now available for $29.95 through the IOM website, www.rasch.org/books.htm. IOM members will receive a 10% discount.

From the Classroom

Matthew Enos, Ph.D.

Editor, *Popular Measurement*

More Classic Ben Wright Handouts

In Volume 3 of *Popular Measurement*, we introduced a feature called "From the Classroom." We reprinted several classic handouts from Ben Wright's University of Chicago courses on Questionnaire Design and Measurement. We would like to continue this feature and augment it with new materials.

On the following pages you will find two more Ben Wright handouts. The first is "Four Minds to Life," from Ben's Questionnaire Design course. The second is "Rasch Analysis for Surveys," which he sometimes used as a transition from Questionnaire Design to his courses on Measurement.

"Four Minds to Life" reminds us of Ben's psychological approach to questionnaire design, which he often described as a conversation between researcher and subject(s). In this handout we see his creative use of insights from a broad spectrum of psychological thought, gathered together in a dynamic view of human interaction. Ben's interest in education is also evident in this early handout.

When we first encountered "Rasch Analysis for Surveys," we were struck by its grace and even purity as a statement of the essence of Rasch measurement. For Ben, ever restless to create something new, it was just one of a treasure trove of materials he had used for a while and then set aside. We insisted that he distribute it to his new classes. Now Ben is allowing us to share his classroom materials with you.

Your Turn

Many of you are also college and university teachers, not a few of you students of Ben Wright himself, and are now carrying on his traditions of instruction in science and research. We are sure that you have created many classroom materials that deserve a wider audience. As you return to battle this fall, keep in mind which of your own favorite handouts — one-page posters and brief explanations — we could all benefit from seeing.

Please send us your own short classroom materials and illustrative handouts for future inclusion in "From the Classroom."

Four Minds to Life

Ben Wright

	MIND'S EYE the illusion of "reality"	INDULGENCE comfort pain	AMBITION pride shame	CONSCIENCE virtue guilt
Freud	ego	id	ego-ideal	super-ego
Voices	I	It/Me/You Us, We	Myself Her/Him	Them/They
What to Learn	facts skills	self-care friendship	opportunity courage	obedience responsibility
How to Learn	listen try out	relax join in	invent discover	memorize comply
How to Teach	show/tell exercise	welcome reward	inspire embody	lecture punish
Rostam Kretschmer Type		digestive pyknic	muscular athletic	respiratory asthenic
Sheldon Body Type Personality Tone		FAT endomorph viscerotonic	STRONG mesomorph somatotonic	BRAINY ectomorph cerebrotonic
Erikson Stage		trust mistrust	autonomy shame	initiative guilt
Identification		possess restore	emulate displace	placate endure
School of Education		permissive free play	progressive democratic	traditional authoritarian
Classroom Style		child centered	learning centered	teacher centered
Contact Mode		touch	sight	sound
Message		feels YES	sees YES	hears NO
Best Teacher		friend	hero	master
Worst Teacher		enemy	villain	tyrant

[1985]

Rasch Analysis for Surveys

Ben Wright

Surveys, questionnaires, and interview protocols that use rating scales to collect psychosocial information can be thought of as structured “conversations” between researchers and subjects. To construct a successful questionnaire, the researcher must develop a clear idea of the aim of the questionnaire, especially the inferences that are to be drawn from its use. The researcher must also be intimate with the language the intended subjects understand and use. Observed responses are local descriptions of a situation as perceived by the subject at a moment in time. From these passing responses, the researcher hopes to induce general inferences concerning reproducible processes of enduring psychosocial significance. The desired generalization requires that the observed responses can be fit into an overall metric, a linear variable, along which more-ness and less-ness have well defined quantitative and qualitative meanings. The Rasch Model meets these criteria.

