# MANY-FACET

## RASCH MEASUREMENT

John M. Linacre

MESA

# MANY-FACET RASCH MEASUREMENT

**John M. Linacre**

*The University of Chicago*

MESA PRESS

Chicago

1989, 1994

**Many-Facet Rasch Measurement**

John M. Linacre

Second Edition



Printed in the United States of America

Library of Congress Catalog Card Number: 94-76939

ISBN 0-941938-02-6

Institute for Objective Measurement, Inc.
155 North Harbor Drive, Suite 1002
Chicago, IL 60601

Telephone 312-616-6705
Fax 312-616-6704
E-Mail: InstObjMeas@worldnet.att.net

# PREFACE

Many-facet Rasch measurement (MFRM) was born out of necessity. A conscientious test director realized that the conventional method of basing decisions for rated performances directly on raw scores is unfair to examinees who encounter severe raters. It is also potentially life-threatening when incompetent examinees in some crucial discipline are certified merely because they encounter lenient raters. This realization motivated the question: "How can fair and meaningful measures be constructed from inevitably dubious ordinal ratings?" MFRM provides the answer.

From the familiar ordinal raw scores, MFRM constructs linear, objective measures of known precision (standard error) and quality (fit), freed as far as statistically possible, from all the local details of the rating situation. MFRM is based on an unachievable theoretical ideal. It does not attempt to describe any particular set of data. In fact, no empirical data fit the model *exactly*, but MFRM prescribes what characteristics data must have if they are to support the construction of fair and meaningful measures. Though MFRM is expressed as a logit-linear model, it deliberately avoids the interaction terms and arbitrary parameterizations of regression analysis.

MFRM does not model the cognitive processes of the rater. How a rater chose the rating is as irrelevant to MFRM as why you gained weight is to your bathroom scale. MFRM merely requires that the original observations, usually ratings, be ordered qualitatively on the variable, dimension, construct or latent trait of interest. A higher rating must be intended to correspond to "more" of the variable. MFRM provides several diagnostic aids for detecting instances where the data fall short of this intention.

If data don't fit MFRM usefully, then the data are ambiguous. What is useful fit? Unfortunately, statistical significance tests won't help you decide this. All they will do is to tell you that the data aren't perfect. But we know this already. Useful fit means that aberrations in the data don't affect the substantive meaning of the measures. If in doubt about the overall quality of your data, remove the worst part. Are the measures substantively different? If not, you have a useful fit. Of course, the higher the quality of the original ratings, the better the resultant measures.

MFRM is a powerful measurement tool. Researchers in many fields have discovered that it solves their measurement problems. Then better measurement leads to clearer thinking, and clearer thinking leads to improvements in both theory and practice. Many "ceiling" and "floor" effects are discovered to be mere artifacts of ordinal raw scales. Rater training is transformed from inculcating mindless conformity into encouraging intelligent self-consistency. Better measurement even leads to improvement in ordinal scales as test constructors discover which rating scale categorizations work to produce valid measurement.

My profound gratitude goes to Prof. Benjamin D. Wright, whose conviction that order must be imposed on chaos initiated this research; to Dr. Mary Lunz of the Board of Registry of the American Society of Clinical Pathologists, who proved the steel of these concepts in contest with real problems; and to the Spencer Foundation, whose grant of a Fellowship assisted in this endeavor.

John M. Linacre
*The University of Chicago*
*June 30, 1989*

# CONTENTS

# ASSAY-PIECE CONSTRUCTING MEASUREMENT WITH A MANY-FACET MODEL

## A.1 RASCH MEASUREMENT MODEL FOR JUDGING

The construction of linear measures from qualitative observations is a conceptual and statistical advance of recent vintage. Rasch (1960/1980) obtains it for dichotomous responses by examinees to test items. Andrich (1978) and Masters (1982) expand the Rasch model to responses in ordered categories, e.g. attitude surveys and partial credit test items. This previous work has focussed on observations resulting from the interaction of two components or "facets", objects and agents. In practice, a third facet is often encountered: a judge, rater or grader whose task is to award a rating to an examinee based on performance on a test item. The "many-facet Rasch model" extends the Rasch model to situations in which more than two facets interact to produce an observation. It enables the construction of a frame of reference in which quantitative comparisons no longer depend on which examinee happened to be rated by which judge on what item.

### A.1.1 The many-facet Rasch model

A many-facet Rasch model for a typical essay examination, using a rating scale of M+1 ordered response categories with "0" labeling the lowest and "M" labeling the highest, is

$$\log\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = B_n - D_i - C_j - F_k \qquad \text{(A.1)}$$

where

$P_{nijk}$ is the probability of examinee n being awarded on item i by judge j a rating of k

$P_{nijk-1}$ is the probability of examinee n being awarded on item i by judge j a rating of k-1

$B_n$ is the ability of examinee n

$D_i$ is the difficulty of item i

$C_j$ is the severity of judge j

$F_k$ is the difficulty of the step up from category k-1 to category k and k=1,M.

It can be seen that this is similar to Andrich's (1978) model, but with an additional judge severity additional parameter. The probabilistic log-linear formulation of this model is distinctive of Rasch rating scale models, since, unlike many approaches to analyzing ratings, the numerical labels attached to the rating scale categories are not assumed to be interval measures. The Rasch approach makes the weaker assumption that the numerical labels indicate the ordering of the categories along the scale.

The absence of interaction terms common in log-linear models is deliberate, since this model is designed to implement the axioms underlying all Rasch models (Rasch 1960/1980):

a) the contribution of each element of each facet to the observations is dominated by a single parameter with a value independent of all other parameters within the frame of reference. It is this parameter value, e.g. an examinee's reading comprehension, which the testing situation is intended to discover. This single parameterization is necessary if examinees are to be arranged in one order of merit, or items indexed by one kind of difficulty on an item bank, or judges calibrated by one kind of severity in a judge management scheme.

b) the parameters combine additively to produce the observations. Additivity implies that all the parameters share one linear scale. Not only does linearity assist the analyst in understanding the meaning underlying the data, it also provides a useful basis for further analysis, because many commonly used statistical procedures assume linearity in their data.

c) the estimate of any parameter is dependent on the accumulation of all ratings in which it participates, but is independent of the particular values of any of those ratings. This axiom of "local independence" allows the statistical estimates of the measures to be as free as possible of which particular judge rated which particular examinee on which particular item, and so to have meaning generalizable beyond the local details of the judging situation.

This model is prescriptive rather than descriptive. The model implements a set of linear measurement intentions on the part of the analyst, so misfit between data and model does not indicate failure on the part of the model. Rather it indicates that the data do not support the construction of interval measures. In the practical application of Rasch models, global fit is misleading, because localized misfit within the data set is just as threatening as global misfit.
Consequently, fit statistics are obtained for each parameter, and the success of the analysis is determined from an inspection of these.

Depending on the nature of the misfit and the motivation for the analysis, the analyst's intentions can be revised, so that, say, each judge is specified to use the rating scale in a different way. Moreover, the data can also be revised to remove or amend aberrant observations such as guesses or ratings by idiosyncratic judges. Though these techniques can improve the fit of the data to the model, they lessen the generality of the parameter values estimated.

With this in mind, the many-facet measurement model can be expressed in many forms to meet the requirements of specific testing situations. Other forms include:

A judge-scale model, in which each judge uses his own interpretation of the rating scale:

(A.2) $$\log(\frac{P_{nijk}}{P_{nijk-1}}) = B_n - D_i - C_j - F_{jk}$$

where

$F_{jk}$ is the difficulty of the step from category k-1 to category k for **judge j**, and k=1,$M_j$

and other parameters are defined as before.

An item-scale model, in which each item is constructed with its own rating scale:

(A.3) $$\log(\frac{P_{nijk}}{P_{nijk-1}}) = B_n - D_i - C_j - F_{ik}$$

where

$F_{ik}$ is the difficulty of the step from category k-1 to category k of the scale unique to **item i**, and k=1,$M_i$

and other parameters are defined as before.

A four-facet model, in which each of the items is modelled to apply to each of a number of tasks:

(A.4) $$\log(\frac{P_{nmijk}}{P_{nmijk-1}}) = B_n - A_m - D_i - C_j - F_k$$

where

$A_m$ is the difficulty of task m

and other parameters are defined as before.

## A.2 THE NON-LINEAR RATING SCALE

Rating scale categories are often labelled with integer values, causing the non-linearity of the ratings to be overlooked. As fig. A.1 illustrates, ratings originate on an ordinal, not an interval, scale. Rating categories are conceptualized as representing qualitatively distinct, but ordered, performance levels. The scale is printed with equal spacing in order to invite raters to devote equal attention to each of the alternatives. The range of performance corresponding to each of the ordered categories can only be discovered empirically by how raters behave. The highest and lowest categories always represent infinite ranges of performance above and below the intermediate categories. The ranges of performance represented by intermediate categories depend on how their definitions are interpreted by the judges. Conventional analysis mistakes the scale designer's choice of numerical labels for measures and so ignores the inevitably non-linear functioning of the scale, but it is how the scale functions, not how it is presented, that produces the measures.

Equation A.1 specifies the intended relationship between the ordered categories of a rating scale and the latent performance continuum they imply. This is a logistic ogive which satisfies both the axiomatic requirements for measurement and the form of the rating scale

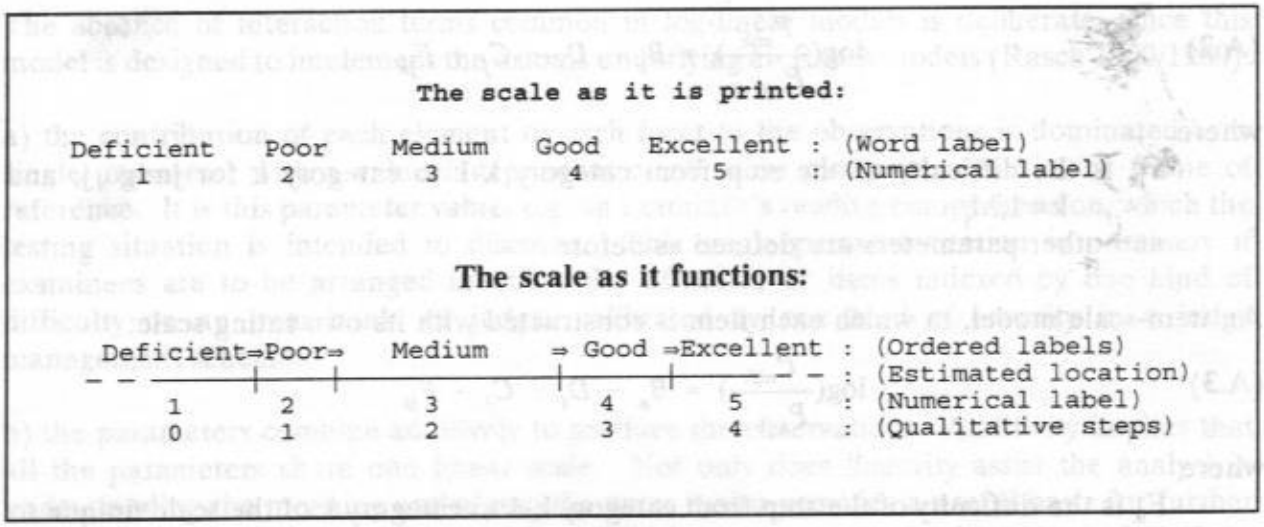


Fig. A.1. Presentation and measurement perspectives on a representative rating scale.

defined by the way the judges use it. The unequal widths of the performance ranges corresponding to the intermediate categories are parameterized and estimated by the $F_k$ terms. The infinite performance ranges at the extremes of the scale are mapped into the corresponding finite top and bottom response categories. Fig. A.2 illustrates the relationship between the average rating awarded each examinee and the examinee's measure based on simulated data. The solid ogive traces the raw score to measure conversion when all judges rate all examinees on all essays. Each X represents a conversion when only some of the judges rate an examinee's performance. Examinee A has a higher measure, but lower average rating label than Examinee B, because A happened to be rated by more severe judges than B.

## A.3 PARAMETER ESTIMATION AND MISSING DATA

The marginal raw score is a sufficient statistic for each parameter. Consequently, a maximum likelihood estimate for each parameter is obtained when the expected marginal score for observations in which the parameter participates is set equal to the observed score. A modelled asymptotic standard error, the estimate's reliability, can also be estimated. The degree to which the parameter estimate is valid is measured by statistics quantifying the fit of the data to the measurement model (Wright & Masters, 1982). This is implemented in the *Facets* computer program (Linacre, 1988).

A valuable by-product of this method of estimation is that there is considerable latitude for missing data. Each parameter is estimated only from the subset of observations in which it participates directly. Accordingly a unique finite estimate for each parameter is obtained provided that

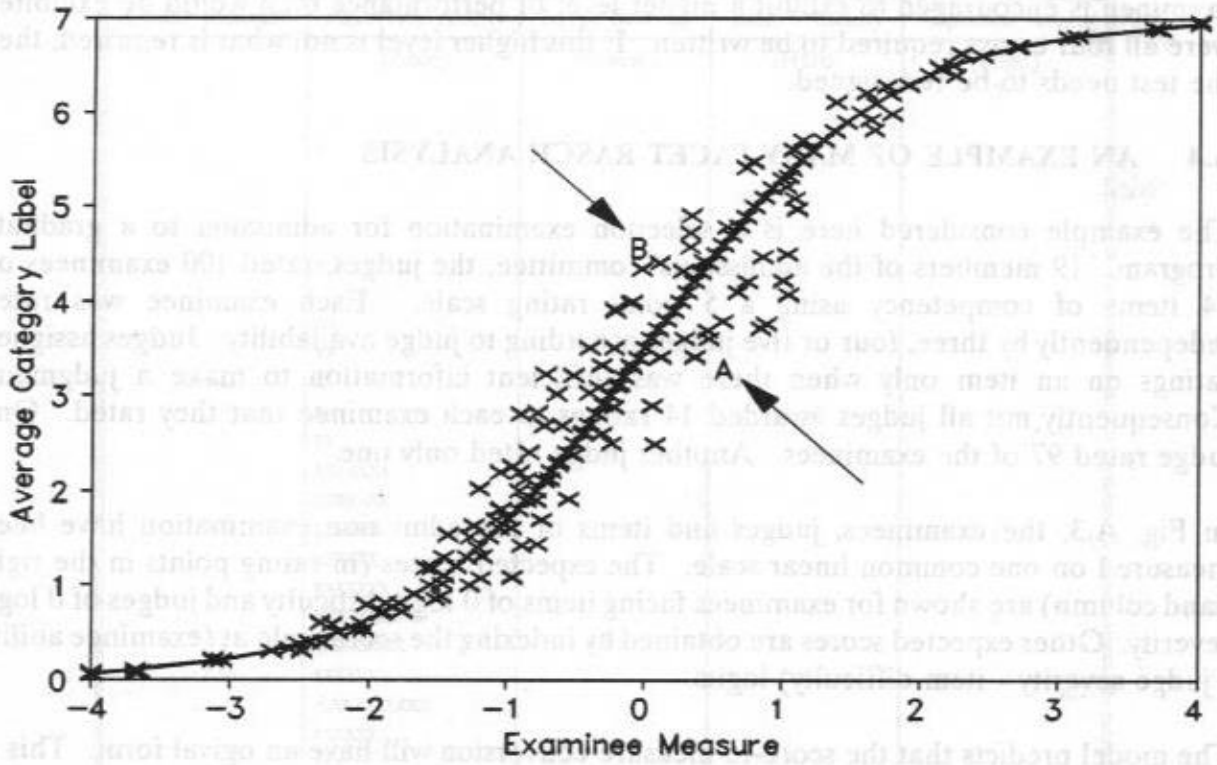


Fig. A.2 Average category labels for examinee performance plotted against estimated logit measures for simulated data.

1) all observations in which the parameter participates are not in the same extreme category. This would require the estimate to be infinite, corresponding to a minimum or maximum possible score. In the FACETS program, infinite estimates can be replaced by estimates corresponding to scores one-third of a score point away from the extreme scores.

2) the observations form a linked network such that every parameter can be estimated unambiguously within the same frame of reference.

An example of ambiguity occurs when Judge B rates all the boys, while Judge G rates all the girls. Then it is unclear whether better performance by the boys is due to their higher ability or Judge B's more lenient judging style. If some children are rated by both judges, this ambiguity would be removed. If the design of a test is such that the examinees are allowed to select which two out of four essays they wish to write, then linkage across essays becomes inevitable as examinees choose different pairs of essays. There is no requirement for statistical randomness in the assignment of judges or selection of essays.

There is no need to impute values to missing observations. In the example of the selection of two essays from four, there is no need to attempt to determine how the examinee would have performed on the two unwritten essays. The structure of the test is such that the

examinee is encouraged to exhibit a higher level of performance than would be exhibited were all four essays required to be written. If this higher level is not what is required, then the test needs to be redesigned.

## A.4 AN EXAMPLE OF MANY-FACET RASCH ANALYSIS

The example considered here is a selection examination for admission to a graduate program. 19 members of the admissions' committee, the judges, rated 100 examinees on 14 items of competency using a 5 point rating scale. Each examinee was rated independently by three, four or five judges according to judge availability. Judges assigned ratings on an item only when there was sufficient information to make a judgment. Consequently not all judges awarded 14 ratings to each examinee that they rated. One judge rated 97 of the examinees. Another judge rated only one.

In Fig. A.3, the examinees, judges and items of the admission examination have been measured on one common linear scale. The expected scores (in rating points in the right hand column) are shown for examinees facing items of 0 logit difficulty and judges of 0 logit severity. Other expected scores are obtained by indexing the score scale at (examinee ability - judge severity - item difficulty) logits.

The model predicts that the score-to-measure conversion will have an ogival form. This is shown in Fig. A.4, where the average rating given an examinee on the admissions test has been mapped against examinee measure. The solid ogive traces the raw score to measure conversion that would have occurred if all judges had rated all examinees on all items. Each point X represents the conversion for an examinee. Its placement depends on which judges rated the examinee's performance. Examinee A has a higher average rating, but a lower measure than Examinee B, because A happened to be rated by more lenient judges than B. Most X's are displaced below the solid ogive because the most lenient judge rated only a few examinees.

## A.5 QUANTIFYING JUDGE UNRELIABILITY

Even the most diligent judge training has failed to produce uniformity among judges (Borman, 1978), but any difference among judges threatens fairness because the examinee raw score depends on which judge awards the rating. Indeed sometimes "there is as much variation among judges as to the value of each paper as there is variation among papers in the estimation of each judge" (Ruggles, 1911). It was this lack of judge reliability that was identified as a chief drawback to judge-dependent tests (Ruch, 1929). Clearly, since differences in judge severity can account for as much ratings variance as differences in examinee ability (Cason & Cason, 1984), objective measurement requires that judge behavior be modelled and statistically controlled.

The Rasch approach recognizes and models two aspects of judge behavior. First, judges are modelled to differ in severity or leniency. This can account for half the variance in judge

| Linear Measure | Examinee Ability | Judge Severity | Item Difficulty | Expected Rating |
|---|---|---|---|---|
| | (Able) | (Severe) | (Hard) | (High) |
| 6.0 | | | | 5 |
| | x | | | |
| 5.0 | | | | |
| | x | | | |
| | xx | | | |
| | xx | | | |
| 4.0 | x | | | 4.5 |
| | x | | | |
| | xx | | | |
| | xxxxxx | | | |
| | xxxxxx | | | |
| 3.0 | xxxxx | | | |
| | xxxxx | | | |
| | xxxxxx | | | |
| | xxxxxxxxx | | | 4 |
| | Bxxxxxx | | | |
| 2.0 | xxxxxx | | | |
| | Axxxxxxxxx | | | |
| | xxxxx | | | |
| | xxx | | | |
| | xxxx | | x | |
| 1.0 | xxxx | x | | 3.5 |
| | xxx | Nx | xx | |
| | | PQMxxx | | |
| | x | xx | xx | |
| | x | x | | |
| 0.0 | | x | x | |
| | | x | x | |
| | | | xxx | |
| | | | xxxx | 3 |
| | | x | | |
| -1.0 | | x | | |
| | | xx | | |
| | | | | 2.5 |
| -2.0 | | | | |
| | | | | 2 |
| | | x | | |
| -3.0 | | | | 1.5 |
| | (Less Able) | (Lenient) | (Easy) | (Low) |
| Linear Measure | Examinee Ability | Judge Severity | Item Difficulty | Expected Rating |

Fig. A.3 Results of a many-facet Rasch analysis of the Admissions data.

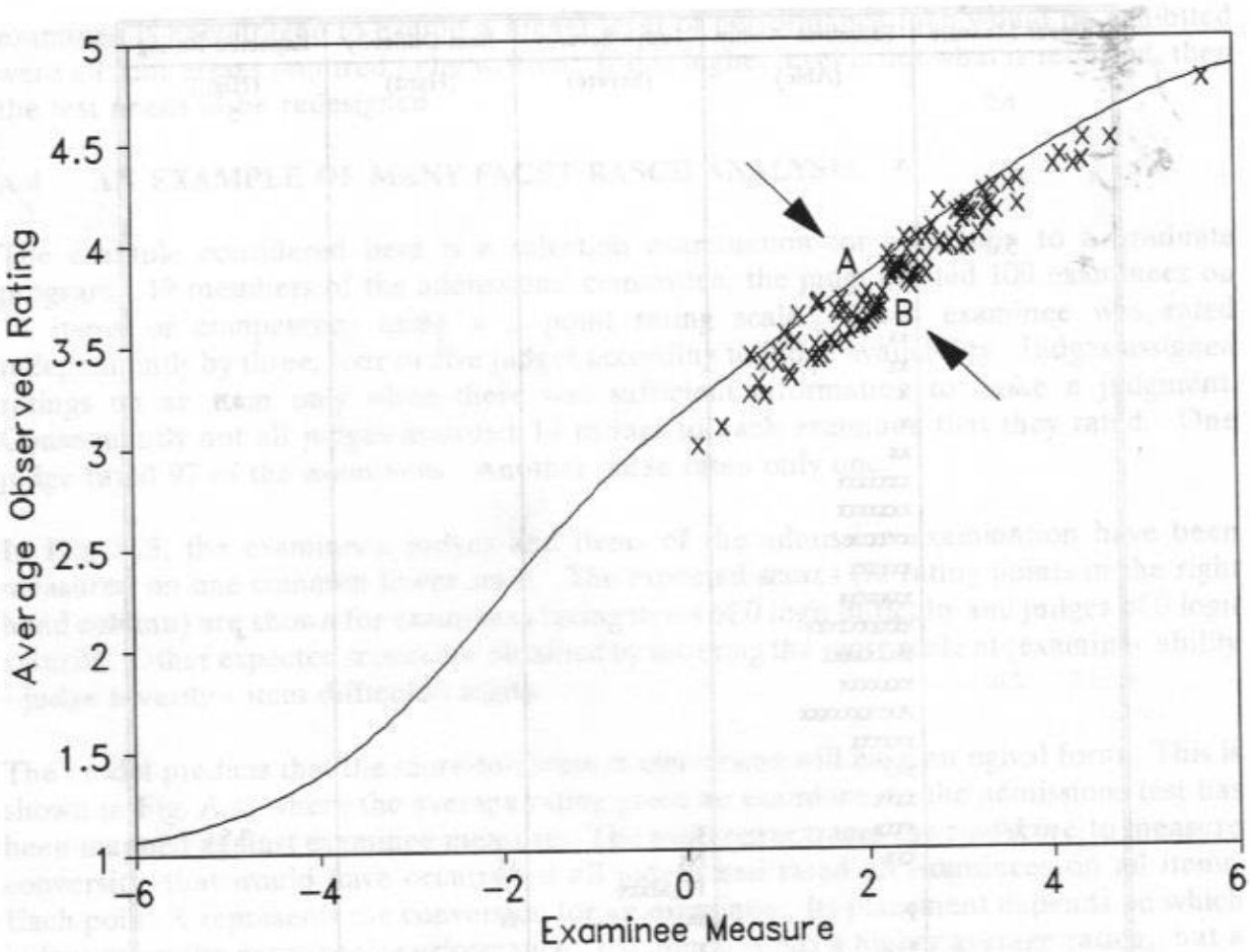


Fig. A.4 Average awarded ratings for examinee performance plotted against estimated logit measures for the Admissions data.

ratings (Edgeworth, 1890). When the dominant judging behavior of a judge is captured in one severity parameter, then that parameter characterizes the judge in exactly the same way as an ability parameter characterizes an examinee and a difficulty parameter characterizes an item.

Second, judges are modelled to exhibit some degree of stochastic behavior when awarding ratings. Every observation is modelled to contain error. Both too much and too little error in the observations contradict the measurement model and are a threat to the validity of the measurement process.

From the measurement model presented in Equation A.1 is derived the model for an observed rating value:

(A.5) $$X_{nij} = E_{nij} \pm S_{nij}$$

where

$X_{nij}$ = the observed rating and the expected rating value is

(A.6) $$E_{nij} = \sum_{k=0}^{M} kP_{nijk}$$

and the modelled "error" variance of the observed rating around its expected value is

(A.7) $$(S_{nij})^2 = \sum_{k=0}^{M} (k-E_{nij})^2 \, P_{nijk}$$

Compare this with the conventional model for ratings, in which the rating category labels are treated as though they were linear measures (c.f. Braun, 1988):

(A.8) $$X_{nij} = E'_{nij} \pm S'_j$$

where

$E'_{nij}$ is the expected score obtained from the sum of the mean values of the elements which generated the rating.