Rasch analysis is a method for constructing from observed counts and categorical responses (like Likert scales) linear systems within which items and subjects can be measured unambiguously. The constructed variables contain the meaning of the structured “conversations.” The measure of a subject on each variable summarizes that subject’s statements about the variable to the extent that the subject shares a definition of the variable with other correspondents. These measures are the most succinct and reproducible report of the information collected by the questionnaire.

Rasch analysis facilitates the transmission of results to subsequent analyses, but now with the advantage of being linear measures with standard errors of the kind required by most statistical analyses. It also simplifies communication of results to therapists, educators, policy makers and the concerned public, in the form of graphical summaries of client populations and detailed individual client profiles.

A unique asset of Rasch analysis is its ability to detect idiosyncrasies — particular, specific departures of subjects and items from the shared understanding that is emerging from the ongoing research. These local departures have powerful diagnostic implications for the treatment of individual subjects. They also suggest new insights into the nature of the proposed variable and new possibilities for improving its definition and measurement.

[1985]

2002 Membership of the Institute for Objective Measurement

Valerie Been Lober, Ph.D.

Executive Manager, IOM

By August 18, 2002, the IOM had 166 members. The IOM would like to acknowledge and thank the following individuals and groups. They have provided generous support to the organization from December 1, 2001 through the present.

Lifetime Members ($5,000 and up)

Jack Stenner
Benjamin Wright

Affiliate (Corporate) Members ($2,500)

American Board of Family Practice

Robert Avant

Affiliate (Professional-Society) Members ($1,000)

North American Spine Society

Tammy Feenstra Banks

Institutional Members ($500)

American Society of Clinical Pathologists

Cindy Brito
Laura Culver-Edgar
Pamela Frommelt
Renata Sekula-Wacura
Donna Surges-Tatum
Kory Ward-Cook

Computer Adaptive Technologies, Inc.

Betty Bergstrom
David Blitz
Deborah Schnipke
John Stahl

Measurement Research Associates, Inc.

James Houston
Patrick Fisher
Mary Lunz
Surintorn Suanthong

National Board of Osteopathic Medical Examiners

Linjun Shen

Sponsoring Members ($300)

Sunhee Chae
Richard Rovinelli

Endorsing Members ($200)

Richard Gershon
Albert Lyons

Introductory IOM Members

University of Illinois at Chicago (contributed $100 for each International Conference on Objective Measurement participant who paid full registration)

Carol Allen
Nancy Bates
Gary Bedell
Svetlana Beltyukova
Suzann Campbell
Shu-Pi Chen
Ryan Deaton
Scott Decker
Daniel Deutscher
Kelly Dineen
Brigid Flood
Christine Fox
William Frey
R. Brian Giesler
Liselotte Hermansson
Gemma Lawton
Ling Liu

David McArthur
Carol Myford
Kyle Perkins
Simon Pickard
Ted Redden
Robert Rush
Karen Rychlik
Noriaki Takepa
Giorgio Vittadini
Elizabeth Zelinski

Regular Members ($80)

Wan Mohd Rani Bin Abdullah
Anne Marie Aish
Rashid Aldosary
Pedro Alvarez Martinez
Ethan Arenson
Judith Babcock
John Barnard
Theresa L.Bender-Pape
Elizabeth Betemps
Nikolaus Bezruczko
Virginia Blankenship
Rita Karwacki Bode
Brian Bontempo
William Boone
Ed Bouchard
Anna Cantagallo
David Cella
Raymond C. K. Chan
Chih-Hung Chang
Christine Chen
Daniel Cipriani
Karen Conrad
Kendon Conrad
Marshall Dahl
Ana Rosa Delgado
Gerald DeMauro
Robert Draba
Arthur Ellen
Marci Enos
Matthew Enos
Howard Everson
William Fisher, Jr.
Robert Florin
Mary Garner
Michael Gilewski
Ines Giorgi
Carl Granger
Elizabeth Hahn
Lisa Hall
Dennis Hart
Allen Heinemann
Gyorgy Horvath
William Elliot Inman
Mark Johnston
Dorthea Juul
Elena Kardanova
Liz Koch
Rense Lange
Ong Kim Lee
John Linacre
Richard Linn
Ross Lober
Marilyn Looney
Peter MacMillan
Trudy Mallinson
Geofferey Masters
H. W. (Bud) Meyers
Mark Moulton
Yuji Nakamura
Jeffrey Nolan
Barry O'Sullivan
Vivian Ott
Marcella Ottonello
Christopher Payne
Gerardo Prieto
Steven Renshaw
Wendy Rheault
Shungwon Ro
Bonnie Roe
Thomas Salzberger
T. Joseph Sheehan
Alan Tennant
Luigi Tesio
Jean Louis Thonnard
William C. Tirre
Agustin Tristan-Lopez
Kazuaki Uekawa