$(S'_j)^2$ is an unexplained error variance.

Even beyond the awkward use of ordinal, non-linear, and locally-defined rating labels as though they were actual linear measures, there are important differences between $S_{nij}$ and $S'_j$. In the Rasch equation, Equation A.5, $(S_{nij})^2$ models the expected error variance for a particular observation. The estimation procedure constructs a linear scale on which the parameter estimates meet the condition that for each parameter, the observed and expected marginal scores coincide. Once this condition is met, the modelled (i.e. expected) values of $S_{nij}$ are also known. The two right-most terms of Equation A.5 are thus obtained explicitly from the model equation, Equation A.1, and the parameter estimates. Error variance is not attributed directly to the judges, but is a consequence of the stochastic nature of the measurement process.

In contrast, the size of $S'_j$ has no antecedent in the model and can only be determined empirically. It is usually thought of as undesired judge-dependent "error" variance. The ideal value of zero can never be observed in a non-trivial situation. Any amount greater than zero threatens validity.

### A.5.1 Fit to the Model

Equations A.1 and A.5 specify the stochastic structure of the data. The modelled, i.e. expected, values of the error variance associated with each rating are explicit. This enables a detailed examination of the data for fit to the model. Not only too much, but also too little, observed "error" variance threatens the validity of the measurement process, and

motivates investigation, diagnosis and remediation of specific measurement problems. The relationships between the modelled "error" variances and the observed "error" variances (sums of squared residuals) are used as partial and global tests of fit of data to model (Wright & Panchapakesan, 1969; Windmeijer, 1990).

TABLE A.1
JUDGE MEASURES AND FIT STATISTICS FOR THE ADMISSIONS DATA

| Judge | Examinees Rated | Total Ratings | % Frequency of Rating | | | | | Mean Rating | Severity Measure | Model Error | Fit Statistics | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | | | | Mean-Square | Standardized (Approximate) |
| P | 12 | 168 | 0 | 0 | 35 | 53 | 11 | 2.8 | 0.62 | 0.13 | 1.0 | 0 |
| Q | 48 | 672 | 0 | 1 | 42 | 42 | 15 | 2.7 | 0.68 | 0.07 | 1.1 | 1 |
| N (Noisy) | 17 | 231 | 0 | 6 | 35 | 41 | 18 | 2.7 | 0.81 | 0.11 | 1.4 | 3 |
| M (Muted) | 73 | 1018 | 0 | 0 | 31 | 61 | 7 | 2.8 | 0.63 | 0.05 | 0.7 | -6 |

An example of Rasch fit statistics for four of the admission examination judges is shown in Table A.1. Their severity measures (in log-odds units, logits) are about equal, but their measures have different modelled standard errors. These indicate the precision or reliability of their measures. The size of these errors is chiefly determined by the number of ratings the judge made. The more ratings a judge makes, the more information there is with which to estimate a severity measure and so the smaller its standard error. A mean-square fit statistic, based on the ratio of observed error variance to modelled error variance, is also reported. It is chi-square distributed on a ratio scale with expectation 1 and range 0 to infinity. Acceptable values are in the range 0.8 to 1.2. A standardized value of the mean-square value is also show to indicate the significance levels of these statistics. Since the success of the standardization depends on the local distribution of the observations, the standardized values do not support strict interpretation.

In Table A.1, Judges P and Q have mean-square outfit statistics close to their expected values of 1. Judge N, however, shows considerable misfit. His mean-square outfit of 1.4 indicates 40% more variance in his ratings than is modelled. Symptomatic of Judge N's behavior is the distribution of his ratings. He awarded considerably more high and more low ratings than Judges P and Q. This wider spread of ratings is unexpected in the light of the rating patterns of the other judges. Judge M, on the other hand exhibits a muted rating pattern. His mean-square statistic of .7 indicates 30% less variance in his ratings than is modelled. Judge M's ratings show a preference for central categories. He reduces the rating scale to a dichotomy and so reduces the variance of his ratings. The fact that Judge M's ratings are more predictable than those of the other raters might be regarded as beneficial in a conventional analysis. In a Rasch analysis, however, Judge M's predictability implies that Judge M is not supplying as much independent information as the other judges on which to base the examinees' measures. Were Judge M perfectly predictable, e.g., always rating in the same category, he would supply no information concerning differences among examinees.

| Judge | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Essay | ABC | ABC | ABC | ABC | ABC | ABC | ABC | ABC | ABC | ABC | ABC | ABC |
| Person | | | | | | | | | | | | |
| 1 | 553 | 686 | 877 | 687 | 777 | 685 | 565 | 667 | 586 | 567 | 776 | 696 |
| 2 | 454 | 542 | 445 | 534 | 334 | 344 | 433 | 526 | 444 | 445 | 533 | 534 |
| 3 | 434 | 544 | 343 | 555 | 433 | 544 | 563 | 443 | 554 | 454 | 443 | 343 |
| 4 | 345 | 426 | 232 | 545 | 445 | 225 | 464 | 456 | 642 | 446 | 445 | 335 |
| 5 | 443 | 548 | 656 | 545 | 657 | 448 | 558 | 466 | 464 | 448 | 547 | 348 |
| 6 | 544 | 846 | 843 | 565 | 633 | 367 | 788 | 673 | 666 | 566 | 564 | 454 |
| 7 | 545 | 665 | 454 | 667 | 755 | 646 | 773 | 785 | 874 | 565 | 745 | 447 |
| 8 | 553 | 763 | 655 | 675 | 775 | 653 | 773 | 656 | 784 | 576 | 573 | 574 |
| 9 | 343 | 643 | 643 | 645 | 534 | 523 | 665 | 674 | 753 | 546 | 545 | 765 |
| 10 | 564 | 766 | 884 | 776 | 655 | 667 | 875 | 778 | 778 | 667 | 649 | 888 |
| 11 | 535 | 524 | 537 | 544 | 545 | 435 | 546 | 557 | 326 | 446 | 456 | 334 |
| 12 | 436 | 644 | 444 | 546 | 666 | 555 | 574 | 445 | 745 | 356 | 763 | 676 |
| 13 | 445 | 486 | 657 | 566 | 246 | 366 | 368 | 448 | 467 | 348 | 569 | 349 |
| 14 | 446 | 533 | 333 | 344 | 545 | 343 | 463 | 353 | 354 | 346 | 462 | 363 |
| 15 | 548 | 855 | 743 | 746 | 766 | 656 | 665 | 765 | 854 | 666 | 862 | 844 |
| 16 | 644 | 653 | 547 | 545 | 643 | 454 | 556 | 467 | 666 | 447 | 558 | 667 |
| 17 | 414 | 817 | 625 | 628 | 536 | 518 | 425 | 618 | 717 | 627 | 639 | 436 |
| 18 | 334 | 655 | 443 | 445 | 243 | 473 | 445 | 747 | 654 | 445 | 435 | 334 |
| 19 | 747 | 745 | 837 | 756 | 755 | 847 | 664 | 688 | 737 | 656 | 847 | 938 |
| 20 | 443 | 666 | 735 | 556 | 557 | 557 | 588 | 667 | 666 | 557 | 476 | 488 |
| 21 | 242 | 443 | 336 | 465 | 245 | 243 | 263 | 245 | 441 | 253 | 342 | 254 |
| 22 | 564 | 765 | 747 | 666 | 864 | 577 | 667 | 576 | 667 | 557 | 667 | 785 |
| 23 | 446 | 566 | 753 | 646 | 444 | 565 | 475 | 388 | 576 | 557 | 557 | 776 |
| 24 | 332 | 422 | 334 | 433 | 322 | 214 | 423 | 223 | 323 | 313 | 233 | 223 |
| 25 | 543 | 664 | 544 | 657 | 646 | 544 | 454 | 448 | 547 | 545 | 456 | 464 |
| 26 | 644 | 764 | 955 | 756 | 545 | 658 | 655 | 867 | 776 | 646 | 756 | 885 |
| 27 | 342 | 346 | 334 | 344 | 346 | 234 | 256 | 256 | 345 | 345 | 256 | 253 |
| 28 | 343 | 463 | 335 | 334 | 465 | 573 | 341 | 475 | 442 | 243 | 462 | 272 |
| 29 | 433 | 444 | 323 | 446 | 334 | 333 | 235 | 336 | 423 | 336 | 323 | 343 |
| 30 | 542 | 564 | 244 | 655 | 445 | 224 | 546 | 575 | 645 | 446 | 432 | 555 |
| 31 | 325 | 514 | 313 | 425 | 315 | 314 | 334 | 225 | 525 | 314 | 324 | 314 |
| 32 | 644 | 744 | 445 | 545 | 533 | 553 | 567 | 584 | 664 | 447 | 556 | 364 |

Fig. A.5 "Complete" judging plan for the Essay data (Courtesy: Robert G. Cameron of the College Board).

In this example, the logistic ogive approximates linearity across the restricted range of the observed ratings, as fig. A.4 indicates. Accordingly, modelling the ratings themselves as interval measures would produce about the same model-data fit as the Rasch approach. This, however, is irrelevant, because the primary intent of Rasch analysis is not to maximize model-data fit, but to construct generalizable linear measures, each of known standard error (reliability) and validity (fit), which supersede the locally-defined structure of the rating scale. Incidentally, without the benefit of the Rasch analysis, the approximation to linearity would have been an assumption, now it is an established property of the data.

### A.5.2 Control of judge behavior

It is always essential to monitor the quality of the ratings being awarded and to direct each judge's attention to those areas in which there is doubt. Conventional analysis has placed

| Linear Measure | Examinee Ability | Judge Severity | Essay Difficulty | Expected Rating |
|---|---|---|---|---|
| | (Able) | (Severe) | (Hard) | (High) |
| 1.0 | | | | |
| | 10 | | | 6.5 |
| | 1 | | | |
| | 19 22 26 | | | 6 |
| | 20 | 1 | | |
| | 15 7 | | | |
| | 13 23 8 | | | 5.5 |
| | 16 5 6 | 3 6 | | |
| | 12 | 5 | B | 5 |
| 0.0 | 25 32 | 10 11 | A | |
| | 17 9 | 12 2 7 9 | C | |
| | | 4 | | 4.5 |
| | 11 30 | 8 | | |
| | 18 | | | 4 |
| | 3 4 | | | |
| | 14 2 28 | | | |
| | 27 | | | 3.5 |
| | 21 29 | | | |
| -1.0 | | | | |
| | 31 | | | 3 |
| | | | | |
| | | | | |
| | 24 | | | 2.5 |
| | | | | |
| -2.0 | (Less Able) | (Lenient) | (Easy) | (Low) |
| Linear Measure | Examinee Ability | Judge Severity | Essay Difficulty | Expected Rating |

Fig. A.6 Results of a many-facet Rasch analysis of the Essay data.

a premium on unanimity across judges concerning the rating to be awarded to each performance on each item. An advantage of the Rasch model is that what is decisive is not the numeric values of the ratings, but what they imply. Consequently judge self-consistency, rather than judge unanimity, is the aim. The effective judge is one who maintains the same severity level and shares the common understanding of the rating scale. Then, if the difference between the observed ratings and their modelled values is only inevitable measurement error, it is irrelevant whether any particular rater or group of raters agreed with any other rater or group of raters as to the precise ratings awarded.

Inspection of the residual differences between the observed and expected ratings enables unexpectedly harsh or lenient ratings to be identified. Further, each judge's bias relating to any particular items, groups of examinees, or the like, can be detected. This has two

| Judge | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Essay | ABC | ABC | ABC | ABC | ABC | ABC | ABC | ABC | ABC | ABC | ABC | ABC |
| Person | | | | | | | | | | | | |
| 1 | 553 | 686 | | | | | | | | | | |
| 2 | | 542 | 445 | | | | | | | | | |
| 3 | | | 343 | 555 | | | | | | | | |
| 4 | | | | 545 | 445 | | | | | | | |
| 5 | | | | | 657 | 448 | | | | | | |
| 6 | | | | | | 367 | 788 | | | | | |
| 7 | | | | | | | 773 | 785 | | | | |
| 8 | | | | | | | | 656 | 784 | | | |
| 9 | | | | | | | | | 753 | 546 | | |
| 10 | | | | | | | | | | 667 | 649 | |
| 11 | | | | | | | | | | | 456 | 334 |
| 12 | 436 | | | | | | | | | | | 676 |
| 13 | 445 | | | | | | 368 | | | | | |
| 14 | | 533 | | | | | | 353 | | | | |
| 15 | | | 743 | | | | | | 854 | | | |
| 16 | | | | 545 | | | | | | 447 | | |
| 17 | | | | | 536 | | | | | | 639 | |
| 18 | | | | | | 473 | | | | | | 334 |
| 19 | 747 | | | 756 | | | | | | | | |
| 20 | | 666 | | | 557 | | | | | | | |
| 21 | | | 336 | | | 243 | | | | | | |
| 22 | | | | 666 | | | 667 | | | | | |
| 23 | | | | | 444 | | | 388 | | | | |
| 24 | | | | | | 214 | | | 323 | | | |
| 25 | | | | | | | 454 | | | 545 | | |
| 26 | | | | | | | | 867 | | | 756 | |
| 27 | | | | | | | | | 345 | | | 253 |
| 28 | 343 | | | | | | | | | 243 | | |
| 29 | | 444 | | | | | | | | | 323 | |
| 30 | | | 244 | | | | | | | | | 555 |
| 31 | rating performed by any available judges | | | | | | | | | | | |
| 32 | rating performed by any available judges | | | | | | | | | | | |

Fig. A.7 "Rotating test-book" judging plan: lower judge load, convenient to administer.

benefits. First, unacceptably idiosyncratic ratings can be intercepted and, if necessary, treated as "missing" without disturbing the validity of the remainder of the analysis. Second, precise feedback to each judge about specific questionable ratings and rating patterns can be used to foster improvements in the judging process.

## A.6 THE JUDGING PLAN

The only requirement on the judging plan is that there be enough linkage between all elements of all facets that all parameters can be estimated without indeterminacy within one frame of reference. Fig. A.5 illustrates an ideal judging plan for both conventional and Rasch analysis. The 1152 ratings shown are a set of essay ratings from the Advanced Placement Program of the College Board. These are also discussed in Braun (1988). This judging plan meets the linkage requirement because every element can be compared directly

| Judge | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Essay | ABC | ABC | ABC | ABC | ABC | ABC | ABC | ABC | ABC | ABC | ABC | ABC |
| Person | | | | | | | | | | | | |
| 1 | | | | | | | 6 | 7 | 5 | | | |
| 2 | | | | 5 | 3 | | | | | 5 | | |
| 3 | 4 | | | | | | | | | 5 | 3 | |
| 4 | | 2 | | | | 5 | | | | 4 | | |
| 5 | | | | | 7 | | | | | 4 | 5 | |
| 6 | 5 | 6 | | | | | | | | | 6 | |
| 7 | | | | | 5 | | 3 | | | | | 4 |
| 8 | | 6 | | | | | | 6 | | 5 | | |
| 9 | | | | | 3 | | 6 | 4 | | | | |
| 10 | | | | 7 | 5 | | | | 7 | | | |
| 11 | | | | | | | | | 6 | | 4 | 3 |
| 12 | 4 | | | 6 | | | | | | | | 7 |
| 13 | 4 | | | 6 | | | | | 4 | | | |
| 14 | 6 | 3 | | | | | 4 | | | | | |
| 15 | 4 | | | | | | | | 4 | 6 | | |
| 16 | | 6 | 4 | | | | | | | | 8 | |
| 17 | | | 2 | 6 | 6 | | | | | | | |
| 18 | | | | | | 4 | 4 | | | | | 4 |
| 19 | | 7 | | 6 | | 4 | | | | | | |
| 20 | | | 7 | 5 | | | | | 6 | | | |
| 21 | | | | | | 2 | 6 | | | 3 | | |
| 22 | | | 7 | | 8 | | | | 6 | | | |
| 23 | | 6 | 7 | | | | | 8 | | | | |
| 24 | | | | | | 2 | 3 | | | | | 2 |
| 25 | | | | | | 4 | 4 | | | | 4 | |
| 26 | | | | | | | | | 7 | 6 | | 8 |
| 27 | | | 3 | | | | | 2 | | | 6 | |
| 28 | 4 | 3 | | | | | | | | | | 2 |
| 29 | | | 3 | | | | | 3 | | | | 3 |
| 30 | 2 | | | | | | | 5 | | | 3 | |
| 31 | | | | 2 | | 4 | | 2 | | | | |
| 32 | | | 5 | | 5 | 5 | | | | | | |

Fig. A.8 Minimal-effort judging plan. Each performance component rated once. Simple to administer.

and unambiguously with every other element. Thus it provides precise and accurate measures of all parameters in a shared frame of reference, as shown in fig. A.6.

Less data intensive, but also less precise, Rasch estimates can be obtained so long as overlap is maintained. Fig. A.7 illustrates such a reduced network of observations which still connects examinees, judges and items. The parameters are linked into one frame of reference through 180 ratings which share pairs of parameters (common essays, common examinees or common judges). Accidental omissions or unintended ratings would alter the judging plan, but would not threaten the analysis. Measures are less precise than with complete data because 83% less observations are made.

Judging is time-consuming and expensive. Under extreme circumstances, judging plans can be devised so that each performance is judged only once. Even then the statistical

requirement for overlap can usually be met rather easily. Fig. A.8 is a simulation of such a minimal judging plan. Each of the 32 examinees' three essays is rated by only one judge. Each of the 12 judges rates 8 essays, including 2 or 3 of each essay type. Nevertheless the examinee-judge-essay overlap of these 96 ratings enables all parameters to be estimated unambiguously in one frame of reference. The constraints used in the assignment of essays to judges were that (1) each essay be rated only once; (2) each judge rate an examinee once at most; and (3) each judge avoid rating any one type of essay too frequently. The statistical cost of this minimal data collection is low measurement precision, but this plan requires only 96 ratings, 8% of the data in fig. A.5. A practical refinement of this minimal plan would allow each judge to work at his own pace until all essays were graded, so that faster judges would rate more essays. A minimal judging plan of this type has been successfully implemented (Lunz et al., 1990).

# 1 THE ROLE OF JUDGEMENT IN OBTAINING KNOWLEDGE

## 1.1 OBJECTIVE KNOWLEDGE AND SUBJECTIVE JUDGEMENT: A CONVERSATION BETWEEN PHYSICAL SCIENCE AND PSYCHOMETRICS

Obtaining measures from data is much the same in "Hard" physical science and "Soft" psychometrics, as the following hypothetical conversation illustrates.

**Physical Science:**
A printed table of physical constants gives the illusion of high precision, pin-point accuracy and absolute truth. Yet, in a later edition of that same table every one of those numbers may have changed (Browne 1987). Some may have altered due to a redefinition of what they signify, but the predominant reason for amending the values of physical constants is that better, but still imperfect, values have been obtained.

*Published definitive values, with their accompanying limits of uncertainty, are not experimental data, but merely the author's inferences from such data. Inferences are always subject to question; they may be criticized, reexamined, and revised at any time (Dorsey 1944, 3).*

**Psychometrics:**
A list of examinees' test scores gives the illusion of pin-point accuracy and absolute truth. If another test of the same material is given, the scores will be different. Scores are often criticized.

**Physical Science:**
The meaning of a physical constant, and even its existence, is determined by the manner in which the Universe is modelled.

*It therefore seems inevitable that we should speak in terms of some definite theoretical model of the world of experience. There appears, however, to be no meaning in supposing there to exist a unique final model that we are trying to discover. We construct a model, we do not discover it.*

*Is it nevertheless true to say that models lead to the discovery of constants in physics? The answer seems to be yes, in the sense that a model may suggest a set of operations that is found to lead to a "determination" of some constant. If so, then ever more refined repeatable observations are found to lead to a more and more "accurate" value (McCrea 1983, 211).*

**Psychometrics:**
To talk of an individual's "ability" implies that a theoretical model of that person's behavior can be constructed in which the parameter corresponding to the term "ability" has some constant value, however transitory. This model is, by no means, an attempt to construct the unique final model which captures the individual's entire behavior.

**Physical Science:**
The methodology for obtaining the value of a physical constant is straight-forward in principle.

*The particular value yielded by a given apparatus, procedure, and observer is of no interest in itself, but only in connection with such a study as will enable one to say with some certainty that the value so found does not depart from the [actual value] by more than a certain stated amount. No investigation can establish a unique value for the [actual value], but merely a range of values centered upon a unique value (Dorsey 1944, 9).*

**Psychometrics:**
The items which make up a particular test and the score made on the test are of no value themselves but only as they provide information relating to an ability defined in a greater context. The purpose of the test is to quantify that ability. No test can provide an exact determination of the ability, but only an estimate of it.

**Physical Science:**
Measurement is a comprehensive term. It is both a process and its outcome.

*Measurement is the assignment of numbers to material things to represent the relations existing among them with respect to particular properties. The number assigned to some particular property serves to represent the relative amount of this property associated with the object concerned. In practice the assignment of numerical magnitudes to a particular property of a thing is ordinarily accomplished by comparison with a set of standards, or by comparison either of the quantity itself [i.e. the thing], or of some transform of it [i.e. some effect of the thing], with a previously calibrated scale (Eisenhart 1969, 23).*

*A measurement process is essentially a production process, the "product" being numbers, that is the measurements. A characteristic of a measurement process is that repeated measurements of the same thing result in a series of non-identical numbers (Pontius and Cameron 1969, 13).*

**Psychometrics:**
Educational measurement is a process involving the generation of numbers by counting the number of right answers, which are essentially comparisons between the ability of the examinee and the performance standard implicit in the difficulty of each item. An examinee responding to numerous items of identical difficulty is not expected either to succeed or to fail on all of them.

The examinee's performance may also be assessed on a rating scale, calibrated in terms of qualitative levels of performance. This assessment may either be obtained directly from the examinee's response to the item, or be made by a judge based on his appraisal of the examinee's performance. Different judges can be expected to award different ratings.

**Physical Science:**
The subjective feeling of the observer is crucial in determining the outcome of an experiment. The experimenter will perform what he feels to be a well-conducted experiment.

*Thus the experimenter presently will feel justified in saying that he feels, or believes, or is of the opinion, that his own work indicates that the [actual value] does not depart form his own definitive value by more than so-and-so, meaning thereby, since he makes no claim to omniscience, that he has found no reason for believing that the departure exceeds that amount (Dorsey 1944, 11).*

**Psychometrics:**
The testing agency, or other judge of performance, feels that the test was valid and that the test scores accurately reflect the examinees' abilities.

**Physical Science:**
The physical sciences rely on the subjective judgements of observers in determining objective knowledge. Those judgements must be informed judgements.

*The experimenter's opinion must rest on evidence, if it is to have any weight. And the only evidence available comes from theory, the series of observations made in the course of the work, and the diligence with which errors were sought (Dorsey 1944, 12).*

**Psychometrics:**
Psychometrics relies on the subjective judgement of experts to decide when the testing process has been successful. They are guided by theories relating to the content and analysis of the tests, the nature of the test responses themselves, and the care with which the whole process has been carried out.

**Physical Science:**
Having carefully constructed an experiment and collected many numbers as carefully as possible, and obtained what we believe to be the definitive measurement, how do we know that it is, in fact, the "true value" that meets our requirements?

*On first thought, the "true value" of the magnitude of a particular quantity appears to be a simple straight-forward concept. On careful analysis, however, it becomes evident that the "true value" of the magnitude of a quantity is intimately linked to the purposes for which knowledge of the magnitude of this quantity is needed, and cannot, in the final analysis, be meaningfully and usefully defined in isolation from these needs (Eisenhart 1969, 22).*

**Psychometrics:**
A concept such as "ability" and its numeric value can be expressed in many ways. Having constructed a test which we feel will enable the particular form of ability in which we are interested to be manifested, we must choose in what way we wish to express its magnitude. There is no "correct" way. A useful way would be to express the ability as a number which is designed to have the properties usually associated with arithmetic, that is to be on a linear scale. Further this number is to be as free as possible from the particular items on the test, the particular judges who may have awarded ratings, and the particular abilities of any other examinees who may have taken the test. Such a number is of as general a nature as the outcome of a test can be. It is this type of measure that is a "specifically objective" measure (Rasch 1966, 21), and the research presented here demonstrates that the many-facet Rasch model is a means of obtaining such measures from judge-awarded ratings of examinees' performances.

## 1.2 PSYCHOMETRIC MEASUREMENT WITH JUDGE INTERMEDIATION

In educational testing, the measurement problem is usually perceived in terms of a simple direct interaction between a person and a set of test items. A test may be composed of a list of multiple-choice questions, or may be presented computer-adaptively, or may be the performance of some task, such as a chemical experiment, or may be in some other form. In all these cases, the test process is regarded as successful when the outcome, for each item, is determined solely by the person's ability and the characteristics of that item.

There are many areas of expertise, however, in which the level of performance can not be determined by a test consisting of multiple-choice, true-false, or other "objective" questions. There are aspects of language skills, artistic talent, and athletic prowess which must be assessed by supposedly qualified judges in a subjective manner. Beyond the educational sphere, judgement by experts is often required to assess the level of proficiency attained. For instance, in measuring employee performance, "by far the most widely used performance measurement techniques are judgmental ones" (Landy and Farr 1983, 57).

The various skills or behaviors on which the examinees are to be rated are here called "items." Judgement is usually expressed in terms of each judge making independent numeric ratings of each examinee's performance on each of several items. The judge may or may not observe all examinees, and may or may not rate all items. Thus the judging situation may be as simple as each judge rating each examinee with 1 for success or 0 for failure on the one assigned item, or as complex as each examinees' performance being assessed on a number of items, each of which has its own rating scale, with individual judges assigned to rate the different examinees and items according to some network of overlapping judgements. The possible ratings on the rating scales are categories defined in ascending order by the level of performance they are intended to represent.

The judging situation may consist of judges rating the performance of examinees on test items, but it may also be expanded to include, say, sets of similar test items relating to a

number of tasks to be performed. For convenience, the various aspects of the judging situation, (be they judges, items, tasks, examinees, or whatever other terms are appropriate in other such situations,) will be called "facets," a term already used in an equivalent manner by Louis Guttman.

A panel of judges may attempt to agree among themselves as to the criteria to be used in judging the items, and the standard that each category on a rating scale represents. Nevertheless, when two judges rate the same performances, their ratings rarely agree perfectly. The differences between their ratings contain both systematic and stochastic elements. Thus, the accurate psychometric measurement of an examinee requires not only the administration of calibrated test items, but also quantitative knowledge of the characteristics of the judges who rate the examinee's performance.

The aim of the testing process is that the measures of performance given to each examinee, as derived from ratings given by judges, be as fair, accurate and useful as possible. Consequently, although the measure given to an examinee is derived from the particular judge or judges who rated that examinee, it must, in meaning and implication, be independent of them. The measure must be "judge-free." Further, when the judging process is repeated for a new cohort of examinees, but criterion standards are to be maintained, or examinee performances are to be compared across judging sessions, there is the requirement that examinee measures be independent not only of the composition of the judging panel, and of the particular items chosen for demonstrating competence, but also of the overall level of performance of the examinees at any session. That is the measures must be "test-free" and "sample-free." Moreover, the uses to which the measures are put include many arithmetical operations, and so the measures must be expressible on a linear scale. Consequently, the most useful measurement model is not that which most completely describes any particular judging situation, but rather that which yields the most generalizable measures, those that are "specifically objective."

## 1.3 THE TYPICAL JUDGING SITUATION

In a typical judging situation, four factors dominate the rating given to an examinee's performance: the ability of the examinee, the difficulty of the item performed, the severity of the judge and the structure of the rating scale. In a competitive diving competition each diver performs a series of dives (the test "items"). Several judges each apply the rating scale to each dive and each award a rating. Judges, however, seldom give the same rating to the same performance. "Everyone *knows* that examiners differ in standard" (Harper and Misra 1976, 2). "In rating examination papers, very great differences of standards appear among supposedly equal judges" (Kelly 1914, 133). Though considerable effort has been expended on diagnosing judge behavior (Guilford 1954, 278) and attempting to modify it, complete agreement among judges is rarely achieved (e.g. Harper and Misra 1976, 11). Nevertheless, in our diving example, the overall goal is to determine the diver of the highest skill, regardless of which dives are actually performed, and which judges happen to rate them.

Under many circumstances, slight differences in the ratings of the same performance by different judges might be acceptable. However, even this level of agreement is hard to obtain.

*In a study designed to see how free from error ratings could be under relatively ideal conditions, Borman selected expert raters, used a carefully designed instrument, and tested it by using videotapes of behavior enacted to present clearly different levels of performance; he succeeded in getting an agreement among the raters of above .80, and yet concluded that ratings are far from perfect (Gruenfeld 1981, 12).*

If there are non-trivial differences between the ratings given by judges, the problem becomes one of determining in what way judges differ, and how these differences can be represented, and hence controlled, in a workable model of the measurement situation which will yield measures with the desirable properties of specific objectivity.

## 1.4 THE STATISTICAL ATTACK

The attainment of this goal of objectivity in judged examinations has motivated considerable research for over a century. An early study reported that
*I find the element of chance in these public examinations to be such that only a fraction - from a third to two-thirds - of the successful candidates can be regarded as safe, above the danger of coming out unsuccessfully if a different set of equally competent judges had happened to be appointed (Edgeworth 1890, 653).*

Another early study reported that
*there is as much variation among the several judges as to the value of each [geography] paper as there is variation among the several papers in the estimation of each judge (Ruggles 1911).*

Edgeworth's 1890 paper is an indication of the lack of real progress made in the development of objective measures from judge ratings in the last 100 years. Edgeworth describes the idea of measurement error which underlies the "true-score" approach to the analysis of judge ratings, but further he reports that in his analysis of the ratings of English composition, error variance is three times the size of variance due to differences in judge severity. His concluding plea is for the further analysis of this problem,

*I submit that the technical treatment of the subject is not to be despised by those who would discover how the element of chance may as far as possible be eliminated, and in what cases our only course is to accommodate ourselves to inevitable uncertainty (Edgeworth 1890, 662).*

For the most part, research is still attempting to eliminate chance, but finding itself forced to accommodate it. Only with the many-facet Rasch model is chance no longer regarded as something to be eliminated but rather as a necessary component of the measurement process.

The development of statistical techniques in the first part of this century led to the first large-scale assault on the arbitrary nature of judged examinations, conducted at the International Conference on Examinations held in May, 1931 at Eastbourne, England, to which six countries sent representatives. Those from the USA included Edward L. Thorndike of Teacher's College, and C. H. Judd of the University of Chicago. At that conference there was unquestioned acceptance of one statistical method of analysis, now known as the true-score method (Hartog and Rhodes 1936, xiv). The only question was which particular formulation was to be preferred.

Since that time many more formulations and adaptations of the true-score method have been suggested, but, as will become clear in the following discussion of many of the main approaches, none of them have achieved the goal of estimating specifically objective measures for the examinees.

The many-facet Rasch model, however, is a conceptually different approach to the problem of variance in judge ratings, and does allow the estimation of specifically objective measures, as will also be demonstrated.

# 2 THE "TRUE-SCORE" APPROACH TO VARIANCE IN JUDGE RATINGS

## 2.1 "TRUE" AS IDEAL

Previous statistical attacks on the problem of disagreement between judges, expressed in the way they award ratings, are now reviewed.

The concept underlying most approaches to the analysis of judged examinations is that each performance can be thought of as meriting a particular "true" score, and that any other observed score must represent a contamination of that "true" score by some additive combination of judgment bias and random error. This concept is a similar to that of the true-score model as applied to the analysis of multiple-choice tests.

The ideal outcome is considered to occur when all judges give exactly the same rating to a performance of a particular merit. This ideal rating would be considered the "true" score. Two benefits immediately follow. First, all ratings would be directly comparable. The same numerical rating by different judges on different items of performance by different examinees would represent the same performance standard. Second, each item of performance need only be rated by one such ideal judge.

Even if such an ideal set of ratings were obtainable, there is still a drawback in this approach. Though equal ratings identify equal performances, it is necessary to make further assumptions about the nature of the rating scale before it is possible to determine the qualitative distance between performances which receive different ratings.

## 2.2 APPROACHES WHICH ASSUME RATING SCALE LINEARITY

The most frequent assertion is that the ratings have been made on a linear scale, so that a rating of "2," given by a judge, is the same distance from a "3" as a "3" is from a "4." Though it is possible for this to be true at the center of a scale, it cannot be true at the ends of a rating scale because each end represents an infinite range of performance, poor at the low end, good at the high end. Even in the center of the scale the definition of categories is arbitrary, and, though the use of consecutive numbers to label the categories represents the intention that they represent equal increments in performance, only a determination of the structure of the rating scale, based on an analysis of the empirical data, can reveal the extent to which this intention has been fulfilled.

Within the true-score framework, several general approaches and many variations of those approaches have been proposed. These all assume that the rating scale either is linear as defined or can be made linear through some arbitrary transformation.

### 2.2.1 Judges only Differ from the "Ideal" in Mean Level of Severity

It has been observed that differences in overall judge severity may account for about as much variance in the numerical ratings given to examinees as do differences in examinee ability (Cason and Cason 1984). An immediate and simple adjustment for this source of variance would be to subtract from each rating given by each judge the mean of all ratings given by that judge.

The fundamental idea underlying this approach has been considerably elaborated upon and widely applied. That idea is that the numerical value of the ratings are equivalent to equally spaced marks on a measuring stick. The problem of resolving differences in judge ratings thus becomes a matter of equating the interval scales corresponding to each judge's different interpretations of the rating scale. Once the equating process has been completed and each judge's interval scale has been aligned to match an "ideal" scale, each examinee's score is the sum total of the aligned ratings. Thus, adjusting the ratings by judge mean rating is an attempt to establish a common origin for the scales.

This approach has been used, but it has not proved satisfactory because, when comparing the way judges award ratings, "the differences in distribution are nearly as large as the differences in mean. Furthermore, they are more important" (Harper and Misra 1976, 253). Realignment implies that judges differ by a fixed amount, so that no attempt at realignment can eliminate the effects of random variation in the scores given by a judge.

### 2.2.2 A Simple True-Score Linear Model of Rating

The existence of random variation is admitted by proponents of the true-score model but decried as undesirable.

*We shall consider an examiner who introduces into his marking random variations of a large order to be less precise in his marking than one whose marks contain less of this element, the perfect examiner being one who introduces no random variation into his marking. Now let us suppose that we had a "perfect" examiner, who could assign to each piece of work the "ideal" mark. The ordinary examiner would differ from him in two ways: his standard might not be the same, and he might introduce random variations into his marking (Hartog and Rhodes 1936, 186).*

Linear true-score models can be constructed to represent this relationship between the observed rating and the ideal or "true" rating. Here is that proposed by Rhodes (Hartog and Rhodes 1936, 315). For each rated performance on a particular test item,

$$(2.1) \qquad X_{nj} = V_j * T_n + M_j + E_{nj}$$

where

$X_{nj}$ = the rating given by judge j to person n on the rated item
$T_n$ = the "true" rating

$V_j$ = adjustment for spread of ratings by judge j
(in most true-score analysis, this is assumed to be 1)
$M_j$ = adjustment for mean of ratings given by judge j
$E_{nj}$ = random (error) component in rating given by judge j to person n, which is assumed to be sampled from a normal distribution with judge-dependent variance.

This has the form of a regression model, and Rhodes presents a method whereby the "true" rating and the other parameters can be estimated (Rhodes and Hartog 1936, 315ff.). The estimate of the "true" rating thus becomes the basis for assessing the examinee's performance, and the standard deviation of the error term distribution indicates the quality of the judge. Though Rhodes assumes that every judge rates every examinee, this technique has also been applied to data sets with many missing observations, with the assumption that the ratings made by any judge are a random sample of his behavior (Lagueux and Amols 1986).

### 2.2.3 More Complex Linear Models of Ratings

One line of development of this linear model has been an attempt to decompose the error terms into different sources of variance. The intention is both to increase the accuracy of the estimate of the "true" rating and also to increase the diagnostic power of the model.

Studies of judge behavior have identified numerous reasons for consistent differences between judges' ratings across all facets, particularly leniency (severity) level, halo effect and central tendency (Guilford 1954, 278-280). Consistent differences are also found in interactions of judges with particular facets of the judging situation. Thus a judge may be consistently lenient on one item, but consistently severe on another. The part of each rating which cannot be accounted for in any systematic way is assumed to be random error.

Following this line of reasoning, numerous studies have attempted to recover the examinees' "true" score from empirically observed ratings. The means for accomplishing this is to construct a model of the form (Saal et al. 1980):

(2.2) Observed Rating = True score + Judge effect + Interaction + Error

Thus, Guilford (1954, 281) presents an expanded version of the true-score model which attempts to quantify various characteristics of judge behavior which have been observed in the field. His model is

$$X_{nij} = T_{ni} + S_j + H_{nj} + C_{ij} + E_{nij} \quad (2.3)$$

where

$X_{nij}$ = rating given to person n on item i by judge j
$T_{ni}$ = "true" rating of person n on item i
$S_j$ = overall severity or leniency of judge j

$H_{nj}$ = "halo" error, the tendency of judge j to over or underrate person n
$C_{ij}$ = "contrast" error, the tendency of judge j to over or underrate item i
$E_{nij}$ = residual error

As is customary, a complete matrix of ratings is assumed. If there are missing observations, "there is no simple, generally applicable solution" (Guilford 1954, 289).

The introduction of interaction terms, such as halo error and contrast error, lessens the generality of the meaning of the ratings given by judges, because these now depend on the particular combinations of judges and examinees that produced the ratings. Nevertheless, Guilford (1954, 284) claims that the judge means and item means provide a base of "objectivity," a topic that will be considered in detail later.

Saal et al. (1980, 413) describe further variations on this approach. The intention of the various researchers is to partition out some particular aspects of rater behavior. But the more precisely the model describes the particular set of ratings being analyzed, the more the results obtained can only relate to the particular judging situation analyzed. If one more person were rated, or one more judge were to rate the performances, the conclusions of the previous analysis would be altered.

### 2.2.4 Linearizing through Transformation

Treating a rating scale as an interval scale ignores the constraint that there is a lowest possible rating and a highest possible rating associated with the scale. Choppin (1982) suggests a transformation of the observed rating scale from numbered categories into a linear logistic scale. De Gruiter (1984) applies this idea to essay examinations and constructs a true-score model in which, for any particular item, each observation includes a logistic term dependent on the ability of the person and the severity of the judge. De Gruiter's model is

$$X_{nj} = T_{nj} + E_{nj} \qquad (2.4)$$

where

$X_{nj}$ = the untransformed rating given person n by judge j
$E_{nj}$ = the random error in the rating

$$T_{nj} = m * \exp(B_n - S_j) / [1+\exp(B_n - S_j)] \qquad (2.4.1)$$

a logistic transformation in which

m = highest category of rating scale, with 0 as the lowest
$B_n$ = ability of person n
$S_j$ = severity of judge j.

In evaluating his model in comparison with an additive linear model of the type described previously, De Gruiter states

*When more accurate results are needed, for example, for intended score corrections, the additive model is clearly inadequate. In such cases the more realistic nonlinear model might be used. Unfortunately not much is known about the properties of the estimates in this model (De Gruiter 1984).*

The cumulative normal ogive has also been proposed as a useful transformation by Cason and Cason (1984), and they report that it reduced error variance by about one third. In practice, but not in theory, this ogive is similar to the logistic ogive.

Though the transformations described here mitigate the most blatant flaw in true-score analysis, which is the non-linear character of the ends of the rating scale, they do not address other irregularities in the structure of the scale, nor do they resolve the problem of random error.

### 2.2.5 Inter-rater Reliability using a Linear Scale

Since attempts to discover the "true" score have not been successful, an alternative approach is to suppose that, if the judges agree in the way they rate examinees, then the observed scores are reliable measures of examinee performance. The inter-rater reliability coefficient quantifies the extent of agreement among judges as to the ratings given to examinees, and has been used as a measure of the quality, and hence accuracy, of the judging process. Such a correlation coefficient has a range of 0 to 1 with higher values regarded as more satisfactory. Then, if the judges are in close enough agreement, differences in ratings are assumed to be trivial and the "average" rating for each examinee, however determined, is assumed to be a good enough estimate of the "true" rating. No adjustment is made for differences in judge severity so that this approach requires that every judge rate every examinee.

Again, the ideal judging situation is that in which all judges agree on every rating, as though the entire panel of judges could be replaced by one judge with no loss of information. The ratings themselves are regarded as measures of the examinee performance, with the inter-rater reliability an indication of how good the measures are. Inter-rater reliability thus informs the decision as to whether the level of agreement is satisfactory or unsatisfactory.

In general, the greater the "true" score variance as compared with the error variance in equation 2.2, the higher the reliability. Thus, depending on the precise formulation of the coefficient being calculated, the reliability is a function of the range of ability of the examinees, the range of severity of the judges, and the range of difficulty inherent in the items. It is thus local to the judging situation and lacks any general significance.

*The extent to which two or more raters independently provide similar ratings on given aspects of the same individuals' behaviors is therefore accepted as a form of consensual validity . . . Several authors, however, have expressed reservations regarding the status of inter-rater reliability or agreement as a criterion of rating quality. Buckner offered data suggesting that high*

*agreement among the ratings assigned the same men by different raters does not necessarily imply predictable or valid ratings, and that disagreement among raters may be associated with predictability and possibly validity (Saal et al. 1980, 419).*

Harper and Misra (1976, 260) support this with the counter-intuitive statement that *examiners who agree are likely to be wrong.*

Here is an example of how this can occur in practice, with a rating scale for a timed task:

```
1.   0 - 5.0 seconds
2. 5.1 - 5.3 seconds
3. 5.4 - 5.6 seconds
4. 5.7 - 6.0 seconds
5. 6.0 and above seconds.
```

If examinees are rated on an item which generally takes about 5.5 seconds, the judges would be expected to rate most examinees in the middle three categories, but not necessarily to show a very high degree of agreement as to which category. But judges slow to start their stop-watches would consistently agree in rating in the lower categories, while those consistently slow in stopping their watches would agree in their ratings of the higher categories. As a result, judge agreement would not imply accuracy in this case.

### 2.2.6 The Factor-analytic Approach

Another approach to judge reliability is factor analysis. Again the ratings are regarded as linear measures. But now variance in judge scores is regarded as caused by the contribution of a number of factors. "The proportion of any observer's variance that is objective in any set of judgments depends upon the extent to which his judgments are determined by common factors" (Guilford 1954, 254). The relationship between the variance attributable to common factors and that which is judge-specific enables an index of reliability to be calculated, which is again used to decide if the empirical scores are a good enough substitute for the "true" scores.

### 2.2.7 The Multitrait-Multimethod (MTMM) Approach

Since the most undesirable feature of observed ratings is felt to be their error variance, attempts have been made to reduce the variance analytically. Campbell and Fiske (1959) and Kavanaugh, McKinney and Wolins (1971) attempt to identify one or more rated items whose ratings are thought to be being highly correlated with a propensity for error. A judge-item correlation matrix is then constructed, and the correlations with the identified items partialled out. The purpose is to improve both the reliability and validity of the ratings by lessening the size of the error term in the linear true-score model. The efficacy of this approach depends entirely on the existence of ratings on items which can be thought to be correlated with the error term, and even its proponents do not suggest it as a "cure" for error prone ratings (Bannister et al. 1987).

### 2.2.8 Generalizability Theory

An undesirable constraint in many analytical approaches to true-score analysis is that every judge rate every examinee on every item. Generalizability theory attempts to overcome this constraint by determining the error variance associated with any particular judge's ratings, so that correction can be made to ratings awarded by a judge when he is the only one to rate an examinee. For this to be useful, examinees must be regarded as randomly sampled from some population of examinees which means that there is no way to correct an individual examinee's score for judge behavior, in a way which would be helpful to an examining board. This approach, however, was developed for use in contexts in which only estimates of population parameters are of interest to the researchers.

Generalizability theory is based on analysis of variance (Cronbach et al. 1972). It is a two-stage approach intended to determine the reliability of a particular judge, or a random judge from a group of judges. Each judge is considered to be a random representative of a population of judges, and the intention is to determine the correlation (generalizability coefficient) between that judge's ratings and the mean ratings of all judges. In the first stage, a "generalizability study," a number of raters rate a small number of examinees and error variances are calculated using the simple linear true-score model described above. In the second stage, the "decision study," the substantive ratings are made (Crocker and Algina 1986, 158ff.) from which the population parameters are estimated. The judges' estimated reliabilities are then used to correct the variance of the population estimates. The precise manner in which the correction is made depends on the design of the judging plan.

As with all these approaches, there are endless opportunities for additional statistical complications. Thus, de Gruiter (1980) considers not only variance but covariance in attempting to determine the structure of the data, and hence its reliability.

A similar approach to generalizability theory, but one in which each judge is calibrated on both severity and error variance, is that of Paul (1981). All judges take part in the generalizability study which is used to estimate a severity and error variance for each judge. In order to do this, the rating scale is assumed to be linear and Bayesian prior distributions are assumed for all parameters. The size of the error variance is used as a guide to selecting the most reliable judges for the decision study. Then each examinee in the decision study can be given a score corrected for the severity of the judge who rated him. This approach lessens, but does not eliminate, the problem of random error.

## 2.3 APPROACHES THAT AVOID ASSUMING RATING SCALE LINEARITY

The problem of determining the numerical nature of the rating scale, together with the errors and adjustments which need to be made to such ratings on account of judge behavior, opens the door to approaches which do not require the rating scale to have arithmetical properties.

### 2.3.1 Rank Order Solutions

An attractive idea is that, even if judges differ in the ratings they award a candidate, the judging process would be successful for many purposes, such as job promotion, if the judges agreed on the ordering of the candidates.

*Examiners who are asked to place answer books in rank order, or order of merit, are asked to do a task which is far simpler for human judgement than is the assigning of absolute marks (Harper and Misra 1976, 255).*

The use of judge-created rankings of the performance of examinees on each test item, instead of judge-awarded scores, removes from the test analysis the severity of the judges, the difficulty of the test items and the arbitrary nature and idiosyncratic implementation of rating scales.

The effect of using rank ordering as the basis of the scoring system is to remove score differences between judges caused by overall differences in judge severity, or due to differences in interpretation of the improvement in observed performance associated with an additional step up the rating scale.

But several benefits of the judging process are lost if the outcome of judging session is one rank-ordered list of the examinees:

1) It is no longer possible to determine what substantive level of competence in individual has reached. Indeed no frame of reference in which levels of competence are specified is available.

2) It is no longer possible to compare individual performances quantitatively, (i.e. in a way corresponding to measurement). Thus being ranked 3 places better than another candidate may, or may not, represent a tangible difference in performance level.

3) "Pass-fail" decisions must be made on a norm-referenced basis (e.g. 75% of candidates pass), because no criterion-referenced basis (demonstrated competence at a pre-determined level) is available.

4) Comparison of examinee performances between judging sessions containing different sample of examinees requires the untestable assumption that performance levels are equivalent.

Nevertheless, the crucial question is that of judge agreement on one ordering of the examinees.

*If there is such agreement, then the disagreement on actual marks is irrelevant, as these can be adjusted by scaling. Thus it is important to know if agreement on order of merit exists (Harper and Misra 1976, 2).*

But the nature of rank ordering is such that the closer two examinees are in ability level, the less likely it is that judges can agree as to their ordering. Consequently, if there are many examinees of similar ability, there will be big numerical differences in the rank order assigned to any one examinee by different judges. On the other hand, if examinees are far apart in ability, there will be agreement in rank ordering any individual examinee, but only small differences in rank between examinees of considerably different abilities.

TABLE 1
RANK ORDERING OF 8 EXAMINEES BY 2 JUDGES
IN TWO EQUALLY VALID WAYS

| Examinee | Examinees' true ability | Examinees as ranked by Judge P | Examinees as ranked by Judge Q | Judges' combined rank order |
|---|---|---|---|---|
| A | 5.0 | 1 | 1 | 1 |
| B | 3.2 | 2 | 2 | 2 |
| C | 3.0 | 3 | 3 | 3 |
| D | 3.0 | 5 | 5 | 4= |
| E | 3.0 | 6 | 4 | 4= |
| F | 3.0 | 4 | 7 | 6 |
| G | 3.0 | 7 | 6 | 7 |
| H | 0.0 | 8 | 8 | 8 |

Table 1 illustrates the equally valid orderings given by two judges. Examinee A is clearly superior and examinee H is clearly inferior, but the rankings do not make this obvious. Indeed examinee B appears to be almost as competent as examinee A. Examinees C through G are equally competent, but the rankings make C look good, while there appears to be considerable disagreement on F. Rank ordering these examinees has not resolved the problem of random variation. In practice, each judge's perception of the ability level of an examinee will not be so precise, so that each ordering will appear to have even more of a stochastic nature.

Since judges do differ in the way they order candidates, a further consideration is the manner in which the orderings are to be combined to yield one overall ordering. However, if the overall ordering is compiled by summing the ranks in the individual judge's orderings, then "it is impossible to assess candidates consistently by only taking rank orders into consideration" (Vassiloglou and French 1982, 191) for the simple reason that removing one candidate or removing one judge can alter the conclusions to be drawn from the overall ranking. In other words, such an overall ranking is not "judge-free" or "sample-free."

Harper and Misra report that constructing a rank ordering for each judge based on the total score he gave each examinee, and then assigning scaled scores to the ranked examinees, based on a normal distribution of the same mean and standard deviation as the complete set of original scores, reduces the range of the scores awarded each examinee by an average

of one-third (Harper and Misra 1976, 258). This finding is in agreement with that of Cason and Cason (1984) who report that 35% of variation in ratings is due to variation in judge severity.

Harper and Misra were required to make the assumption that examinee abilities are normally distributed for their analysis, but this is contradicted by their own empirical observation of marked skewness to the right in their chief data set (Harper and Misra 1976, 130). They do not appear to have noticed this inconsistency, but it is the reason for their conclusion that "the highest agreement of marks was for the weakest candidates, but the highest agreement on ranks is for the best candidates" (Harper and Misra 1976, 14). If the distribution of candidate abilities is somewhat normal, "organizing scores into rank orders tends to conceal the magnitude of ability differences between ranks at the extremes (highest or lowest ranks) and to exaggerate small differences in performance near the middle" (Stanley and Hopkins 1972, 22).

More complex non-parametric models based on rank-ordering, paralleling in conceptual structure the linear models described above, have also been presented in the literature (Wall 1980).

Rank ordering as a device for eliminating random variation has not been successful. An alternative approach, which capitalizes on the random variation, is presented later as a special case of many-facet Rasch measurement.