Suzy Vance
Ning Wang
John Ware
Walter Wengel
Louise White
J. Ivan (Jack) Williams
Dennis Wisniewski
Donald Witzke
Frederick Wolfe
Paul Woodward
Ron Zybura

Student Members ($35)

Sarah Ailey
Sarah Austin
Kirk Becker
Barry Brown
Violeta Carrion
Chyi-Kong Karen Chang
Edward Clark
Sharon Foley
Peter Hagell
Roberta Henderson
Jeremy Hobart
Terrence Jackson
Terri Lynne Jackson
Lena Krumlinde-Sundholm
Nantawadee Lee
Edson Matsubayashi
John Mukasa-Ssebaana
David Poradzisz
Eric Paul Rogers
Anna Simone
Cheryl Smithgall
Lana Snider
Donald Straube
Bharat Kumar Vallurupalli
Kwai Ming Wa
Grant Wynn
Futoshi Yumoto
Jason Zedaker

Additional Contributions

University of Illinois at Chicago ($2,500)
American Society of Clinical Pathologists ($494)
Robert Draba ($20)

IOM Chapters

Italian (Italian Study Group on Outcome and Person Measures)
Kansas City Metropolitan (USA)
Korean
Chicago Metropolitan (USA)

In-kind Donors

The IOM would also like to thank its "in-kind" donors. From December 1, 2001 (beginning of IOM's fiscal year) to the present, the following individuals have made gifts to the IOM:

Matthew Enos
- Services as editor of *Popular Measurement*

John Michael Linacre
- Services as webmaster for IOM website

Ross Lober
- Office space for IOM headquarters
- Photography for IOM

Geofferey Masters
- Permission to reprint 500 copies of *Rating Scale Analysis*

Agustin Tristan
- Donation of 10 copies of *Análysis de Rasch para Todos* to the IOM Book Sales Program

Institute for Objective Measurement, Inc.
155 North Harbor Drive, Suite 1002
Chicago, IL 60601
312-616-6705 (voice); 312-616-6704 (fax)
InstObjMeas@worldnet.att.net www. rasch.org

Membership Application

Name (please print) ____

Institution ____

Address Line 1 ____

Address Line 2 ____

City ______________________________ State __________ Zip ____________

Work Telephone ______________________ Fax ___________________________

E-mail ______________________________ @ ______________________________

(*Important for receiving announcements*)

Membership categories (check one) *(Dues are tax-deductible)*

Traditional

□ Student $35 □ Regular $80 □ Endorsing $200

□ Sponsoring $300 □ Institutional $500 □ Lifetime $5,000

Affiliate

□ Professional-society $1,000 □ Corporate $2,500

Total amount of dues: $ ______________

Method of payment

Paid by (circle one): VISA MC CHECK MONEY ORDER

Credit Card Number: □□□□ □□□□ □□□□ □□□□

Credit Card Expires: _____ - _____
Month *Year*

If paying by credit card, you may fax your application to: **312-616-6704**

Signature ________________________________ Date ____________________

(Required, if paying by credit card)