### 2.3.2 Ratings on Nominal Scales

Instead of considering ratings such as "positive, neutral, negative" as three categories on a continuum, they can be regarded merely as three distinct and mutually exclusive categories. Inter-judge reliability then becomes a statistic reporting to what extent the judges agree on the categories they choose when rating the same performances. Numerous indices have been suggested for contingency tables constructed on this basis (Zwick 1986, Hubert and Golledge 1983). This approach is deficient for performance assessment since judges whose severity differs by one point of the rating scale may have perfect correlation between the ratings they award, but never award the same category. Their substantive agreement is perfect, but their nominal agreement is nil.

A probabilistic latent-class model, in which different classes of judges are allowed different judging styles, has also been suggested (Dillon and Mulani 1984). With such a model, the concern is no longer to measure the facets but only to classify them, and to determine types and degrees of agreement that exist among the judges.

In nominal scale analysis, the problem of false agreement by judges, explained above, is ignored, and decisions relating to examinee performance are an after-thought.

## 2.4 AN ASSESSMENT OF ATTEMPTS TO REMOVE THE EFFECTS OF RANDOM ERROR

Though much ingenuity and statistical expertise has been expended on the problem of obtaining error-free ratings from judges, there is consensus that the solution has not yet been found. "It is therefore imperative that psychologists pursue research that is designed to maximize the desirable psychometric characteristics of ratings and to minimize or eliminate the undesirable characteristics" (Saal et al. 1980, 426). The greatest single problem perceived in judge ratings is random error. The oft proposed solution is "better training programs" (e.g. Bannister et al. 1987), but no examination board has achieved, or can be realistically expected to achieve, the ideal of no random component in judge ratings. Obvious solutions, such as attempts to lessen inter-judge variance by better defining the performance level associated with each step up the rating scale, do not necessarily improve the situation (Kelly 1914, 133).

Even if perfect ratings were to be obtained, and all judges awarded the same rating to each examinee, numerous problems would still remain. Here are some:

1) How much better is a "3" on item 1 than a "2." What is the form of the rating scale?

2) Is a "2" on item 1 better than a "2" on item 2?

How do the items compare in difficulty?

3) Is a "2" on item 1 this year better than a "2" on item 1 last year? Have the judges become more severe as a group, or the examinees more able, or the items easier?

Thus even perfect judge agreement, and so perfect judge reliability, would not mean perfect test interpretability.

## 2.5 THE PRACTICAL PROBLEM OF JUDGING PLANS

In view of the fact that random error does exist in the ratings, an often unstated assumption in many of the analytical techniques mentioned here is that every judge rate every examinee on every item, or that a statistically equivalent, and so generally complex and hard to manage, judging design be followed. An example of such a judging plan is a balanced incomplete block design (Fleiss 1981). This requirement either imposes an impractically large burden on the individual judges, or an impractically difficult management problem on the judging process. Even the most carefully managed judging studies contain mistakes. "Substantial effort was devoted to insuring that the plan was followed exactly . . . The final analysis file [of 2934 ratings] contains only three missing values" (Braun 1986). Nevertheless, values for these missing data had to be imputed in order for the analysis to

continue, and consequently could have been a point of dispute were these data to be used for substantive decisions based on the performance of individual examinees. Large scale judged examinations, or those conducted in less controlled circumstances, would be considerably more prone to analytical problems due to errors in implementing the judging plan.

In practice, judging is an arduous time-consuming task, and the universal desire of all judging panels is to minimize the number of ratings that must be made. The ideal judging plan would be one in which each performance is rated only once, and yet distributive assumptions can also be avoided. Such plans are feasible with the many-facet Rasch model.

## 2.6 FINAL WORDS ON THE TRUE-SCORE METHOD

The multitude of different methods proposed is a indication of the failure of the true-score method to provide a satisfactory solution to the problem of variance in judge scores. The time has come to take a very different conceptual approach, one in which judge variance is an aid rather than a hindrance to measurement. That approach is realized in the many-facet Rasch model.

# 3 THE INEVITABILITY OF MEASUREMENT ERROR

## 3.1 AN INVESTIGATION OF JUDGE RATINGS

The discussion of the true-score method has revolved around the concept of measurement error, which has been presented as inevitable. There is always, however, the hope that some new form of analysis, or some new judge training technique, will remove the element of chance from judge ratings, and produce perfect inter-rater reliability. This hope is vain, as an investigation of judge behavior will make clear.

| | Examinees 1 2 3 4 5 6 |
|---|---|
| Judge A | 4 3 4 2 1 3 |
| Judge B | 4 3 4 2 1 3 |

Fig. 3.1. Perfect agreement in judges' ratings of six examinees on some item. The rating scale has 5 categories in ascending order of performance level, 0,4.

## 3.2 THE QUEST FOR PERFECT RELIABILITY IS DOOMED TO FAILURE

The ideal judging situation would appear to be that in which all judges agree on every rating of examinees that they share in common. An example of this is shown in fig. 3.1. Examinees have been given the same ratings by two independent judges. For the purposes of this immediate discussion, the rating scale is assumed to be an equal interval scale, so that the usual arithmetic operations can be performed on it. This is an assumption in most analysis of rating scales, such as Guilford (1954, 278-301).

Whether perfect agreement is, in fact, the ideal has been questioned by a number of researchers. On the one hand, from the empirical viewpoint, "when two examiners award different marks, the average is more likely to be correct, or nearly correct, than it is when they award the same mark" (Harper and Misra 1976, 262). On the other hand, from the theoretical viewpoint,

*it is usually required to have two or more raters who are trained to agree on independent ratings of the same performance. It is suggested that such a requirement may produce a paradox of attenuation associated with item analysis, in which too high a correlation between items, while enhancing reliability, decreases validity (Constable and Andrich 1984).*

| | Examinees 1 2 3 4 5 6 |
|---|---|
| Judge A | 4 3 4 2 1 3 |
| Judge C | 3 2 3 1 0 2 |

Fig. 3.2. Perfect inter-rater reliability between judges in the rating of six examinees on some item. Judge C is one score-point more severe than Judge A.

The topic of complete agreement, however, is a moot point, because it cannot be expected to occur in any large-scale examination situation. Given that raters do differ, let us consider the question of perfect judge reliability. Fig. 3.2 gives an example of this, under the same conditions and with the same six examinees as fig. 3.1.

It can be seen that the ratings given by the judges are perfectly correlated, and so perfectly reliable according to indices based on product-moment correlation. However, according to indices based on nominal agreements in the ratings, such as Cohen's (1960) Kappa, there is no agreement at all. It can be seen that these judges do agree on the rank order of the examinees, so that to report that they have no agreement at all is clearly misleading. Accordingly we will consider these judges to have perfect inter-rater reliability, but we discern that judge A is one score-point more lenient than Judge C.

| | Examinees 1 2 3 4 5 6 |
|---|---|
| Judge A | 4 3 4 2 1 3 |
| Judge D | 3 3 3 2 0 3 |

Fig. 3.3. Ratings given by two judges when one judge is 0.5 score-points more lenient than the other.

The possibility of perfect inter-rater reliability of judges, whose severity differs by one score point, raises the question of what would be the ratings given by a perfectly reliable judge who is only 0.5 score points more severe than Judge A. Fig. 3.3 makes one suggestion. Judge D must accommodate his behavior to the predefined rating scale, and thus his 0.5 point difference is expressed by awarding half the examinees a rating one point lower than Judge A, and the other half of the examinees the same rating as Judge A. Consequently, two judges, who, in intention, have perfect reliability, are observed to have a correlation coefficient of 0.895.

| | Examinees 1 2 3 4 5 6 |
|---|---|
| Judge A | 4 3 4 2 1 3 |
| Judge D | 3.5 3.5 3.5 2.5 0.5 3.5 |

Fig. 3.4. Ratings given by two judges after a correcting for a judge severity of 0.5 score-points.

Let us say that, as a result of some analysis, Judge D has been determined to be 0.5 score points more severe than Judge A, and a correction of 0.5 points is made in all Judge D's ratings. The outcome is shown in fig. 3.4. We can now test the hypothesis that "adjusting scores for the differences between [judges] should improve the reliability" (Braun 1988, 8). In fact, the inter-rater reliability has not changed, nor, after rounding to the nearest integer category, has the nominal agreement in categories. The correction for judge severity has made no improvement in the reliability of this set of ratings.

| | Examinees 1 2 3 4 5 6 |
|---|---|
| Judge A | 4 3 4 2 1 3 |
| Judge E | 4 2 4 1 1 2 |

Fig. 3.5. Further example of ratings given by two judges when one judge is 0.5 score-points more severe than the other.

Another judge, E, who is also 0.5 score points more severe that Judge A, now awards his ratings and these are shown in fig. 3.5. Again, Judge E expressed his severity by awarding half the examinees a rating one score point below that of Judge A, and the other half the same rating as judge A. Again their agreement, in intention, is perfect but their correlation coefficient is 0.895. Now compare Judges D and E, who are both 0.5 score points more severe than Judge A. Their ratings are shown in fig. 3.6.

| | Examinees 1 2 3 4 5 6 |
|---|---|
| Judge D | 3 3 3 2 0 3 |
| Judge E | 4 2 4 1 1 2 |

Fig. 3.6. Comparison of the ratings awarded by two judges of equal severity

We know that both Judge D and Judge E have the same degree of severity and agree, in intention, as to the standard of performance of the examinees, but the constraints of the rating scale have caused them to express their opinions differently. Judge D and Judge E are reported to have a correlation coefficient 0.645, but, from the point of view of an examining board, even this understates the problem. If a rating of 3 or 4 constituted a pass, and 0, 1 or 2 constituted a failure, then Judge D passes 4 examinees and fails 2, and Judge E passes 2 examinees and fails 4. Judge E's two passes, however, are reported as maximum scores of 4. In traditional analysis, Judge D would be reported as being more severe, but generally in agreement with Judge A, while Judge E would be reported as having a marked inversion of "central tendency."

This paradox of lack of reliability has been presented in terms of one judge being 0.5 score points more severe than another. The very same situation arises, however, when one examinee is 0.5 score points less able than another, even when there is a perfect correlation of judge intentions. Indeed, since the process of measurement is based on the concept that there is a continuum of examinee performance, examinees will always be found who perform, for any particular judge, at or near the transition between adjacent categories.

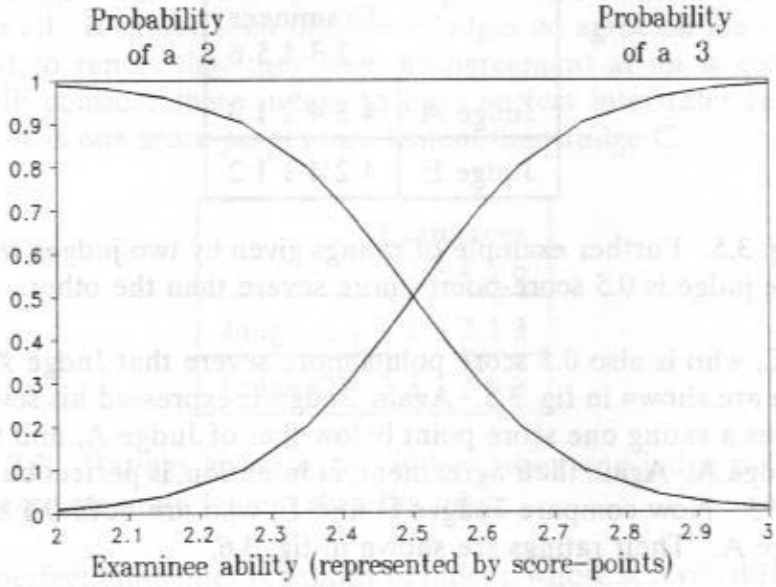


Fig. 3.7. Probability of rating expected to be given to an examinee of given ability.

Even given ideal judges, it is clear that a performance at a level of 2.5 score-points will be awarded 2 points or 3 points with approximately equal frequency. However, with real judges, however well-trained and experienced, it is not clear what will happen to a performance at a level of 2.49 score-points. It cannot be expected that the judges will have such precise discrimination that this performance will always be awarded a rating of 2, and

never a rating of 3. Indeed we expect a greater frequency of 2's than 3's, but not a very great difference. By extension of the same argument, the situation in fig. 3.7 can be expected to result. What appeared to be a deterministic decision by judges is revealed to be a probabilistic one. This stochastic element in rating is what dooms the quest for perfect inter-rater reliability.

## 3.3 MOVING FROM RELIABILITY TO OBJECTIVITY: THE NATURE OF THE RATING SCALE

The fact that the scale we have been discussing has categories numbered 0, 1, 2, 3, and 4 does not force the categories to represent equal increments in performance. In the next judging session, the examining board might decide to introduce a new category between previous categories 2 and 3, and then renumber the scale as 0, 1, 2, 3, 4, and 5. If the old scale were linear, then the new scale isn't, and vice-versa. The numbers assigned to categories are merely a convenience of labelling which enables an ordering of performance levels, but they are not a direct expression of the amount of performance each category represents. Fig. 3.8 illustrates this in the context of a realistic rating scale. Category 0, "Unacceptable," represents all levels of performance below category 1, and so represents an infinite range of performance. Similarly category 4, "Superior" represents all levels of performance above category 3, another infinite range.

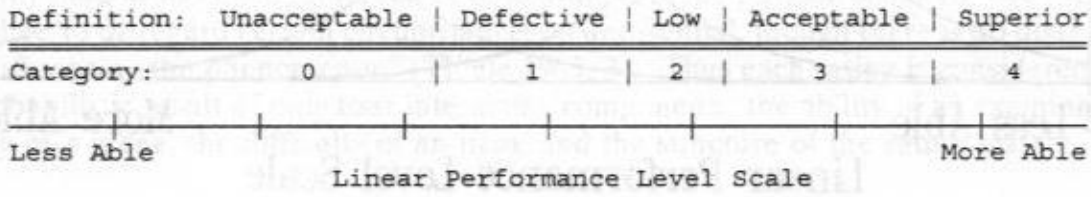


Fig. 3.8. Relationship between category numbers and performance level.

Consequently, no matter how the categories are defined, it is impossible for all of them to represent equal ranges of performance, and thus it is impossible for a numerical scale of untransformed category numbers to be linear.

We have seen illustrated in fig. 3.7 that there is a stochastic element to the awarding of categories. If we extend this finding to fig. 3.8, it can be seen that category 2 represents such a narrow range of real performance, that a "true" or latent performance level on the threshold between a 1 and a 2 could well be rated not only as a 1 or a 2, but even as a 3. In fact, the stochastic nature of judge rating implies that whatever the examinee's performance level, there is some probability that the judge may award a rating in any of the categories, though, for a well designed scale, the category nearest the examinee's performance level has the highest probability. Fig. 3.9 depicts what occurs. This stochastic behavior is what has caused the quest for reliability to fail, but it is this very behavior which provides the key to objectivity and so to fair and sample-free measures.

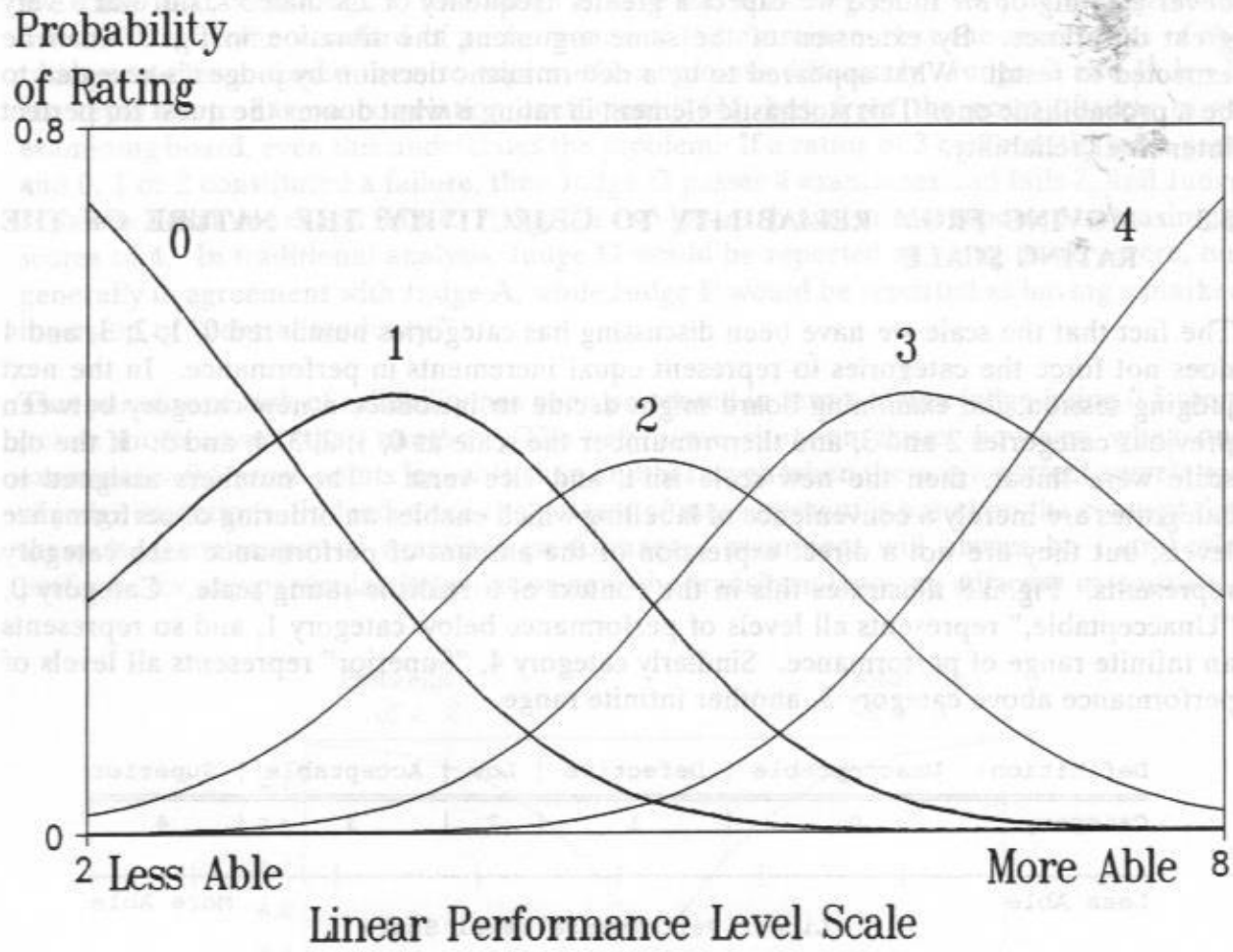


Fig. 3.9. The probabilistic nature of the awarding of ratings by judges.

# 4 THE PURPOSE OF THE MANY-FACET MODEL

The nature of the solution which the many-facet Rasch model offers to the problem of obtaining measures from judge-awarded ratings is now considered.

## 4.1 THE AIM OF THE JUDGING PROCESS

For an examination in which judges rate examinees on test items, the ultimate goal of the judging process, from the viewpoint of an examining board, is not to determine some "true" rating for an examinee on each item, on which ideal judges would agree, but rather to estimate the examinee's latent ability level, of which each judge's rating is a manifestation. This is the very essence of objectivity. In order to supersede the local particularities of the judging situation, each judge must be treated as though he has a unique severity, each examinee as though he has a unique ability, each item as though it has a unique difficulty, and the rating scale as though it has one formulation applied identically by all the judges.

This means that many interesting aspects of behavior must be regarded as incidental, in the same way that, as has been remarked of observations in the physical sciences, "we are often compelled to disregard certain circumstances as unessential, though there is no doubt as to their influencing the phenomenon" (Thiele 1903, 3). Thus each rating is considered to be the probabilistic result of only four interacting components: the ability of an examinee, the severity of a judge, the difficulty of an item, and the structure of the rating scale.

With these assumptions, it is possible to obtain the outcome that the examining board desires, which is an estimate of the ability of each examinee, freed from the level of severity of the particular judges who happened to rate the performance and also freed from the difficulties of the items and the arbitrary manner in which the categories of the rating scale have been defined. The more that incidental aspects of behavior are in evidence in the ratings, the more uncertainty there is in the estimate of each examinee's ability, and the less confidence there is that the aim of the judging process has been realized in the judges' ratings. Thus accurate measurement depends not on finding the one "ideal" judge but in discerning the intentions of the actual judges through the way in which they have replicated their behavior in all the ratings each has made.

## 4.2 THE MANY-FACET RASCH MODEL AND SPECIFIC OBJECTIVITY

The statistical model which permits objective measures to be estimated from judge-awarded ratings is the many-facet Rasch model. But, before deriving the precise form of that model, the requirements it is to meet will be clarified.

Georg Rasch (1960, 1964, 1966, 1968, 1977) discusses the concept of specific objectivity which motivates the construction of model. In my reading of Rasch, I find three key axioms, but this is open to dispute. Fischer, quoted in Roskam and Jansen (1984, 296), states "it is difficult to give a completely general definition of specific objectivity," and Roskam and Jansen (1984, 296) present an axiomatization based on ordering. The axiomatization presented here is in accord with Schleibechner, as quoted in Roskam and Jansen (1984, 296).

First, each separate component of the test situation is represented by a parameter whose value is independent of the other parameters. These parameters could represent examinees, items and judges, and the independent set of parameters representing the structure of the rating scale. These then combine to give the probability that any particular rating will be awarded by any particular judge to any particular examinee's performance on any particular item. In essence, this is the concept underlying Thurstone's Law of Comparative Judgement (Thurstone 1927a and 1927b).

Second, the method of combination of the parameters is additive. Since the rating scale represents increasing amount of the same quality, this implies that the parameters can be expressed on a shared linear scale. In other words, the test is uni-dimensional.

Third, the estimate of any parameter is dependent only on the accumulation of all the ratings in which it participates, but is not affected by the particular values of any of them. In other words, the sum of the ratings in which it participates is the sufficient statistic for the estimation of the parameter (Fisher 1925, 27). Given a particular total score, an unexpectedly high score by one judge does not inflate an examinee's ability estimate, nor does an unexpectedly low score by another judge deflate it. Such scores do, however, threaten the validity of the estimates and so would tend to cause misfit of the data to the model. This axiom appears to have been assumed by Rasch, and is always associated with Rasch models (Andersen 1977), but whether a weaker axiom can be substituted is unclear. "It seems that the measurement argument of specific objectivity and the statistical argument of sufficiency are not well separated in the literature of the Rasch model" (Roskam and Jansen 1984, 296).

These three axioms combine to give parameter estimates with the greatest possible independence, not only from each other, but also from the particular way in which they are manifested by the empirical data. Further the linear, additive, properties of the parameters are those which permit arithmetical manipulations of the estimates to maintain their conventional meanings.

The ratings actually awarded by the judges to the examinees on the items are a realization of the probabilities resulting from the combination of the parameters and may be used to estimate their values. The distribution of the ratings across examinees, judges, items are the means of determining the ability of each examinee, the severity of each judge, the difficulty of each item, and the structure of the rating scale. In this approach, identical

replications of an "ideal" judge are not informative. The modelled parameters can only be estimated when there is some variance in the ratings awarded by the judges.

The axioms given here are equivalent to the requirements for "fundamental" measurement in the physical sciences (N. Campbell 1920), which as Hempel (1952) explains, is not just a matter of establishing operational rules for measurement but also of establishing general laws and theories within which the measures have meaning. The theory underlying Rasch measurement is that of an underlying variable, often referred to as the "latent trait." The parameters are quantitative representations on the underlying variable of abilities, severities, difficulties or the like, whose meaning is determined by the qualitative nature of the test. Thus the parameters estimated from the responses to a properly constructed math test will represent the examinees' ability in math and the difficulty of the math test items. The parameter estimates will not consist of random values, but, when ordered, will be seen to represent increasing quantities of the underlying variable, which can be expressed as increasing ability in math for the examinees, increasing difficulty for the math items, and increasing severity for the judges.

## 4.3 THE MANY-FACET RASCH MODEL AS A STATISTICAL MODEL

The Rasch family of models are not intended as descriptive statistical models (either exploratory or confirmatory) but as measurement models. The analyst designs a particular specification of the model which is intended to extract, from the data, estimates of measures which have meaning outside of the particularities of the data. Consequently the criterion for a successful analysis is not that the model fits the data, but that the data fit the model.

In fact, it can be stated that no set of empirical observations, of any considerable size, ever fits the Rasch model perfectly, just as, in the physical sciences, no measurement is ever perfectly accurate. Nevertheless, the degree of fit of the data to the model can be calculated and the acceptability of the measures, for any particular purpose, determined. This approach must be followed even in the physical sciences, for

*considering now the invention, elaboration, and use of theories which are inconsistent, not just with other theories, but even with experiments, facts, observations, we may start by pointing out that no single theory ever agrees with all the known facts in its domain (Feyerabend 1975, 55).*

An axiom of the Rasch model is that the variable to be measured is uni-dimensional. General lack of fit of the data to the model is an indication that the facets of the model do not combine to realize one underlying variable, in which greater competence by the examinees is expressed in terms of higher ratings awarded by the judges. General lack of fit is not unusual in practical examinations in which conflicting requirements such as speed, accuracy and neatness are combined. Under these circumstances, any decision based on a cumulative score involves an arbitrary weighting of the various non-homogeneous content components by means of the way that they have been included in the test by the testing agency. Unlike true-score models which tend to hide the underlying conflict, Rasch analysis brings this situation to the attention of the examination board, on whom the subjective

policy decisions concerning the relative importance of these different content components rests.

Significant misfit associated with particular parameter estimates, or unexpectedly large residuals for particular observations, are aids to the identification and diagnosis of anomalies in the behavior of the facets of the judging situation. Various forms of remedial action become available and errant (highly improbable) observations can be omitted or re-rated prior to reporting results.

## 4.4 EXTENDING THE RASCH MODEL TO MANY-FACET SITUATIONS

In many assessment situations, the awarding of a score (category on a rating scale) must be based on a judge's subjective appraisal of the performance. In test situations where independence from any particular judge's foibles is desired, several judges are used in an attempt to balance out differences in overall severity and other systematic or random judge effects. The intention is to obtain a measure for each of the examinees as independent as possible of the particular personal attributes of the judges who happen to rate any particular performance. This is to be done by quantifying the attributes of each judge and using this information to eliminate the idiosyncracies of individual judges from the final measures of the examinees.

In "Probabilistic Models for Some Intelligence and Attainment Tests," Georg Rasch writes that "we shall try to define simultaneously the meaning of two concepts: degree of ability and degree of difficulty" (Rasch 1980, 73) and he presents a model to enable this. Rasch's initial model has since been expanded to allow for rating scales through the inclusion of parameters which describe the structure of the rating scale, or scales, underlying the observations (Wright and Masters 1982, Masters 1982).

With the inclusion of judges in the measurement process, it is useful to define simultaneously not only the ability of the examinee, and the difficulty of the test item, but also the severity of the judge. This is accomplished by expanding the rating scale model to include parameters describing each judge's method of applying the rating scale. This introduces the additional facet of judges into the previous framework of examinees and items. The traditional paper-and-pencil test is a two-facet test, and the intermediation of a judge makes the testing situation a three-facet one.

The introduction of facets need not stop at three, and the extra facets need not refer to a judging situation. In principle, there is no limit to the number of facets that can interact to produce a rating, but, in actual testing situations, each of the facets must have a specific meaning for the test developer and user. For instance, there are situations in which each test item is viewed as comprised of several tasks. Thus each answer on an language test could be graded for correct spelling as one task, and correct grammar as another task. The overall difficulty of each item is determined by the nature of the question posed, and the measure for each task is determined by the difficulty of that task, e.g. "providing the correct

spelling" or "using the correct grammar," relative to the overall difficulty characterizing the item. These tasks can be specified to have the same relationship with every item, but each task and each item is allowed to have its own inherent difficulty along some common unidimensional underlying variable.

Another example of a many-facet situation, which does not involve judgment, is that in which soldiers are to hit a variety of targets at several distances using several types of gun. The underlying variable could be denoted as "accuracy." The targets, each with its own distance and size, would be the test items, and the types of gun would represent tasks. The further the target is away, the harder it is to hit. However, it could be hypothesized that one type of gun is inherently more accurate than another, and so makes the target easier to hit. The measure for each soldier would quantify his level of marksmanship, and the calibrations for the tasks (types of gun) would indicate which is the more accurate.

## 4.5 A THREE-FACET MEASUREMENT MODEL FOR JUDGING

In tests involving ratings made by judges, differences in judge severity have frequently been observed, even though they have been hard to quantify or adjust for. The three-facet model allows for the estimation of the difference in severity between judges, and consequently for the corresponding adjustment of the measures of the test-takers, so that the effect of this judge "bias" is eliminated.

"True score" methods have approached the judging situation in terms of the ideal of trying to perfect the accuracy of each rating given by each judge. The ultimate purpose of judging, however, is to obtain an accurate measurement of each examinee. Of course, quality of ratings is important, and it is desirable to minimize the amount of random and systematic error in the ratings. Nevertheless, error will always exist and consequently needs to be regarded as inherent in the judging and, hence, measurement, process. The intention of the judging situation is to quantify some or all facets of the judging situation so that they yield results, that is to say measures, which are sufficiently general to be susceptible to further analysis or interpretation, and ultimately to decision making.

Consider the usual target of the judging process, the examinees. The intention is that whichever judges rate an examinee and however differing be their ratings, the measures obtained, as opposed to the actual rating given, should always be statistically equivalent. Further, whichever judges rate whichever examinees, the examinee measures should always bear the same relationship to one another. This is that specific objectivity already discussed.

The measurement model which meets this requirement is an extension of that invented by Rasch (1960), for the problem of obtaining objective measures from paper-and-pencil tests comprised of dichotomous items and thus is part of the family of probabilistic measurement models described by Rasch (1961). Judge severity, item difficulty, examinee ability and rating scale thresholds can be expressed as real numbers on a common interval scale. They combine additively to provide an expectation of the logarithm of the odds of a judge

awarding a rating in one category to an examinee's performance on an item, as compared to that same judge awarding the rating in the next lower category.

An example of judging to which this three-facet measurement model could be apply is competitive diving, in which 3 or more judges rate contestants on a variety of dives over a number of rounds. The type of dive would be the item (each diver performs a different selection of dives). Each judge has a unique judging style. The outcome of the measuring process is intended to be an objective measure of each diver's performance, enabling divers to be compared.

## 4.6 THE RASCH MODEL AS AN AID TO INFERENCE

The ratings given by judges, or the observations however they were obtained, are manifestations of a local interaction between the facets such as examinees, items and judges. These data remain ambiguous with respect to inference, unless they can be transformed into measures with general meaning which transcends the data from which they were estimated. It is essential for inference that the measures estimated from the observations be independent of the particular sample of examinees and items comprising the test situation. This requirement is especially apparent when examinees do not face identical testing situations. In circumstances where examinees are rated by different judges, respond to different sets of test items, or perform different demonstrations of competence, measures must be independent of the particular local interactions have any meaning outside the particularities of the analysis. The construction of this independence is always necessary when the intention is to compare examinees on a common scale, and to draw conclusion from it. The many-facet Rasch model provides a measurement model with these required properties. Whether, in fact, any particular data set can be used to support general inferences is always a matter for empirical investigation beyond the scope of the original data.

# 5 DERIVATION OF THE MANY-FACET RASCH MODEL FROM OBJECTIVITY

In the previous chapter, the axioms of specific objectivity were discussed. In this chapter they are applied to developing a measurement model for judge-awarded ratings, the many-facet Rasch model.

## 5.1 DERIVING THE GENERAL MANY-FACET RASCH MODEL VIA A THREE-FACET EXAMPLE

The derivation of any particular form of the many-facet model demonstrates the general principles by which any other particular form of the many-facet model can also be derived. The particular model to be derived here is applicable to a three-facet test in which each judge of a panel of judges awards a rating to each examinee on each item.

Consider the performance of two examinees, $O_n$ and $O_m$, as rated by a judge, $J_j$, on replications of the same item, $A_i$. In whatever way the ratings were originally recorded, they have been recoded into K+1 categories ordinally numbered from 0 to K, with each higher numbered category representing a higher level of perceived performance, and with each category having a non-zero probability of occurrence.

The administration of the numerous replications of item $A_i$ is the "test." The performance levels of examinees $O_n$ and $O_m$ can be compared by their relative frequencies of being rated in the various categories of the rating scale. Part of their performance can be summarized by a 2x2 cross-tabulation of counts of ratings in categories k and l of the rating scale, chosen so that category k is numerically greater than category h and so represents a higher performance level. This is depicted in fig. 4.1.

| | | Examinee $O_n$ | |
|---|---|---|---|
| | Categories: | k | h |
| Examinee $O_m$ | k | $F_{kk}$ | $F_{hk}$ |
| | h | $F_{kh}$ | $F_{hh}$ |

Fig. 4.1. Frequency distribution of judge-awarded ratings. $F_{kh}$ represents the count of the number of times that examinee $O_n$ is awarded rating k, when examinee $O_m$ is rated a h, by judge $J_j$ across numerous replications of item $A_i$, where k > h.

When both examinees are given the same rating, which occurs $F_{kk}$ times for a rating of k, and $F_{hh}$ times for a rating of h, their performance levels are indistinguishable. When the examinees are rated differently, which occurs $F_{kh}$ and $F_{hk}$ times, the examinee with the greater relative frequency of ratings in category k, the higher category, is perceived to have the higher ability. In comparing performance levels, we intend that the numeric result be independent of the number of replications. Thus, if the test were to be repeated again, and were of the same length, we would expect to get approximately the same numeric result.

Moreover, if the two tests were then to be concatenated, we would again expect to obtain about the same result. The division of the two frequencies, $F_{kh}$ and $F_{hk}$, is compatible with this expectation because we expect this ratio to be about the same when the test is repeated, and also when the two tests are concatenated. Consequently, the comparative levels of performance of examinees $O_n$ and $O_m$ can be identified by the ratio, $F_{kh}/F_{hk}$.

$$(5.1) \qquad \frac{\text{Performance level of } O_n}{\text{Performance level of } O_m} = \frac{F_{kh}}{F_{hk}}$$

The ratio of the empirically observed frequencies, $F_{kh}/F_{hk}$, is an approximation, which is never exact (Thiele 1903, 133), to the ratio of the probabilities, $P_{kh}/P_{hk}$, where $P_{kh}$ is the probability of examinee On being given a rating of k, when examinee $O_m$ is given a rating of h on one replication of item i. $P_{hk}$ is similarly defined. This unobservable ratio $P_{kh}/P_{hk}$ is defined to be the ratio of the examinee's performances.

$$(5.2) \qquad \frac{\text{Performance level of } O_n}{\text{Performance level of } O_m} = \frac{P_{kh}}{P_{hk}}$$

Now, for objectivity, the ratings given examinees $O_m$ and $O_n$ must be independently awarded by the judge. Consequently,

$$(5.3) \qquad P_{kh} = P_{nijk} * P_{mijh}$$

and

$$(5.4) \qquad P_{hk} = P_{nijh} * P_{mijk}$$

where $P_{nijk}$ is the probability of examinee $O_n$ being given a rating of k on item $A_i$ by judge $J_j$. $P_{nijh}$, $P_{mijk}$, and $P_{mijh}$ are similarly defined. Then

$$(5.5) \qquad \frac{\text{Performance level of } O_n}{\text{Performance level of } O_m} = \frac{P_{kh}}{P_{hk}} = \frac{P_{nijk}}{P_{nijh}} * \frac{P_{mijh}}{P_{mijk}}$$

Furthermore, also for objectivity, the relative performance of examinees $O_n$ and $O_m$ must be independent of which particular item is used to compare them. Thus, though performance levels are initially defined in terms of item $A_i$, the relative performance levels

must have the same value when defined in terms of any conceptually equivalent item $A_{i'}$. That is

$$(5.6)\qquad \frac{\text{Performance level of } O_n}{\text{Performance level of } O_m} = \frac{P_{nijk}}{P_{nijh}} * \frac{P_{mijh}}{P_{mijk}} = \frac{P_{ni'jk}}{P_{ni'jh}} * \frac{P_{mi'jh}}{P_{mi'jk}}$$

then

$$(5.7)\qquad \frac{P_{nijk}}{P_{nijh}} = \frac{P_{mijk}}{P_{mijh}} * \frac{P_{ni'jk}}{P_{ni'jh}} * \frac{P_{mi'jh}}{P_{mi'jk}}$$

For objectivity, this ratio of the probabilities of examinee $O_n$ being rated in categories k and h must be independent of whichever examinee $O_m$ is used in the comparison. So, let us consider examinee $O_0$ with performance level at the local origin of the ability sub-scale. Similarly the ratio must also be independent of whichever item $A_{i'}$ is used for the comparison. Thus it must also be hold for item $A_0$ chosen to have difficulty at the local origin of the item sub-scale.

$$(5.8)\qquad \frac{P_{nijk}}{P_{nijh}} = \frac{P_{0ijk}}{P_{0ijh}} * \frac{P_{n0jk}}{P_{n0jh}} * \frac{P_{00jh}}{P_{00jk}}$$

If, instead of comparing performance levels by means of items $A_i$ and $A_i$', we compare performance levels by means of the ratings given by judges $J_j$ and $J_{j'}$ over numerous replications of item $A_i$, then again we expect the relative performance levels of the examinees to be maintained.

$$(5.9)\qquad \frac{\text{Performance level of } O_n}{\text{Performance level of } O_m} = \frac{P_{nijk}}{P_{nijh}} * \frac{P_{mijh}}{P_{mijk}} = \frac{P_{nij'k}}{P_{nij'h}} * \frac{P_{mij'h}}{P_{mij'k}}$$

so that

$$(5.10)\qquad \frac{P_{nijk}}{P_{nijh}} = \frac{P_{mijk}}{P_{mijh}} * \frac{P_{nij'k}}{P_{nij'h}} * \frac{P_{mij'h}}{P_{mij'k}}$$

Again this must be true if judge $J_{j'}$ is chosen to be judge $J_0$ with severity at the local origin of the severity scale, and examinee $O_m$ is examinee $O_0$, and when item $A_i$ is replaced by item $A_0$. Therefore,

$$(5.11)\qquad \frac{P_{n0jk}}{P_{n0jh}} = \frac{P_{00jk}}{P_{00jh}} * \frac{P_{n00k}}{P_{n00h}} * \frac{P_{000h}}{P_{000k}}$$

Furthermore, for objectivity, the relative severity levels of judges $J_j$ and $J_{j'}$ must be maintained whether the judging takes place over numerous replications of the administration of either item $A_i$ or item $A_i$' to the same examinee $O_n$.

$$(5.12)\qquad \frac{\text{Severity level of } J_{j'}}{\text{Severity level of } J_j} = \frac{P_{nijk}}{P_{nijh}} * \frac{P_{ni'jh}}{P_{ni'jk}} = \frac{P_{nij'k}}{P_{nij'h}} * \frac{P_{ni'j'h}}{P_{ni'j'k}}$$

then

$$(5.13)\qquad \frac{P_{nijk}}{P_{nijh}} = \frac{P_{ni'jk}}{P_{ni'jh}} * \frac{P_{nij'k}}{P_{nij'h}} * \frac{P_{ni'j'h}}{P_{ni'j'k}}$$

Again this must be true if judge $J_{j'}$ is judge $J_0$ chosen at the origin of the severity scale, and examinee $O_n$ is examinee $O_0$, and item $A_i$' is item $A_0$.

$$(5.14)\qquad \frac{P_{0ijk}}{P_{0ijh}} = \frac{P_{00jk}}{P_{00jh}} * \frac{P_{0i0k}}{P_{0i0h}} * \frac{P_{000h}}{P_{000k}}$$

Substituting equations 5.14 and 5.11 in 5.8, and simplifying,

$$(5.15)\qquad \frac{P_{nijk}}{P_{nijh}} = \frac{P_{n00k}}{P_{n00h}} * \frac{P_{0i0k}}{P_{0i0h}} * \frac{P_{00jk}}{P_{00jh}} * \left(\frac{(P_{000h})^2}{(P_{000k})}\right)$$

which gives a general form in which each term is an expression of the relationship between a component of a facet and the local origin of a sub-scale, in the context of a particular pair of categories.

## 5.2 THE THREE-FACET DICHOTOMOUS MODEL

Considering equation 5.15 as a dichotomous model in which k=1, meaning "right," and h=0, meaning "wrong," then this equation expresses the ratio of the probabilities of the possible outcomes as a product of terms which relate each component of each facet with the local origin of its sub-scale. These terms are independent of whichever other examinees, items and judges are included in the test situation. To consider these terms in an additive way, we can take logarithms and assign a numerical direction to each term in accordance with conventional interpretation. Let

$$(5.16)\qquad B_n = \log(P_{n001}/P_{n000}),$$

which is defined to be the ability of examinee $O_n$,

$$D_i = \log(P_{0i00}/P_{0i01}),$$

which is defined to be the difficulty of item $A_i$,

$$C_j = \log(P_{00j0}/P_{00j1}),$$

which is defined to be the severity of judge $J_j$.

We also define the relationship of the sub-scales, such that the probability of "original" examinee $O_0$ being rated a "1" by judge $J_0$ on item $A_0$ is 0.5, so that the last term of equation 5.15 becomes 1. Consequently, reparameterizing $P_{nij}$ to be $P_{nij1}$, so that 1 - $P_{nij}$ becomes $P_{nij0}$, results in equation 5.17, the three-facet Rasch model for the dichotomous case:

$$\log(P_{nij}/(1-P_{nij})) = B_n - D_i - C_j \qquad (5.17)$$

## 5.3 THE THREE-FACET RATING SCALE MODEL

When we consider examinees $O_n$ and $O_m$ in the more general circumstances of a rating scale, we do not wish the comparison of their abilities to depend on which particular pair of categories of the rating scale are used for the comparison. So we return to equation 5.5 which stated:

$$\frac{\text{Performance level of } O_n}{\text{Performance level of } O_m} = \frac{P_{nh}}{P_{mh}} = \frac{P_{nijk}}{P_{mijh}} * \frac{P_{mijh}}{P_{mijk}} \qquad (5.18)$$

We wish to generalize this equation to any pair of categories. The rating scale categories, however, are not independent but structured. In order to determine the structure in an objective manner, we require that performance levels are invariant when they are compared using any pair of adjacent categories in ascending order. This is the only possible objective structuring, since defining invariance not over adjacent categories, but over some pairing of non-adjacent categories, results in a contradiction or indeterminacy in the rating scale structure. Thus, if a rating scale has 3 categories and performance levels are to be invariant only when the top and bottom categories are used for the comparison in equation 5.15, then performance levels based on the middle category are indeterminate, and so are not objective.

Invariance in relative performance when categories are chosen such that k is one greater than h, and also k' is chosen one greater than h', yields

$$\frac{\text{Performance level of } O_n}{\text{Performance level of } O_m} = \frac{P_{nijk}}{P_{mijh}} * \frac{P_{mijh}}{P_{mijk}} = \frac{P_{nijk'}}{P_{mijh'}} * \frac{P_{mijh'}}{P_{mijk'}} \qquad (5.19)$$

so that

$$\frac{P_{nijk}}{P_{mijh}} = \frac{P_{nijk'}}{P_{mijh'}} * \frac{P_{mijk}}{P_{mijh}} * \frac{P_{mijh'}}{P_{mijk'}} \qquad (5.20)$$

Since we want this result to be generalizable, we must be able to substitute examinee $O_0$, item $A_0$, and judge $J_0$,

$$(5.21) \qquad \frac{P_{n00k}}{P_{n00h}} = \frac{P_{n00k'}}{P_{n00h'}} * \frac{P_{000k}}{P_{000h}} * \frac{P_{000h'}}{P_{000k'}}$$

Reordering the terms,

$$(5.22) \qquad \frac{P_{n00k}}{P_{n00h}} = \left(\frac{P_{n00k'}}{P_{n00h'}} * \frac{P_{000h'}}{P_{000k'}}\right) * \frac{P_{000k}}{P_{000h}}$$

The two terms in parentheses are invariant over changes in choice of pairs of categories and so are independent of the local structure of the rating scale, but they are not independent of the choice of object, so we can accordingly rewrite them as $P_{n00}$, so that

$$(5.23) \qquad \frac{P_{n00k}}{P_{n00h}} = P_{n00} * \frac{P_{000k}}{P_{000h}}$$

Similar equations hold for $P_{0i0k}/P_{0i0h}$ and $P_{00jk}/P_{00jh}$, so that, substituting into equation 5.15,

$$(5.24) \qquad \frac{P_{nijk}}{P_{nijh}} = P_{n00} * P_{0i0} * P_{00j} * \frac{P_{000k}}{P_{000h}}$$

This is an equation in which the ratio of the probabilities of particular outcomes is the product of terms which depend only on a single component and the local origin of its sub-scale, combined with a term dependent on the pair of categories used for the comparison. Let

$$(5.25) \qquad B_n = \log(P_{n00}),$$

which is defined to be the ability of examinee $O_n$,

$$D_i = -\log(P_{0i0}),$$

which is defined to be the difficulty of item $A_i$,

$$C_j = -\log(P_{00j}),$$

which is defined to be the severity of judge $J_j$,

$$F_k = -\log(P_{000k}/P_{000h}),$$

which is defined to be the difficulty of the step from category k-1, which is h, to category k of the rating scale.

Then equation 5.24 becomes the equation for the three-facet Rasch rating scale model:

$$(5.26) \qquad \log(P_{nijk}/P_{nijk-1}) = B_n - D_i - C_j - F_k$$

In equation 5.26, the parameters relating to the particular examinees, items, and judges interacting to make each rating have been separated. Thus equation 5.26 immediately meets two of the requirements of objective. First, each component of the judging situation has been characterized by a parameter which is independent of the other parameters. Second, the parameters combine additively in creating the probabilities of the empirical ratings. The third axiom of specific objectivity relates to estimation and will be seen to be satisfied in the course of a later discussion. The log-odds formulation of equation 5.26 determines that the parameter values are expressed in log-odds units, "logits."

To determine the probability associated with category k, $P_{nijk}$, an alternative, exponential, form of equation 5.26 is useful. Summing model statement equation 5.26 for category k with all those below it down to category 1, the probabilities of the intermediate steps are eliminated so that:

$$\log(P_{nijk}/P_{nij0}) = k(B_n - D_i - C_j) - \sum_{s=1}^{k} F_s \qquad (5.27)$$

which leads to

$$P_{nijk} = P_{nij0} * \exp( k(B_n - D_i - C_j) - \sum_{s=1}^{k} F_s ) \qquad (5.28)$$

But some rating in one of the categories 0,K must be awarded, so that

$$\sum_{k=0}^{K} P_{nijk} \equiv 1 \qquad (5.29)$$

Therefore, summing equation 5.28 for categories 1,K with $P_{nij0}$ for category 0,

$$1 \equiv P_{nij0} + \sum_{k=1}^{K} P_{nij0} * \exp( k(B_n - D_i - C_j) - \sum_{s=1}^{k} F_s ) \qquad (5.30)$$

which gives the formulation for the probability of category 0,

$$P_{nij0} = 1/( 1 + \sum_{k=1}^{K} \exp( k(B_n - D_i - C_j) - \sum_{s=1}^{k} F_s )) \qquad (5.31)$$

and, for categories k=1,K,

$$P_{nijk} = \frac{\exp( k(B_n - D_i - C_j) - \sum_{s=1}^{k} F_s )}{1 + \sum_{h=1}^{K} \exp( h(B_n - D_i - C_j) - \sum_{s=1}^{h} F_s )} \qquad (5.32)$$

If the hypothetical step difficulty up to category 0 is defined to be $F_0$, then both the numerator and the denominator in equation 5.32 can be multiplied by $\exp(-F_0)$ without loss of generality, and the equation rewritten, so that for k = 0, K,

$$(5.33) \qquad P_{nijk} = \frac{\exp\left( k(B_n - D_i - C_j) - \sum_{s=0}^{k} F_s \right)}{\sum_{h=0}^{K} \exp\left( h(B_n - D_i - C_j) - \sum_{s=0}^{h} F_s \right)}$$

Equation 5.33 is the exponential form of the many-facet Rasch model.

## 5.4 THE CONVENTIONAL ORIGINS OF THE SUB-SCALES

In the course of the derivation of the many-facet model, the term "local origin of the sub-scale" occurred frequently. In psychometrics, as in many scales in the physical sciences (Kinston 1985), the decision as to where to place the origin of a scale is a matter of the convenience of the investigator. When measuring the height of a table, the floor is a convenient local origin, but when measuring the height of a mountain, sea-level is a different, but equally convenient, local origin. There is no unique, "correct," local origin to the height scale. Neither in the Rasch model are there unique local origins. The local origin of a sub-scale, such as that for ability, is not determined by absence or non-existence of ability, nor yet at some point of minimum intensity (Guttman 1950). Rather it is some level of ability that can be conveniently represented by an arithmetical zero.

Inspection of equation 5.26 reveals that the parameter value estimate of $B_n$, say, could be inflated by an arbitrary amount so long as the other parameter estimates were deflated accordingly, which is a convenient property when equating scales across tests (Wright and Bell 1984, Kelderman 1986). Thus the actual placement of the examinee, item, judge and step sub-scales within the common frame of reference is arbitrary. Nevertheless it is clear that once the origins of all but one sub-scale have been fixed, the origin of the last sub-scale is forced to a unique position on the linear scale as the values of the parameters, relative to their respective origins, must combine to create specific probabilities.

By convention, local origins for each sub-scale are chosen such that the mean calibrations of the items, of the judges, and of the rating scale structure are each zero for the set of ratings being analyzed. The local origin of the examinee abilities' sub-scale is then defined uniquely by the model. In algebraic terms, this means that

$$(5.34) \qquad \sum D_i = 0,$$

where the summation is over all items.

$$\Sigma C_j = 0,$$

where the summation is over all judges.

$$F_0 \equiv 0,$$

where 0 represents lowest category.

$$\Sigma F_k = 0,$$

where the summation is over all categories.

## 5.5 ALTERNATIVE ORIGINS OF THE SUB-SCALES

There are circumstances in which the conventional frame of reference may not be the most useful. For instance, a different origin may be preferred in order to simplify the comparison of examinee abilities between two test administrations. If the second administration uses a subset of items from the first administration, then setting the overall local origin of the item sub-scale to be the mean calibration of the common set of items, has the effect of equating the item sub-scales for the two administrations.

As has been discussed in the context of equation 5.26, the redefinition of the origin of any sub-scale may be accommodated by the adjustment of the origins in one, some, or all of the other sub-scales. To see how this is done, let us consider two sets of equivalent parameter estimates corresponding to the same parameters, with one set established in the conventional frame of reference and the other set in a different frame of reference. Thus, in the conventional frame of reference, we can use the exponential form of the model, equation 5.35,

$$P_{nijk} = \frac{\exp\left( k(B_n - D_i - C_j) - \sum_{s=0}^{k} F_s \right)}{\sum_{h=0}^{K} \exp\left( h(B_n - D_i - C_j) - \sum_{s=0}^{h} F_s \right)} \qquad (5.35)$$

where

$$\Sigma D_i = 0,\ \Sigma C_j = 0,\ F_0 = 0,\ \Sigma F_k = 0.$$

We can also create the same form with different origins, such that the parameter value for examinee n, which is conventionally $B_n$, is here $B_{n'}$. Then

$$P_{nijk} = \frac{\exp\left( k(B_n - D_i - C_j) - \sum_{s'=0}^{k} F_{s'} \right)}{\sum_{h=0}^{K} \exp\left( h(B_n - D_i - C_j) - \sum_{s'=0}^{h} F_{s'} \right)} \qquad (5.36)$$

The terms $\exp(-F_0)$ or $\exp(-F_{0'})$ occur as multipliers of every term in both the numerator and denominator of their respective equations, and so the probabilities (but not the numerical calculations) are independent of whatever values are assigned to either of them.

The equations, 5.35 and 5.36, must produce the same value of $P_{nij0}$, so that

$$P_{nij0} = \frac{\exp(-F_0)}{\sum_{h=0}^{K} \exp\left(h(B_n - D_i - C_j) - \sum_{s=0}^{h} F_s\right)} = \frac{\exp(-F_{0'})}{\sum_{h=0}^{K} \exp\left(h(B_n - D_i - C_j) - \sum_{s=0}^{h} F_s\right)} \qquad (5.37)$$

that is

$$\sum_{h=1}^{K} \exp\left(h(B_n - D_i - C_j) - \sum_{s=1}^{h} F_s\right) = \sum_{h=1}^{K} \exp\left(h(B_n - D_i - C_j) - \sum_{s=1}^{h} F_s\right) \qquad (5.38)$$

The model equations must also produce the same value for $P_{nijk}$ the probability of awarding category k, so that, since the $F_0$ terms cancel, and the denominators are equal by equation 5.38, then the remaining terms in the numerator must also be equal. Thus

$$k(B_n - D_i - C_j) - \sum_{1}^{k} F_s = k(B_n - D_i - C_j) - \sum_{1}^{k} F_s \qquad (5.39)$$

But, for category K, since $\Sigma F_s = 0$, and dividing by K,

$$B_n - D_i - C_j = B_n - D_i - C_j - \left(\sum_{1}^{K} F_s\right)/K \qquad (5.40)$$

Then, substituting equation 5.40 into equation 5.39, and considering category 1,

$$F_1 = F_{1'} - \left(\sum_{1}^{K} F_{s'}\right)/K \qquad (5.41)$$

and, thus, by consideration of the categories of the rating scale in ascending order, the result is obtained that

$$F_k = F_{k'} - \left(\sum_{1}^{K} F_{s'}\right)/K \qquad (5.42)$$

with the value of $F_0$ being independent of that of $F_{0'}$ as previously discussed.

For the other parameters, the translation is simpler. Since the conventional local origin of the item sub-scale is the mean item difficulty, then

$$D_i = D_{i\cdot} - \sum_{1}^{L} D_{i\cdot}/L \qquad (5.43)$$

where the items, $A_i$, are numbered from 1 to L.

Since the conventional local origin of the judge sub-scale is the mean judge severity, then

$$C_j = C_{j\cdot} - \sum_{1}^{J} C_{j\cdot}/J \qquad (5.44)$$

where the judges, $J_j$, are numbered from 1 to J.

Then, by equation 5.40, the examinee ability translation is,

$$B_n = B_{n\cdot} + \sum_{1}^{L} D_{i\cdot}/L + \sum_{1}^{J} C_{j\cdot}/J - \sum_{1}^{K} F_{s\cdot}/K \qquad (5.45)$$

Thus parameter values estimated with respect to one set of local origins can be restated with respect to any other set of local origins.

## 5.6 FURTHER EXAMPLES OF MANY-FACET MODELS

The example used for derivation was a three-facet rating scale model. This can be expanded to as many facets as are required. Thus, in an ice-skating competition, skaters are judged on two aspects ("technical merit" and "artistic impression") of two items ("the short program" and "the long program"). This is a four-facet model: skaters, judges, aspects, items. Algebraically, one expression of such a four-facet model is:

$$\log (P_{nmijk}/P_{nmijk-1}) = B_n - E_m - D_i - C_j - F_k \qquad (5.46)$$

where

$P_{nmijk}$ is the probability of examinee $O_n$ being awarded, on aspect $T_m$ of item $A_i$ by judge $J_j$, a rating of category k.

$P_{nmijk-1}$ is the probability of examinee $O_n$ being awarded, on aspect $T_m$ of item $A_i$ by judge $J_j$, a rating of category k-1

$B_n$ is the ability of examinee $O_n$, where n = 1,N

$E_m$ is the difficulty of aspect $T_m$, where m= 1,M

$D_i$ is the difficulty of item $A_i$, where i = 1,L

$C_j$ is the severity of judge $J_j$, where j = 1,J

$F_k$ is the height of step k of the rating scale on a scale where there are K+1 categories, labelled ordinally 0,K in ascending order of perceived quality.

The many-facet model can be expressed in other ways. The examples given so far have the premise that just one rating scale structure, and so one set of step estimates, $F_k$, applies to all combinations of the other facets. The term, $F_k$, which is only subscripted by the category number, k, specifies that the rating scale has the same structure for each observation, regardless of the examinee, aspect, item or judge concerned. This is the "common step" model, and is always the case when every observation is a dichotomy, in which case k=0 or 1 so that $F_0 = 0$ and $F_1 = 0$.

There are many ways in which the structure of the rating scale can be related to other facets of the model. These include:

The "item-step" model:

$$\log (P_{nijk}/P_{nijk-1}) = B_n - D_i - C_j - F_{ik} \quad (5.47)$$

where $F_{ik}$, with $k = 1,K_i$, implies that each item, i, has its own step structure, with categories numbered 0, $K_i$. This permits the inclusion in one test of items with different numbers of differently defined response categories. Each item has an overall difficulty, $D_i$, and also an additional difficulty associated with each step of its own rating scale, $F_{ik}$. The "partial credit model" (Wright and Masters, 1982) is a two-facet example of the item-step model.

The "judge-step" model:

$$\log (P_{nijk}/P_{nijk-1}) = B_n - D_i - C_j - F_{jk} \quad (5.48)$$

where Fjk implies that each judge has his own way of using the rating categories. Not only may the judges vary in severity, $C_j$, but also in the way they use the response categories, Fjk. For instance, one judge may favor extreme categories, while another may favor central categories. Further, a particular judge may not use certain categories at all. A judge who, though allowed to use the same rating scale as the other judges, only uses the even-numbered categories of a 10 category scale may be regarded as using his own five category scale.

The "judge-item-step" model:

$$\log (P_{nijk}/P_{nijk-1}) = B_n - D_i - C_j - F_{ijk} \quad (5.49)$$

where Fijk implies that each judge has a different way of using the response categories of each item. This may be due to a combination of the factors mentioned in the item-step and judge-step models, or else due to the idiosyncratic behavior of the judges.

These different formulations are particular examples of a more general model, which is useful in practical applications:

The "group-step" model:

$$\log (P_{nijk}/P_{nijk-1}) = B_n - D_i - C_j - F_{gk} \quad (5.50)$$

where Fgk implies that observations may be grouped into G groups, g=1,G, and each group has its own rating scale structure. Thus if the items are designed to have two different rating scale structures, then group 1 may be all three-category items, and group 2 may be all four-category items. Similarly, if all the judges are found to behave similarly except one, group 1 may contain that idiosyncratic judge, and group 2 all other judges. A group model is applicable whenever it is tenable that differences within a group, however defined, can be regarded as random, while differences between groups are systematic.

# 6 ESTIMATING THE PARAMETERS OF THE RASCH MODEL

Having derived the form of the model, an examination is now made of the methods available to estimate its parameters.

The parameters of the many-facet Rasch model cannot be observed or estimated directly, but must be reached through their interaction with one another to produce the observations which comprise the data. Thus, the severity of judges, difficulty of items, ability of examinees, etc. can only be estimated from ratings given by judges to examinees as they respond to items.

The numerical method of obtaining parameter estimates for a particular model from data is at the discretion of the analyst. For instance, the Rasch model can be regarded as a special kind of log-linear model, and the method of iterative proportional fitting applied to the cells of the data matrix (Kelderman and Steen 1988). However, as Neyman and Scott remark,

*we have the almost universal method of obtaining estimates that are not only consistent but asymptotically efficient, in the sense that (1) they tend to be normally distributed as the number of observations grows and (2) their variances tend to zero at least as fast as those of any other estimates. Of course, the method in question is that of maximum likelihood (Neyman and Scott 1948, 2).*

It is this method of maximum likelihood, due to Fisher (1922), that is the estimation procedure discussed here.

Various estimation equations can be derived from the many-facet measurement model equations. Wright's Afterword (Rasch 1980, 188) provides a short description of some methods for the two-facet case. The most widely used of these, as well as some later methods, are discussed in this chapter.

## 6.1 CONSISTENCY AND STATISTICAL BIAS

Not all parameter estimates are equally good. Consistency and bias are two aspects of the quality of statistical estimates. A convenient definition of consistency is: "A statistic is said to be a consistent estimate of any parameter, if when calculated from an indefinitely large sample it tends to be accurately equal to that parameter" (Fisher 1925, 702). Consistency is thus an asymptotic property. On the other hand, for any finite group of observations, the degree to which the mean value of an estimate differs from the corresponding parameter value is the statistical bias. Bias itself is not a fatal defect in an estimate since "a knowledge of the exact form of the distribution of [an estimate] will enable us to eliminate any

disadvantages from which [an estimate] might seem to suffer by reason of bias" (Fisher 1925, 703). Further, "In some cases the bias may be ignored if it is known to be small enough not to invalidate any inference drawn by using the estimate" (Rao 1952, 152).

A clear example of these concepts is given by estimates of the mean and variance of a normal distribution obtained from the mean and variance of a sample from that distribution. The mean of the sample is both a consistent and an unbiased maximum likelihood estimate of the mean of the distribution. The variance of the sample is a consistent, but biased, maximum likelihood estimate of the variance of the distribution. Multiplication of the biased variance estimate by a simple correction factor based on the sample size makes it an unbiased estimate of the variance of the distribution.

Consistency is an asymptotic property and so can only exit when the number of observations in which a parameter appears is conceptually unbounded. Some data sets may be thought of as generated by two types of parameters, incidental and structural (Neyman and Scott 1948). For instance, an arithmetic test is composed of a finite set of items, but could be administered to an unlimited number of examinees. The parameter corresponding to an items could thus appear in an unlimited number of observations, one for each examinee, and this type of parameter is termed "structural." The parameter corresponding to an examinee, however, can appear in only as many observations as there are items on the test, a finite number. This type of parameter is termed "incidental." Since incidental parameters appear in only a finite number of observations, they can have no asymptotic properties and so cannot be consistent.

The maximum likelihood estimates of structural parameters need not be consistent if incidental parameters are also present in the probability expressions underlying the observations. Nevertheless, if the probability expression can be reformulated without the incidental parameters, in other words, if the incidental parameters can be conditioned out, and certain other regularity conditions are met, the resulting maximum likelihood estimates for the structural parameters are consistent, but may be biased (Neyman and Scott 1948).

## 6.2 WHICH PARAMETERS ARE STRUCTURAL AND WHICH ARE INCIDENTAL?

It can be argued that observations (test responses) obtained from conventional two-facet tests are the manifestation of interactions between a strictly finite number of structural parameters, representing the test items, and a conceptually infinite number of incidental parameters, representing the examinees (e.g. Andersen 1973). But the determination of which parameters are incidental and which are structural is doubtful even in this two-facet case.

The traditional view is that a test is a fixed composition which can be administered to an indefinite, and potentially unlimited, number of people. Consider, on the other hand, a computer-adaptive test in arithmetic in which the computer has been given a set of rules for the construction of an unlimited number of appropriate test items. This test is to be given

to a specific group of examinees. Then the number of examinees is a fixed finite quantity, but the number of test items is conceptually infinite. The examinee abilities now become the structural parameters and the item difficulties become the incidental parameters.

In fact, the very concept of fundamental measurement implies a psychological continuum (Thurstone 1927a, 273) of both person ability and item difficulty parameters of which only a sample is manifested in the observations (Wright and Masters 1982, 6). The arbitrary assignment of some parameters as structural and others as incidental is thus contrary to the conceptual basis of fundamental measurement, though it may often be a convenience for the purpose of parameter estimation. Essentially, all parameters are structural.

The arbitrary nature of the decision as to which parameters are structural and which are incidental is even more pronounced in the case of a three-facet examination involving examinees, judges and test items. Examinee parameters could be declared as incidental, and item parameters as structural. But the status of judge parameters remains equivocal, and may well depend on the judging plan to be followed. Indeed, if the study is one of judge behavior, then an indefinite number of judges may observe the same finite number of examinees' behavior on the given set of test items (say, on video tape). In which case, the judge parameters become incidental and the examinee and item parameters become structural.

Paradoxically, it is often the case that it is the parameters generally labelled as incidental that are the desired outcome of the analysis, such as the measurement of examinee ability or judge severity. If the incidental parameters have been conditioned out, so that only the structural parameters have been estimated, then the incidental parameters must be estimated in a second analysis, which will have its own analytical characteristics.

## 6.3 A SAMPLER OF ESTIMATION TECHNIQUES

### 6.3.1 Marginal maximum likelihood (MML) and Bayesian approaches

These estimation methods have been applied to two-facet situations generally with the viewpoint that the item difficulty parameters are structural and the examinee ability parameters are incidental.

*The effect of the incidental ability parameters [is essentially removed] by assuming these values constituted a random sample from a population distribution and then integrating over that ability distribution. The item parameters of a test are then estimated in the marginal distribution of ability, thus the nomenclature "marginal maximum likelihood estimation" (Harwell et al. 1988, p.244).*

As thus described, MML is a mixed effects model in which ability parameters are assumed to be sampled from a distribution but the item parameters have fixed values. The distribution of the ability parameters may be obtained from external information in a

Bayesian manner, or determined from the data itself in an empirical-Bayesian manner, or arbitrarily asserted to have some well-defined form such as a standard normal distribution.

According to Harwell et al., the successful use of MML depends on both a correct specification of the ability distribution and a correct choice of the item-response theory (IRT) model. The problem of a correct choice of IRT model is one that Harwell et al. do not consider in depth. It is clear, however, that they consider MML to be applicable to a descriptive model, rather than a measurement model, as they state that "the metric of the item parameter estimates is defined by the location and scale parameters [of the ability distribution]" (Harwell et al. 1988, 258). Though they presume that the sample being tested is a random sample from the entire population, and consequently that the item parameter estimates are generalizable to that wider population, this is more a matter of fortunate specification of the descriptive model than due to a deliberate choice of a measurement model. "MML, however, has a serious drawback: violation of the normality assumption may distort the item parameter estimates" (Verhelst and Molenaar, 1988, p. 277).

An intriguing variation on MML is the Normal Approximation Estimation Algorithm (PROX) derived by L. Cohen (1979) and fully described in Wright and Stone (1979, 21). It is based on the assumption that both the examinee ability parameters and the item difficulty parameters of the two-facet dichotomous Rasch model are normally distributed with mean and variance to be determined from the observations in an empirical-Bayesian manner. The data is thus formed by the interaction of a random sample of examinee ability parameters and a random sample of item difficulty parameters, and the mean and variance of the ability and difficulty distributions is obtained from the marginal scores of the data.

If this constraint is approximated, then the resultant metric established by the Rasch measurement model is common to both examinee ability and item difficulty estimates, and is item-free and sample-free, once an origin is specified, say, by the mean difficulty of the items. The metric is established by the measurement properties of the model and not by an arbitrary specification of variance of the ability distribution. Nevertheless, once item calibration has been successful (which can be verified, see Wright and Stone 1979, 21ff.), there is no requirement that later samples of examinees have the same ability distribution as the calibrating sample, only that the later samples continue to be relevant.

Though assumptions about the distribution of parameters within each facet may be useful for obtaining initial estimates of parameter values, their further validity is questionable in most many-facet situations, especially those in which some facets have only a few components deliberately selected by an examination board. The distribution of missing data, which in many-facet situations can have either a systematic or a random nature, may also significantly disturb MML parameter estimates.

### 6.3.2 Pair-wise Estimation Algorithm (PAIR)

This method was presented for the dichotomous two-facet model by Rasch (1980, 171-172), developed by Choppin (1968), and generalized to include rating scales by Wright and Masters (1982, 69).

Conceptually, all the ratings in which one parameter, say m, of a facet participated are collected, and similarly all the ratings in which another parameter of the same facet, say n, participated are collected. Then each rating for m is matched with a rating for n in which all the other interacting parameters are identical. Unmatched ratings are ignored.

A "pair-wise" table is constructed containing a row for every possible combination of ratings in which the first rating is greater than the second. Thus, if the rating scale categories are 0,1,2,3, the rows correspond to (1,0), (2,0), (2,1), (3,0), (3,1) and (3,2). Then, for each row, all pairs of the matched ratings are identified whose two categories correspond to those of the row, disregarding order. In each row, counts are entered in two columns. The first column contains the count of the number of matched pairs, identified with that row, in which m is rated higher than n, and the second column contains counts of the number of pairs in which m is rated lower than n.

The ratio of the pair of counts in each row is an empirical manifestation of the difference between the parameters, m and n. Rows in which one or both of the counts is zero are ignored. The logarithms of the ratios are combined to give an estimate of the distance between the parameter values. If there are no matching ratings or if no row has non-zero counts in both columns, the estimate is missing.

This procedure is repeated for all pairs of parameters within each facet, so that estimates of the distances between all pairs of parameters are obtained. The non-missing estimates for the paired differences are then combined arithmetically to give an estimate for each of the parameters in each facet relative to the origin of the sub-scale for that facet.

Further analysis must be performed to determine the numerical adjustment to be applied to the calibrations in each facet in order to align sub-scale origins within one overall frame of reference, and also to estimate the parameters of the rating scale.

Wright mentions, in the context of the two-facet model, that "the theoretical shortcoming of [the pair-wise method] is its neglect of item information not used in the pair-wise analysis" (Rasch 1980, 189). This shortcoming is even more apparent in the estimation of the many-facet model, where the proportion of ratings contributing to the estimation of parameter differences may be small, particularly if a judging plan is followed which specifies that each item of an examinee's performance is only rated by one judge. A further drawback is the lack of an asymptotic standard error estimate (Wright and Masters 1982, 72). These shortcomings become insurmountable in the general many-facet case.

### 6.3.3 The fully conditional estimation algorithm (FCON)

In the conditional method for conventional two-facet data, the examinee ability parameters (or, with equal justification, the item difficulties) are treated as incidental. The procedure is that, for each non-extreme examinee score on the test, a likelihood is obtained, which is the probability of that examinee's observed item response vector, divided by sum of the probabilities corresponding to all possible different item vectors which obtain that same score. The examinee ability parameter is common to all terms and so cancels out of the likelihood expression. The incidental examinee parameter has thus been conditioned out of the estimation by means of its marginal score. Simultaneously maximizing the likelihood expressions for all non-extreme score groups over the empirical data yields consistent estimates of the structural parameters (Andersen 1973). In general, however, these estimates are biased, as will be demonstrated later.

As with all non-Bayesian Rasch estimation, the fully conditional algorithm eliminates extreme scores for both structural and incidental parameters from the data. It also eliminates the possibility of extreme scores corresponding to the incidental parameters from the likelihood formulation. No attempt is made to eliminate the possibility of extreme scores corresponding to structural parameters, since each score group is treated as similarly parameterized but otherwise independent.

Though this estimation algorithm has been regarded as most accurate (Jansen, Van Den Wollenberg, and Wierda, 1988), its practical drawbacks are profound. The fundamental mathematical operation is the calculation of the probability of every way in which an incidental parameter could have obtained its observed marginal score. Then the estimates of the structural parameters are adjusted so that the likelihood of the observed incidental marginal score is maximized. The actual estimation procedure is iterative, and there is the requirement to revise, at each iteration, the probability of every way in which an incidental marginal score could be obtained. This necessitates the repeated calculation of symmetric functions whose number increases exponentially as the number of parameters increases. In spite of possible computational short-cuts, there remains the necessity of calculating a large number of terms with a high degree of precision. This is too resource-consuming for most situations. Fortunately, in these larger situations, the theoretical advantages claimed for FCON also disappear.

### 6.3.4 The unconditional estimation algorithm (UCON)

This approach, originally presented by Wright and Panchapakesan (1969), maximizes simultaneously the likelihood of the marginal score corresponding to every parameter, without making assumptions as to the distribution of the parameters. Since the computations are not conditioned by the marginal scores corresponding to any particular parameters, the technique is referred to as "unconditional." In the two-facet case it has proven to be computationally efficient since, for a given rating scale, the number of computations increases, at worst, linearly with the number of empirical observations.

Further, for tests which are otherwise well-constructed, computational problems relating to loss of precision are rarely encountered.

If, in the two-facet case, examinee ability parameters are regarded as incidental and item difficulty parameters are regarded as structural, i.e. of only a finite number, then the joint estimation of incidental and structural parameters leads to estimates that are not consistent as sample size increases (Neyman and Scott 1948).

On the other hand, if N is the number of ability parameters and L is the number of item parameters, then the item parameters are consistent if log(N)/L approaches 0 as N and L approach infinity with N>=L and some other weak constraints (Haberman 1977, 835). Similarly by symmetry, the person parameters are consistent if log(L)/N approaches 0 as N and L approach infinity with L>=N. Thus, if both sets of parameters are regarded as structural, then estimates for both sets of parameters are consistent provided that the number of parameters in both facets approach infinity together. Thus Andersen's conclusion that "the maximum likelihood estimator for [UCON estimates] is not consistent" (Andersen 1973), deduced from consideration of a special case, is not generally true.

Haberman further points out that the source of bias is the fact that, whatever the true parameter values, there is always some possibility that the empirical data will contain extreme scores. These render the corresponding parameters inestimable, and cause bias in the estimates of the other parameters. The degree of bias is reduced as the probability of extreme scores is reduced, which occurs, in general, as the number of persons and items in a two-facet test increase.

The unconditional estimation procedure, perforce, eliminates extreme scores from the data to be used in estimation. But the statement of the likelihood of the data is based on all possible values of every observation. The likelihood space thus includes the likelihood of vectors of extreme scores. It is this contradictory treatment of extreme scores that gives rise to bias. Nevertheless, for persons and items observed at least 30 times each in a two-facet data set, the existence of bias has proved to have no meaningful practical significance (Wright and Douglas 1976).

### 6.3.5 An extra-conditional estimation algorithm (XCON)

An algorithm, termed extra-conditional and abbreviated as XCON, is proposed here which combines the weaker requirements for consistency of the fully conditional algorithm with the computational simplicity of the unconditional algorithm. XCON explicitly eliminates the possibility of extreme score vectors from the likelihood equations for all facets by means of an approximation. This approximation holds unless the parameter values are such that there is a high probability of extreme score vectors in several facets simultaneously. The mathematical estimation algorithm is derived in Chapter 8. The precise nature of the bias of XCON estimates is discussed in chapter 9, but it is of the order of the bias in FCON estimates.

## 6.4 LIMITATIONS TO ESTIMATION

There are numerous reasons why the estimates of measures obtained from an analysis of data are not identical with the underlying parameter values or with previous estimates of those same parameters. Some reasons of particular relevance to Rasch model estimates are discussed here.

### 6.4.1 The Possibility of Extreme Score Vectors

When the vector of observations corresponding to a parameter is extreme, that is when all observations have the highest possible value or all observations have the lowest possible value, that parameter becomes inestimable. Only its direction is known (e.g. very easy or very hard).

More subtle is the manner in which the possibility of extreme vectors is taken into account by the estimation algorithm. No estimation procedure discussed here is completely satisfactory in this regard. Each algorithm contains some bias, but all estimation procedures tend to give better estimates as the number of observations per parameter becomes larger meaning that the probability of extreme scores diminishes.

Another source of bias in estimating parameters for rating scale data is the fact that, whatever the true parameter values, there is some probability that any category of the scale is not observed. If a category were missing from the empirical data, the structure of the rating scale would change and consequently the estimates of the parameters would be

| | Item responses | | |
|---|---|---|---|
| Person responses | # | Extreme high person vectors | # |
| | Extreme low-item vectors | Vectors of non-extreme observations | Extreme high-item vectors |
| | # | Extreme low person vectors | # |

Fig. 6.1. Sample space of all possible vectors of observations for two-facet data. Only vectors in the central double-lined rectangle yield finite measure estimates.

biased. No estimation algorithm takes account of the probabilities of empty categories, and so all produce estimates based on the assumption that the categories observed in the empirical data exist in all possible data sets. This source of bias is not discussed further in this paper.

In summarizing the manner in which the estimation algorithms treat the possibility of extreme scores, fig. 6.1 is a useful guide for the two facet case. The outside thick rectangle is the space of all possible response the vectors. Only the inner double-lined rectangle of non-extreme response vectors yields finite estimates of measures.

### 6.4.2 Vector space for the unconditional algorithm (UCON)

The sample space for unconditional (joint) maximum likelihood estimation is represented by the thick outside rectangle of fig. 6.1. Vectors containing all possible combinations of observations are included in the likelihood equations. But only non-extreme response vectors are actually used in estimation. This lack of congruence between sample space and estimation space produces statistical bias in UCON estimates.

### 6.4.3 Vector space for the fully conditional algorithm (FCON)

If the person parameters are regarded as incidental, the FCON algorithm explicitly excludes extreme person response vectors from its sample space. But the possibility that every person with a non-extreme score responded in the bottom category of a particular item is not excluded. Thus extreme item vectors are included in the likelihood equations as illustrated in fig. 6.2.

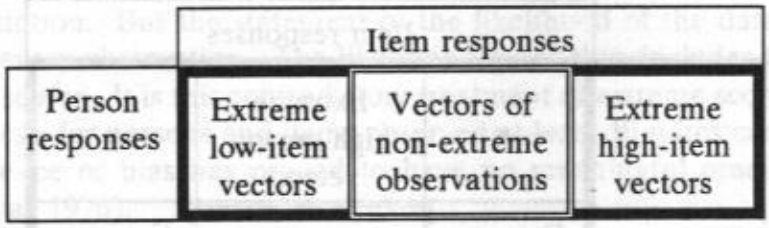


Fig. 6.2. FCON item estimation. Sample space of possible vectors of observations for estimation from two-facet data. Only vectors in the central double-lined rectangle are empirically available.

Alternatively, if the item parameters are regarded as incidental, the vector space for conditional estimation is represented by the vertical rectangle in fig. 6.3, which includes the possibility of extreme person vectors. This lack of congruity in the vector spaces of both of the conditional formulations of the estimation equations shows that the FCON estimates for the two sets of parameters are not exactly compatible and are also slightly statistically biased. The bias in FCON estimates, however, is generally negligible. FCON estimates, however, are consistent, because, when the number of incidental parameters increases

without limit, the probability of an extreme vector for a structural parameter vanishes to zero.

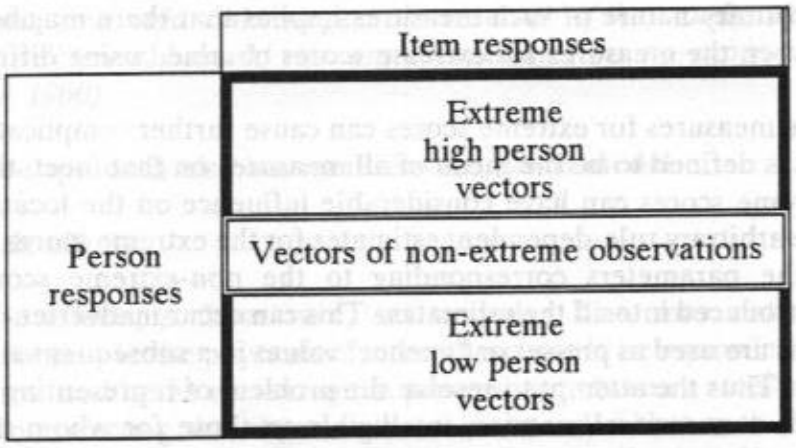


Fig. 6.3. FCON person estimation. Sample space of possible vectors of observations for estimation from two-facet data. Only vectors in the central double-lined rectangle are empirically available.

### 6.4.4 Vector space for the extra-conditional algorithm (XCON)

The vector space for extra-conditional estimation is represented by the internal rectangle of fig. 6.1. The possibility of all vectors corresponding to all extreme scores are explicitly, if approximately, omitted from the likelihood equations. The areas marked "#" in two-facet fig. 6.1, which represent the probability that two or more facets may have extreme response vectors simultaneously, are removed twice in this two-facet example. A correction is made in XCON for this double removal, but not for the triple and higher-order removals implicit in the many-facet implementation of XCON. This is because the probabilities associated with the simultaneous occurrence of extreme vectors in more than two facets are generally negligible.

### 6.4.5 Vector space for the pair-wise algorithm (PAIR)

The PAIR algorithm does not use entire response vectors in its estimation, but rather employs a comparison of the responses made by persons to pairs of items. It takes no account, however, of the fact that the same response may be used in many comparisons, and so bias due to this possibility remains.

## 6.5 MEASUREMENTS CORRESPONDING TO EXTREME SCORES

The maximum-likelihood parameter estimate corresponding to an extreme score is infinite. An infinite estimate, however, is not useful in many practical situations. Consequently numerous approximations, finite substitutes for infinite estimates, have been suggested, such

as those by Swaminathan and Gifford (1982), Wright (1986). Most have a Bayesian justification but little other agreement apart from the fact that the measure for an extreme score must lie between the measure for a "one but extreme score" and the corresponding infinity. The arbitrary nature of such measures implies that there may be little numerical agreement between the measures for extreme scores obtained using different approaches.

Estimating finite measures for extreme scores can cause further complications. If the local origin of a facet is defined to be the mean of all measures on that facet, then the measures assigned to extreme scores can have considerable influence on the location of the origin. Moreover, if the arbitrary rule-dependent estimates for the extreme scores participate in the estimation of the parameters corresponding to the non-extreme scores, an arbitrary component is introduced into all the estimates. This can occur inadvertently when estimates from one analysis are used as pre-set or "anchor" values in a subsequent analysis, item bank or adaptive test. Thus the attempt to resolve the problem of representing the measures for extreme score in a numerical manner, intelligible to those for whom the measures are intended, may itself cause discrepancies in other measures.

## 6.6 THE DATA DO NOT REPRESENT THE INTENDED PARAMETERS

The data, though intended or perceived to manifest the measures governing the measurement model may not, in fact, represent them. From a descriptive viewpoint, this would be regarded as incorrect specification of the model. From a measurement viewpoint, this is a failure of the intended measurement process, and is recognized by a lack of fit of the data to the measurement model. The degree to which lack of fit compromises the meaning and use of the estimated measures is dependent on the purpose for determining measures.

Any measurement situation, in either the physical or social sciences, clearly has numerous sources of variance in the measures which are neither intended nor desired. In an attempt to obtain measures that are generalizable to a wider situation, any measurement model must embody a simplified and useful description of the measurement situation. Two particular types of aberration in the data are of note:

### 6.6.1 Outliers

The existence of outlying data points in a measurement process is a threat to validity, but does not necessarily invalidate the measurement model. Such outliers raise questions about the manner in which the data manifest the desired measurement parameters. In many testing situations, such outliers are the result of behavior which digressed from the trait under examination, e.g. behavior such as guessing or carelessness. The decision to include such data is a deliberate choice which weakens the measurement properties of the model in use.

Often the reasons for the occurrence of outliers are not known. In a measuring process in the physical sciences, Kruskal applies the following principle:
*My own practice in this sort of situation [of wild observations of unknown causality] is to carry out an analysis with and without the suspect observations. If the broad conclusions of the two analyses are quite different, I should view any conclusions from the experiment with very great caution (Kruskal 1960).*

This conservative approach also has merit for the Rasch model.

### 6.6.2 Systematic effects

Some systematic effects, such as race, sex or judge bias, may noticeably distort the measurement process. In the physical sciences, these would be equivalent to an interaction between elements used in the measurement process, as between a steel measuring rod and temperature. In collecting the observations, it is intended that such disturbances be of negligible proportions and randomly distributed in the observations.

When such effects are noticeable and systematic, the analyst can either augment the measurement model to include a measurement of the undesired extra effect, or perform numerous analyses on different subsets of the data to obtain measures generalizable to situations containing different degrees of the extra effect.

## 6.7 THE ARBITRARY NATURE OF THE LOCAL ZERO

Parameter estimates are always determined with reference to some arbitrary local zero of the measurement scale. This local zero is usually first a default of the estimation procedure and then a decision by the analyst. Unnoticed differences in the initial placements of local zeroes may yield numerically different, but conceptually identical, parameter estimates.

## 6.8 THE ARBITRARY NATURE OF CONVERGENCE

Since a closed form solution to the estimation equations only exists in trivial cases, an iterative method must be used. The criteria used to terminate the iterative procedure are at the discretion of the analyst and hence can lead to significant differences in the estimates obtained. Though differences of less than one score point between expected and observed marginal scores cannot be observed, termination of the iteration process at such a coarse difference tends to bias the estimates towards the starting values of the iterative procedure. Indeed, the convergence criteria could be unintentionally made so coarse by the analyst that the starting values require no alteration in order for convergence to be declared, and so the starting values, however arbitrary, become the estimates of the measures.

Further, the nature of the estimation of rating scale parameters is such that the estimates of the step difficulties comprising the rating scale structure may change considerably even after the marginal expected and observed scores, corresponding to the parameters of each

facet, have met the analyst's convergence criteria. This occurs primarily because any arbitrary set of rating scale parameters generates a set of parallel ogives, which can then be used to obtain parameter estimates for the facets with the property that the observed marginal scores match the expected scores. Any arbitrary set, however, will not yield a close match between the observed and expected count of ratings in each category. Consequently, a match to the marginal scores is obtained more quickly than a match to the count of ratings in each category.

# 7 DERIVATION OF THE UNCONDITIONAL ESTIMATION EQUATIONS

## 7.1 ESTIMATION EQUATIONS FOR THE MANY-FACET MODEL

The unconditional estimation algorithm follows the maximum likelihood approach of Fisher (1922), in which it is assumed that the differences between the empirical observations and their expected theoretical values, based on the unobservable true parameter values, are normally distributed. Strictly speaking, Fisher assumed observations on a continuous variable, but his method has also proved effective for discrete observations. In any case, the assumption of normally distributed differences can be tested, post hoc, by examination of the residual differences between the observed ratings and their expected values based on the final parameter estimates. The particular derivation given here closely follows that of Wright and Masters (1982, 73). The generalized many-facet model can be written:

$$\text{(7.1)} \qquad \text{Prob}(x|\{\theta\}) = \frac{\text{Exp}(F(x|\{\theta\}))}{\sum_{k=0}^{K} \text{Exp}(F(k|\{\theta\}))}$$

where

x is an observed datum (ordinally counted)
Θ is a parameter value
{Θ} is the set of all parameter values
Prob(x|{Θ}) is the probability of the datum given the parameter values
F() is a linear function which includes only the parameters which combined to generate the observed datum, x
K is the maximum possible value of each observed datum, in terms of which x is the corresponding value empirically observed
k is each of the possible values of that observed datum 0,K.

Thus Prob(x|{Θ}) is the probability of observing x, given all the parameters. In general, however, at any given time, we are interested in the probability relative to a particular parameter, Θ, which will be written, PxΘ.

The likelihood of the observed data, ∩, is the product of the probabilities of all data. Thus

$$\text{(7.2)} \qquad \cap = \prod_{\Omega} \text{Px}\theta$$

where Ω is the set of all the observed data.

The log-likelihood of the data, ¥, becomes:

$$\log(\cap) = \Psi = \sum^{\Omega} \log (Px\theta) \tag{7.3}$$

The maximum likelihood of this data, for different estimates of parameter $\Theta$, occurs when this likelihood function is at a maximum relative to changes in the parameter estimate of $\Theta$. This occurs when the partial differential of the likelihood function relative to the parameter estimate is zero. The partial differential is:

$$\frac{\delta\Psi}{\delta\theta} = \sum^{\Omega} \frac{\delta Px\theta/\delta\theta}{Px\theta} \tag{7.4}$$

To determine when this partial differential is zero, we may follow Fisher (1922) and apply the Newton-Raphson estimation equation for $\Theta$, which, ignoring covariance, is

$$\theta' = \theta - \frac{\delta\Psi}{\delta\theta} / \frac{\delta^2\Psi}{\delta\theta^2} \tag{7.5}$$

where $\Theta'$ is an improved estimate for the value of $\Theta$ when $\delta\Psi/\delta\Theta = 0$.

Differentiating equation 7.4 gives

$$\frac{\delta^2\Psi}{\delta\theta^2} = \sum^{\Omega} \frac{\delta^2 Px\theta/\delta\theta^2}{Px\theta} - \sum^{\Omega} \left(\frac{\delta Px\theta/\delta\theta}{Px\theta}\right)^2 \tag{7.6}$$

However, particularizing equation 7.1 for the parameter $\Theta$ currently being estimated,

$$Px\theta = \frac{\exp(F(x|\theta))}{\sum_{k=0}^{K} \exp(F(k|\theta))} \tag{7.7}$$

and since there is the general result that

$$\frac{\delta \exp(F(..))}{\delta\theta} = \exp(F(..)) * \frac{\delta F(..)}{\delta\theta} \tag{7.8}$$

then

$$(7.9)\qquad \frac{\delta Px\theta}{\delta\theta} = \mathrm{Exp}(F(x|\theta))*\frac{\delta F(x|\theta)}{\delta\theta} \Big/ \sum_{k=0}^{K}\mathrm{Exp}(F(k|\theta))$$

$$- \mathrm{Exp}(F(x|\theta))*\sum_{k=0}^{K}\mathrm{Exp}(k|F(\theta))*\frac{\delta F(k|\theta)}{\delta\theta} \Big/ \left(\sum_{k=0}^{K}\mathrm{Exp}(F(k|\theta))\right)^2$$

$$= Px\theta*\left(\frac{\delta F(x|\theta)}{\delta\theta} - \sum_{k=0}^{K}\left(Pk\theta*\frac{\delta F(k|\theta)}{\delta\theta}\right)\right)$$

Consequently, substituting in equation 7.4,

$$(7.10)\qquad \frac{\delta\Upsilon}{\delta\theta} = \sum_{\Omega}\frac{\delta Px\theta/\delta\theta}{Px\theta} = \sum_{\Omega}\left(\frac{\delta F(x|\theta)}{\delta\theta} - \sum_{k=0}^{K}\left(Pk\theta*\frac{\delta F(k|\theta)}{\delta\theta}\right)\right)$$

Differentiating again gives

$$(7.11)\qquad \frac{\delta^2\Upsilon}{\delta\theta^2} = \sum_{\Omega}\left(\frac{\delta^2F(x|\theta)}{\delta\theta^2} - \sum_{0}^{K}\left(Pk\theta*\frac{\delta^2F(k|\theta)}{\delta\theta^2}\right)\right.$$

$$\left. - \sum_{0}^{K}\left(Pk\theta*\left(\frac{\delta F(k|\theta)}{\delta\theta}\right)^2\right) + \left(\sum_{0}^{K}\left(Pk\theta*\frac{\delta F(k|\theta)}{\delta\theta}\right)\right)^2\right)$$

Since $F(x|\Theta)$ is a linear function of $\Theta$, its second derivative with respect to $\Theta$ is zero. Thus the two leading terms are zeroes, and the two trailing terms contain constant multipliers of $Pk\Theta$.

The estimation equations for $\{\Theta\}$ are thus obtained by substituting equations 7.10 and 7.11 in 7.5. The effect of ignoring covariance in equation 7.5, however, is to render the divisor in the Newton-Raphson equation generally slightly larger than it should be. Thus, when equation 7.5 is used iteratively to improve the parameter estimates, the change in the estimated values at each iteration is somewhat too small. Nevertheless convergence is usually reached within 20 iterations. Further, following Fisher's reasoning, the asymptotic standard error should also take account of covariance. In practice, however, it is approximated by the square root of the reciprocal of the absolute value of equation 7.11 which is generally smaller than, though close to, its theoretical value.

### 7.1.1 Estimation Equations for the Three-facet Model

For convenience, we will consider the particular case of a three-facet model with a common rating scale, but the same technique can be extended to any other many-facet model. The three-facet model to be considered is:

$$(7.12)\qquad \log(P_{nijk}/P_{nijk-1}) = B_n - D_i - C_j - F_k$$

with the same notation as before.

Define, for convenience of later manipulation,

$$G_k = F_0 + \ldots + F_k \tag{7.13}$$

then, generally,

$$F_k = G_k - G_{k-1} \tag{7.14}$$

and rewrite the model in exponential form, with the constraint that the sum of the probabilities for all the categories of the rating scale be 1,

$$P_{nijx} = \frac{\exp(k*(B_n - D_i - C_j) - G_x)}{\sum_{k=0}^{K} (\exp(k*(B_n - D_i - C_j) - G_k))} \tag{7.15}$$

Then, consider the likelihood of the set of observations in terms of the model:

$$\text{Likelihood } \{X_{nij}\} = \prod^{\Omega} (P_{nijx}) \tag{7.16}$$

where

$X_{nij}$ is the rating given by judge j to person n on item i,

$P_{nijx}$ is the probability of this interaction being rated as $X_{nij}$.

so that, taking logarithms,

$$\text{¥} = \log(\text{likelihood}) = \sum^{\Omega} \log \left( \frac{\exp(x*(B_n - D_i - C_j) - G_x)}{\sum_{k=1}^{K} (\exp(k*(B_n - D_i - C_j) - G_k))} \right) \tag{7.17}$$

where x is written for $X_{nij}$.

$$\text{¥} = \sum^{\Omega}(x*B_n) - \sum^{\Omega}(x*D_i) - \sum^{\Omega}(x*C_j) - \sum^{\Omega}(G_x) - \sum_{k=0}^{\Omega} \log\left(\sum^{K}(\exp(k*(B_n - D_i - C_j) - G_k))\right) \tag{7.18}$$

By substituting with equation 7.5, in the same way as before, the estimation equations become, for $B_n$,

$$(7.19)\qquad B_n' = B_n - \frac{R_n - \sum_{(\Omega|n)}\left(\sum_{k=1}^{K} kP_{nijk}\right)}{\sum_{(\Omega|n)}\left(\sum_{k=1}^{K} k^2P_{nijk} - \left(\sum_{k=1}^{K} kP_{nijk}\right)^2\right)}$$

where

$R_n$ is the sum of all observed ratings in which n took part, so that $R_n$ is the observed marginal score corresponding to $B_n$.

$(\Omega|n)$ means include all data points which are the outcome of an interaction involving $B_n$.

The asymptotic standard error for $B_n$ is approximated by the reciprocal of the square root of the denominator in this estimation equation.

Similar estimation equations can be obtained for $D_i$ and $C_j$, but with a change of sign because $D_i$ and $C_j$ are defined to be in the opposite direction to $B_n$. Thus, for $D_i$,

$$(7.20)\qquad D_i' = D_i - \frac{\sum_{(\Omega|i)}\left(\sum_{k=1}^{K} kP_{nijk}\right) - R_i}{\sum_{(\Omega|i)}\left(\sum_{k=1}^{K} k^2P_{nijk} - \left(\sum_{k=1}^{K} kP_{nijk}\right)^2\right)}$$

A further constraint is imposed on these estimates by the manner in which the frame of reference is determined. This affects the covariance but not the difference between the estimates for the different parameters within the frame of reference.

The parameters corresponding to the categories, the step difficulties, of the rating scale are of a rather different nature to the other parameters. They cannot be independent, and the construction of a frame of reference imposes a much stronger constraint on the step estimates than it does on the item difficulties or parameters of the other facets. Indeed, in the dichotomous case, the step difficulty is defined to be 0 and the steps disappear from the estimation equations. For rating scales in general, however, there are two degrees of freedom fewer than there are categories, and this has a considerable effect on the covariance. Nevertheless, we will ignore covariance for the purposes of estimation, since a simulation study indicates there is no meaningful difference in the estimates.

The estimation equation for the cumulative step difficulties is:

$$(7.21)\qquad G_k' = G_k - \frac{\sum_{\Omega} P_{nijk} - R_k}{\sum_{\Omega}\left(P_{nijk} - P_{nijk}^2\right)}$$

where $R_k$ is the count of all responses in category k of the rating scale and the probabilities are summed over all the observations, $\Omega$, because the rating scale is common to them all.

If any category, k, is not observed in the data, then its parameters, $G_k$ and $F_k$, are inestimable. In this case, $P_{nijk} \equiv 0$ for all n, i, j. Simply skip over category k when estimating the other parameters (Wilson 1991).

An approximation to the asymptotic standard error of $G_k$ is, ignoring covariance,

$$(7.22) \qquad S.E.\ (G_k) = \sqrt{\ 1/(\sum_{\Omega}(P_{nijk} - P_{nijk}^2))\ }$$

A small simulation study indicates this is close to, but generally larger than the true standard deviation of identical replications.

The step difficulties relating to each category of the rating scale, $F_k$, are given by

$$(7.23) \qquad F_k = G_k - G_{k-1}$$

where $F_0 \equiv 0$,

and, pretending the steps of the rating scale are independent, and so ignoring the effect of the structure of the rating scale, an approximation to the standard error of $F_k$ is

$$(7.24) \qquad S.E.\ (F_k) = \sqrt{\ (S.E.(G_k))^2 + (S.E.(G_{k-1}))^2\ }$$

producing

$$(7.25) \qquad S.E.\ (F_k) = \sqrt{\ \sum_{\Omega} 1/(P_{nijk} - P_{nijk}^2) + \sum_{\Omega} 1/(P_{nijk-1} - P_{nijk-1}^2)\ }$$

with S.E. $(F_0) \equiv 0$.

Instead of estimating $F_k$ via $G_k$, it can be estimated directly, as in Wright and Masters (1982, 77), which yields the standard error estimate,

$$(7.26) \qquad S.E.\ (F_k) = 1/\sqrt{\ \sum_{h=k}^{\Omega}\left(\sum_{h=k}^{K} P_{nijk} - \left(\sum_{h=k}^{K} P_{nijk}\right)^2\right)\ }$$

A small simulation study indicates that, in general, equation 7.25 yields an estimate of the standard error that is larger than the standard deviation of identical replications, while equation 7.26 gives a standard error estimate close to that standard deviation.

## 7.2 "MISSING" DATA AND "PERFECT" SCORES

It is a useful property of these estimation equations that there is no requirement that the set of observations be complete, but only that they be sufficiently varying and sufficiently connected to be unambiguously estimable.

They will not be estimable if the marginal score corresponding to any parameter is perfect, which is to say that it is zero or the maximum possible value, because then the corresponding parameter estimate will be infinite. Accordingly, in actual analysis, the observations corresponding to such complete scores are either omitted from the data set or some other constraint as to the nature of extreme scores is imposed on the parameter estimation procedure.

The effect of missing responses is to lessen the information about the parameters and so increase the standard errors. If it is felt that missing responses should be regarded as "wrong" answers, then it may be appropriate to follow the advice of Kruskal (1960) and estimate the parameters twice, once with missing responses omitted and once with missing responses recoded as wrong. Comparison of the resulting estimates and fit statistics will then provide guidance as to the action to take.

Missing data can cause the data to disintegrate into disjoint subsets. Under these circumstances, the meaning of the measures is unclear.

# 8 DERIVATION OF THE EXTRA-CONDITIONAL ESTIMATION EQUATIONS

## 8.1 THE LOGIC BEHIND THE EXTRA-CONDITIONAL ALGORITHM

Parameter estimates corresponding to extreme scores are not obtainable under the Rasch model without the introduction of external assumptions. Generally the vectors of observations corresponding to extreme scores are removed from the data before parameter estimation is attempted. To be completely compatible with this approach, the likelihood equation for the data must also eliminate the possibility of extreme score vectors. The extra-conditional algorithm attempts to do this in a comprehensive yet computationally efficient way, by introducing into the unconditional estimation algorithm the extra condition that the probability of extreme scores be removed from the estimation equations.

## 8.2 THE DICHOTOMOUS TWO-FACET EXTRA-CONDITIONAL ESTIMATION ALGORITHM

Consider the modelled probability, $P_{nix}$, of person n scoring $X_{ni}$ on dichotomous item i, so that $X_{ni}$ is 0 or 1. This is illustrated in fig. 6.4. Even though all observations in fig. 6.4 are possible based on the parameter values, observations in those areas marked "extreme" cannot be used in estimating parameter values from the empirical data because they are included in vectors of extreme scores.

| | UCON | | XCON |
|---|---|---|---|
| Extreme high 1 | $\hat{}$ | $V_{ni1}$ | |
| Non-extreme 1 | $P_{ni1}$ v | $\hat{}$ $P_{ni1} - V_{ni1}$ v | $\hat{}$ $E_{ni1}$ v |
| Non-extreme 0 | $\hat{}$ $P_{ni0}$ v | $\hat{}$ $P_{ni0} - V_{ni0}$ v | $\hat{}$ $E_{ni0}$ v |
| Extreme low 0 | v | $V_{ni0}$ | |

Fig. 8.1. Sample space of the possible responses to dichotomous item i by person n

The modelled probability of an extreme low score vector is the probability that every observation included in the vector is an extreme low observation. Algebraically, the probability of the extreme low-score item vector is given by:

$$(8.1) \qquad V_{i0} = \prod_{n=1}^{N} (P_{ni0})$$

where

$V_{i0}$ is the probability that the score vector corresponding to item i is an extreme low vector.

N represents all persons who responded to item i.

$P_{ni0}$ is the probability that $X_{ni}$, the observation, is 0.

Similarly, the probability of the extreme low score person vector is:

$$(8.2) \qquad V_{n0} = \prod_{i=1}^{L} (P_{ni0})$$

where

$V_{n0}$ is the probability that the score vector corresponding to person n i is an extreme low vector.

L represents all items responded to by person n.

$P_{ni0}$ is the probability that $X_{ni}$, the observation, is 0.

$V_{i0}$ and $V_{n0}$ are not independent because they both contain $P_{ni0}$, which implies that an observation of 0 may be removed twice because it is part of both item and person vectors which are extreme low scores. The combined probability of $P_{nix}$ being eliminated as part of either extreme low-score vector thus becomes:

$$(8.3) \qquad V_{ni0} = V_{n0} + V_{i0} - (V_{n0} * V_{i0}) / P_{ni0}$$

where

$V_{ni0}$ is the probability that observation $X_{ni}$ is part of an extreme low score vector.

$V_{ni0}$ is thus that part of fig. 6.4 labelled "Extreme low 0." $V_{i1}$, $V_{n1}$ and $V_{ni1}$ are similarly defined as the probabilities of $X_{ni}$ being part of an extreme high-score vector. $V_{ni1}$ is that part of fig. 6.4 labelled "Extreme high 1."

After elimination of the probabilities associated with extreme score vectors, the probabilities of the observations that take part in the estimation procedure are that part of fig. 6.4 labelled "Non-extreme." These are comprised of the two values of $E_{nix}$, $E_{ni1}$ and $E_{ni0}$, where $E_{ni1}$ is the probability that $X_{ni}=1$ in the data used for estimation and $E_{ni0}$ is the probability that $X_{ni}=0$. From fig. 6.4, with the constraint that $E_{ni1} + E_{ni0} = 1$ in the data used for estimation, then

$$(8.4) \qquad E_{ni1} = \frac{P_{ni1} - V_{ni1}}{1 - V_{ni1} - V_{ni0}}$$

gives the probability of observing a correct response in the data used for estimation, and is the foundational probability statement of the extra-conditional estimation algorithm.

The likelihood equation for the data used in estimation is thus a function of $E_{nix}$ for all $X_{ni}$, and has a complicated from. But, by comparison with the unconditional estimation algorithm, equations 7.10 and 7.11, the partial derivative of its log-likelihood, ¥, with respect to ability parameter $B_n$ is

$$(8.5) \qquad \frac{\delta ¥}{\delta B_n} = R_n - \sum_{i=1}^{L} \left( \frac{P_{ni1} - V_{ni1}}{1 - V_{ni1} - V_{ni0}} \right)$$

where $R_n$ is the observed score corresponding to parameter $B_n$ in the data used for estimation.

The maximum likelihood thus occurs when the observed and expected marginal scores coincide for the data used in estimating the parameters. A similar formulation holds for the item parameters, $D_i$.

## 8.3 EXTRA-CONDITIONAL ESTIMATION OF THE TWO-FACET MODEL FOR RATING SCALES

In a data set containing rating scales, only extreme ratings can form part of extreme vectors. Intermediate ratings cannot form part of extreme scores and so are immune from elimination from either the observed data or the likelihood equations provided that at least one occurrence of each category is contained in the empirical data. There is some probability, however, that no observations occur in a particular category. This source of bias is not considered here.

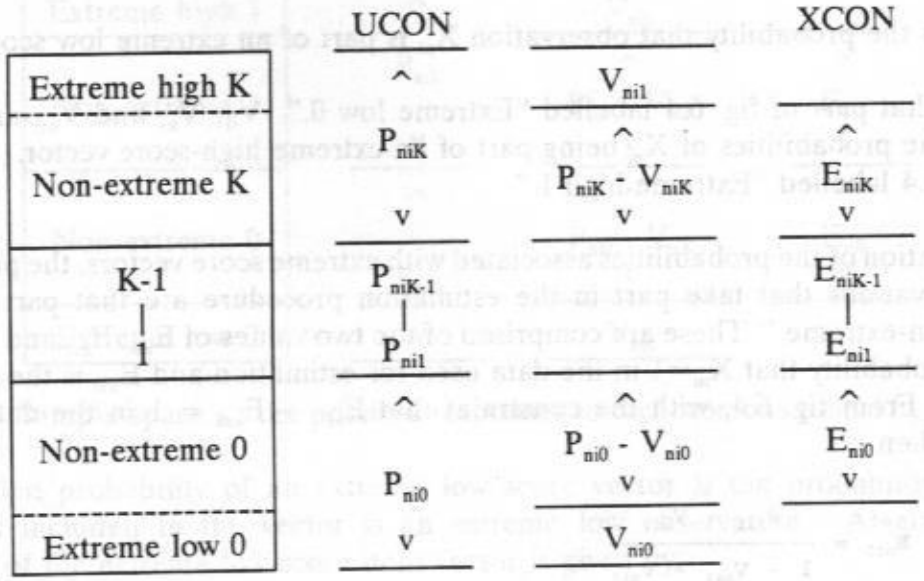


Fig. 8.2. Sample space of the $X_{nik}$ (0,K) responses to polytomous item i by person n.

Thus, in generalizing the dichotomous extra-conditional model to rating scales, only the probabilities relating to the extreme categories are restricted by probabilities associated with extreme scores. So, by reference to the notation of fig. 6.5, the following probabilities are observed:

$$E_{nik} = (P_{nik} - V_{nik}) \,/\, (1 - V_{ni0} - V_{nik}) \tag{8.6}$$

$$E_{nik} = P_{nik} \,/\, (1 - V_{ni0} - V_{nik}) \tag{8.7}$$

$$E_{ni0} = (P_{ni0} - V_{ni0}) \,/\, (1 - V_{ni0} - V_{nik}) \tag{8.8}$$

The derivative of the maximum likelihood equation for $B_n$ thus becomes:

$$\frac{\delta \Psi}{\delta B_n} = R_n - \sum_{i=1}^{L} \left( \frac{\sum_{k=1}^{K} (kP_{nik}) - K*V_{nik}}{1 - V_{nik} - V_{ni0}} \right) \tag{8.9}$$

where $R_n$ is the observed score corresponding to parameter $B_n$ in the data used for estimation, and similarly for $D_j$, where again the maximum likelihood occurs when the observed marginal score equals the expected marginal score for the estimated data.

The second derivative of the log-likelihood, from which the asymptotic standard error is obtained, is given by

$$\frac{\delta^2 \Psi}{\delta B_n^2} = -\sum_{i=1}^{L} \left( \frac{\sum_{k=1}^{K} (k^2 P_{nik}) - P_{ni}^2 - K\left(\frac{\delta V_{nik}}{\delta B_n}\right)}{1 - V_{nik} - V_{ni0}} + \frac{(P_{ni} - K*V_{nik}) * \left(\frac{\delta V_{nik}}{\delta B_n} + \frac{\delta V_{ni0}}{\delta B_n}\right)}{(1 - V_{nik} - V_{ni0})^2} \right) \tag{8.10}$$

where

$$P_{ni} = \sum_{k=1}^{K} (k_{Pnik}) \tag{8.11}$$

$$\frac{\delta V_{nik}}{\delta B_n} = V_{nk} * \sum_{j=1}^{L} (K - P_{nj}) \left(1 - \frac{V_{ik}}{P_{nik}}\right) + V_{ik}*(K - P_{ni}) \tag{8.12}$$

$$\frac{\delta V_{ni0}}{\delta B_n} = V_{n0} * \sum_{j=1}^{L} (-P_{nj}) \left(1 - \frac{V_{i0}}{P_{ni0}}\right) + V_{i0}*(-P_{ni}) \tag{8.13}$$

using the same notation, $V_{n0}$ etc., as in the dichotomous formulation.

The maximum likelihood condition for $G_k$, the cumulative step difficulty, then becomes, by comparison with the unconditional estimation algorithms,

$$(8.14) \qquad \frac{\delta \Psi}{\delta G_k} = R_k - \sum_{n=1}^{N} \sum_{i=1}^{L} \left( \frac{P_k}{1 - V_{ni0} - V_{niK}} \right)$$

where

$R_k$ is the count of observations in category k
if k = 0, then $P_k = P_0 = P_{ni0} - V_{ni0}$
if k = K, then $P_k = P_K = P_{niK} - V_{niK}$
else $P_k = P_{nik}$

with the corresponding second derivatives for the cumulative step difficulties being

$$(8.15) \qquad \frac{\delta^2 \Psi}{\delta G_k^2} = - \sum_{n=1}^{N} \sum_{i=1}^{L} \left( \left( \frac{\frac{\delta P_k}{\delta G_k}}{1 - V_{ni0} - V_{niK}} \right) + \left( \frac{P_k * \left( \frac{\delta V_{niK}}{\delta G_k} + \frac{\delta V_{ni0}}{\delta G_k} \right)}{(1 - V_{ni0} - V_{niK})^2} \right) \right)$$

here

when k = 0,

$$(8.16) \qquad \frac{\delta P_k}{\delta G_k} = P_{ni0} * (1-P_{ni0}) - \frac{\delta V_{ni0}}{\delta G_k}$$

when k = K,

$$(8.17) \qquad \frac{\delta P_k}{\delta G_k} = P_{niK} * (1-P_{niK}) - \frac{\delta V_{niK}}{\delta G_k}$$

otherwise,

$$(8.18) \qquad \frac{\delta P_k}{\delta G_k} = P_{nik} * (1-P_{nik})$$

and, when k=K,

$$(8.19) \qquad \frac{\delta V_{niK}}{\delta G_k} = V_{nK} * \sum_{j=1}^{L} (1 - P_{njK}))(1- V_{nK}/P_{niK})$$

$$+ V_{iK} * \left( \sum_{m=1}^{N} (1 - P_{miK}) \right) (1- V_{iK}/P_{niK})$$

$$+ V_{nK} * V_{iK} * (1 - P_{niK}) / P_{niK}$$

otherwise

$$(8.20)\qquad \frac{\delta V_{nix}}{\delta G_k} = V_{nx} * \sum_{j=1}^{L} (-P_{njk}))(1- V_{nx}/P_{nix}) + V_{ix} * (\sum_{m=1}^{N} (-P_{mik}))(1- V_{ix}/P_{nix}) + V_{nx} * V_{ix} * (-P_{nik})/P_{nix}$$

and, when k = 0,

$$(8.21)\qquad \frac{\delta V_{ni0}}{\delta G_0} = V_{n0} * \sum_{j=1}^{L} (1 - P_{nj0}))(1- V_{n0}/P_{ni0}) + V_{i0} * (\sum_{m=1}^{N} (1 - P_{mi0}))(1- V_{i0}/P_{ni0}) + V_{n0} * V_{i0} * (1 - P_{ni0})/P_{ni0}$$

otherwise

$$(8.22)\qquad \frac{\delta V_{ni0}}{\delta G_k} = V_{n0} * \sum_{j=1}^{L} (-P_{njk}))(1- V_{n0}/P_{ni0}) + V_{i0} * (\sum_{m=1}^{N} (-P_{mik}))(1- V_{i0}/P_{ni0}) + V_{n0} * V_{i0} * (-P_{nik})/P_{ni0}$$

In order to establish a frame of reference for the rating scale, two of the $G_k$ are given fixed values. The usual convention is that $G_0 = G_K = 0$.

The terms in the first and second derivatives, though complex in appearance, add very little computational burden above that of the unconditional estimation algorithm. The only significant extra arithmetic is the creation of the running products of the probabilities of extreme score vectors, $V_{n0}$ and $V_{i0}$. These will change slowly as convergence is reached, and so the values calculated during one iteration may be used as the current values during the next iteration with little loss of estimation efficiency, but considerable gain in computational economy.

## 8.4 EXTRA-CONDITIONAL ESTIMATION OF THE MANY-FACET MODEL

The two-facet estimation equations can be rewritten for the many-facet situation in a manner similar to the equivalent unconditional estimation equations. Thus the term $V_{ni0}$ becomes, for the three-facet model, $V_{nij0}$, where

$$V_{nijo} = V_{no} + V_{io} + V_{jo} - (V_{no}*V_{io} + V_{no}*V_{jo} + V_{io}*V_{jo})/P_{nijo} \quad (8.23)$$

and its partial derivatives also have the consequent extra terms.

A more precise form of $V_{nijo}$ would include a term modelling the probability that all three score vectors for a datum are extreme simultaneously, namely $V_{no}*V_{io}*V_{jo}/(P_{nijo}*P_{nijo})$. This term and the even higher order terms of more complex many-facet models are generally significantly smaller than the other terms in equation 8.23, and, in general, increase the computational burden without yielding commensurate benefits.

# 9 NUMERICAL BIAS IN ESTIMATION

Considerable thought has been given to the matter of inconsistency and bias in Rasch model estimates (Andersen 1973, Wright and Douglas 1976, 1977, Haberman 1977, Jansen, Van Den Wollenberg and Wierda 1988). As discussed in chapter 6, the Rasch model estimators of interest here, namely FCON, UCON, XCON and PAIR are all consistent under ideal conditions. However, for any finite set of parameters, they are all biased. The numerical nature of that bias is discussed here, in the context of dichotomous data.

## 9.1 THE TWO-ITEM MULTIPLE-PERSON DICHOTOMOUS TWO-FACET TEST

Let us consider a two-facet test with two items, I1 and I2, of difficulties D1 and D2. These are responded to by a number of persons of unknown ability. Fig. 9.1 illustrates the possible outcomes. N00 and N11 merely tell us that those people found both items either too hard or too easy and so tell us nothing of their relative difficulty.

| | Item 2 | |
|---|---|---|
| | Failed | Succeeded |
| Item 1 Failed | N00 | N01 |
| Succeeded | N10 | N11 |

Fig. 9.1 The outcome of a two-item multiple-person test. N00, N01, N10, N11 are the counts of persons in each cell of the 2x2 mutually exclusive classification.

The comparison of the difficulty of the items becomes a comparison between N01 and N10. These provide an estimate, D1'-D2', of the relative difficulty of the two items, D1 - D2, through the Rasch model estimation equation,

$$D1 - D2 \approx D1' - D2' = \log(N01/N10) \qquad (9.1)$$

where

D1' and D2' are the parameter estimates

N01 is the count of the number of times objects (persons) failed on item 1 and succeeded on item 2,

N10 is the count of the number of times objects (persons) succeeded on item 1 and failed on item 2.

For the purposes of comparing estimation algorithms, we will consider all N persons to be of equal ability, where N = N01 + N10. Thus all N persons scored 1 on the two item test. Any persons included in N00 or N11, in fig. 9.1, scored 0 or 2 on the two item test and so have extreme scores. The measures corresponding to such scores are not estimable without further assumptions outside of the measurement model, and so are eliminated from further consideration here.

## 9.2 METHODS OF ESTIMATING THE RELATIVE DIFFICULTIES OF TWO ITEMS

### 9.2.1 The log-odds estimator (LOE)

An immediate estimate of the relative difficulty of the two items is given by direct application of objectivity,

```
(9.2)     D1' - D2' = log (N01/N10)
```

as described above.

### 9.2.2 The conditional estimator (FCON)

The conditional maximum likelihood solution to the Rasch equations for the dichotomous two-facet case can be deduced from by Wright and Masters (1982, 86). The maximum likelihood conditions for a two item test are, for item 1,

```
(9.3)     N10 = (N10+N01) * exp(-D1')/(exp(-D1')+exp(-D2'))
```

and, for item 2,

```
          N01 = (N10+N01) * exp(-D2')/(exp(-D1')+exp(-D2'))
```

Therefore, dividing the two equations and taking logarithms,

```
(9.4)     D1'-D2' = log(N01/N10)
```

agreeing with the log odds estimate (LOE).

### 9.2.3 The pair-wise estimator (PAIR)

Following Wright and Masters (1982, 72), the dichotomous two-facet PAIR maximum likelihood conditions for two items are, for item 1,

```
(9.5)     N10 = (N10+N01) * exp(-D1')/(exp(-D1')+exp(-D2'))
```

and, for item 2,

```
(9.6)     N01 = (N10+N01) * exp(-D2')/(exp(-D1')+exp(-D2'))
```

so, dividing these and taking logarithms,

(9.7) $$D1' - D2' = \log(N01/N10)$$

agreeing with the log odds estimate (LOE).

### 9.2.4 The unconditional estimator (UCON)

Following Wright and Master (1982, 77), the maximum likelihood conditions for the two item test are, for item 1,

(9.8) $$N10 = (N10 + N01)\ \exp(B'-D1')/(1+\exp(B'-D1')$$

where B' is an estimate of the person ability corresponding to a score of 1,

and, for item 2,

(9.9) $$N01 = (N10 + N01)\ \exp(B'-D2')/(1+\exp(B'-D2')$$

Rewriting these,

(9.10) $$\exp(D1' - B') = (N01+N10)/N10 - 1 = N01/N10$$

(9.11) $$\exp(D2' - B') = (N01+N10)/N01 - 1 = N10/N01$$

and logarithms and subtracting

(9.12) $$D1' - D2' = \log(N01/N10) - \log(N10/N01) = 2*\log(N01/N10)$$

Thus the UCON estimate for a two-item test is always twice the log-odds (LOE) or conditional (FCON) estimate, a result previously deduced by Andersen (1973) for the asymptotic case. This result also agrees with the correction for bias suggested by Wright and Douglas (1976), namely (L-1)/L, where L is the number of items. In this case the correction factor would be 0.5, and the corrected UCON estimate would agree with the LOE estimate.

### 9.2.5 The extra-conditional estimator (XCON)

The maximum likelihood condition for the general dichotomous model is

(9.13) $$S_i = \sum_{n=1}^{N} \left( \frac{P_{nil} - V_{nil}}{1 - V_{ni0} - V_{nil}} \right)$$

where $S_i$ is N10 for item 1, and N01 for item 2, and the other notation is that of chapters 7 and 8.

In the two-facet case here, the terms become, for parameter estimate, $D_i'$, which is $D_1'$ or $D_2'$, when all N persons have the same measure, B', corresponding to a score of 1,

$$V_{ni0} = V_{i0} + V_{n0} - V_{i0}*V_{n0}/P_{ni0} \quad (9.14)$$

$$V_{ni1} = V_{i1} + V_{n1} - V_{i1}*V_{n1}/P_{ni1} \quad (9.15)$$

where

$$V_{i1} = P_{ni1}^{N} \quad (9.16)$$

$$V_{i0} = P_{ni0}^{N} \quad (9.17)$$

$$V_{n0} = P_{n10} * P_{n20} \quad (9.18)$$

$$V_{n1} = P_{n11} * P_{n21} \quad (9.19)$$

This is not susceptible to the simple manipulation of the other estimators. But, as the number of persons, N, becomes large, $V_{i1}$ and $V_{i0}$ become very small. Therefore, the asymptotic value of $D_i'$ as N becomes large satisfies the approximate simplified equation,

$$S_i \approx N*\left(\frac{P_{ni1} - P_{ni1}*P_{n21}}{1 - P_{n10}*P_{n20} - P_{n11}*P_{n21}}\right) \quad (9.20)$$

Substituting for i, and the observed marginal scores of the items, the maximum likelihood conditions for the two items become

$$\frac{N10}{N} \approx \frac{P_{n10}*P_{n21}}{1 - P_{n10}*P_{n20} - P_{n11}*P_{n21}} \quad (9.21)$$

$$\frac{N01}{N} \approx \frac{P_{n10}*P_{n21}}{1 - P_{n10}*P_{n20} - P_{n11}*P_{n21}} \quad (9.22)$$

dividing these gives

$$\frac{N01}{N10} \approx \frac{P_{n10}*P_{n21}}{P_{n11}*P_{n20}} \quad (9.23)$$

But, since $P_{n11}/P_{n10} = \exp(B'-D_1')$ and $P_{n21}/P_{n20} = \exp(B'-D_2')$ by the definition of the Rasch model,

$$N01/N10 \approx \exp(B'-D_2')/\exp(B'-D_1') \quad (9.24)$$

Rearranging, it is seen that this agrees with the LOE estimator,

$$D_1' - D_2' \approx \log(N01/N10) \quad (9.25)$$

## 9.3 COMPARISON OF ESTIMATORS FOR THE TWO-ITEM TEST

Fig. 9.2 shows the possible estimable outcomes when two persons take two items. Under these circumstances, all the estimators under consideration estimate D1' - D2' to be zero.

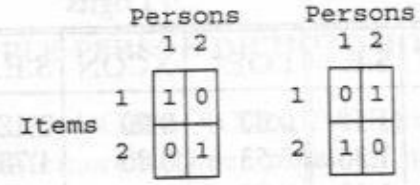


Fig. 9.2 Possible outcomes of two persons taking two items

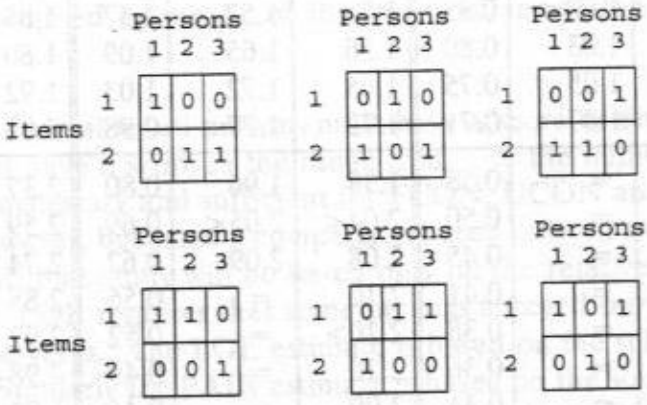


Fig. 9.3 Possible outcomes for three persons taking two items.

The possible data sets for three persons taking a two item test are shown in fig. 9.3. It can be seen that the person score is always 1, and so all persons are estimated to have the same ability. The item scores can be either 1 or 2. Over numerous replications of the same two-item three-person test, the frequency of occurrence of the different possible outcomes is expected to follow a binomial distribution defined by the latent parameter values of the Rasch model. Thus, if the difficulty of item 1, D1, is 1 logit, and the difficulty of item 2, D2, is -1 logit, and all three persons have ability, B, of 0 logits, then, the empirical data set has a 93% probability of being inestimable (a person or an item has an extreme score vector). Of the estimable data sets, item 1 has an 88% probability of an empirical score of 1 and a 12% probability of a score of 2. When item 1 has a score of 1 in an estimable data set, then item 2 has a score of 2 and vice versa. The estimated difference D1' - D2' is thus determined by weighting its value for each possible estimable outcome by the probability of that outcome. Only estimable outcomes are considered.

As the number of persons increase, the number of possible data sets also increase, but their asymptotic distribution is still expected to follow a binomial distribution based on the latent parameters of the persons and items.

TABLE 2
BIAS IN ESTIMATORS FOR A TWO ITEM TEST

| Number of person ratings | Generating Difference | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1.0 Logits | | | 2.0 Logits | | | 3.0 Logits | | |
| | LOE* | XCON | S.E. | LOE* | XCON | S.E. | LOE* | XCON | S.E. |
| 2 | 0.00 | 0.00 | 1.59 | 0.00 | 0.00 | 2.18 | 0.00 | 0.00 | 3.33 |
| 3 | 0.32 | 0.48 | 1.30 | 0.53 | 0.80 | 1.78 | 0.63 | 0.95 | 2.72 |
| 4 | 0.56 | 0.73 | 1.13 | 0.88 | 1.14 | 1.54 | 1.02 | 1.32 | 2.35 |
| 5 | 0.73 | 0.87 | 1.01 | 1.13 | 1.34 | 1.38 | 1.29 | 1.53 | 2.10 |
| 6 | 0.85 | 0.95 | 0.92 | 1.31 | 1.47 | 1.26 | 1.50 | 1.68 | 1.92 |
| 7 | 0.93 | 1.00< | 0.85 | 1.45 | 1.57 | 1.17 | 1.66 | 1.81 | 1.78 |
| 8 | 0.98 | 1.03 | 0.80 | 1.56 | 1.65 | 1.09 | 1.80 | 1.91 | 1.66 |
| 9 | 1.02< | 1.05 | 0.75 | 1.65 | 1.72 | 1.03 | 1.92 | 2.01 | 1.57 |
| 10 | 1.04 | 1.07 | 0.71 | 1.72 | 1.77 | 0.98 | 2.02 | 2.09 | 1.49 |
| 15 | 1.08> | = | 0.58 | 1.94 | 1.96 | 0.80 | 2.37 | 2.40 | 1.21 |
| 20 | 1.07 | = | 0.50 | 2.04< | 2.05< | 0.69 | 2.59 | 2.61 | 1.05 |
| 25 | 1.05 | = | 0.45 | 2.08 | 2.09 | 0.62 | 2.74 | 2.75 | 0.94 |
| 30 | 1.04 | = | 0.41 | 2.10 | = | 0.56 | 2.85 | = | 0.86 |
| 35 | 1.04 | = | 0.38 | 2.10> | = | 0.52 | 2.92 | = | 0.80 |
| 40 | 1.03 | = | 0.36 | 2.10 | = | 0.49 | 2.98 | = | 0.74 |
| 45 | 1.03 | = | 0.34 | 2.09 | = | 0.46 | 3.02< | = | 0.70 |
| 50 | 1.03 | = | 0.32 | 2.08 | = | 0.44 | 3.05 | = | 0.67 |
| 100 | 1.01 | = | 0.23 | 2.04 | = | 0.31 | 3.10> | = | 0.47 |
| 150 | 1.01 | = | 0.18 | 2.03 | = | 0.25 | 3.08 | = | 0.38 |

Table 2 gives the bias of the estimators. The Table was produced by determining the estimated difference between the difficulties of the two items corresponding to every possible estimable outcome for the number of person replications. The estimated difference for each different outcome was weighted by the probability of that outcome, based on the generating parameter values.

In Table 2, estimator bias shown by the ratio of each estimate to an unbiased estimate. Thus an unbiased estimate is shown as 1.00, an overestimate is greater than 1.00, and an underestimate is less than 1.00. The column headed LOE* represents the estimates given by LOE, FCON, PAIR and corrected UCON estimators. S.E. is a modelled standard error based on the generating parameters. "=" indicates that the XCON estimate is numerically indistinguishable from the LOE estimate. ">" indicates transition from underestimate to overestimate of generating difference. ">" indicates the maximum overestimate.

As can be seen in Table 2, for small numbers of person replications the estimators underestimate the difference between the items. For larger numbers, there is an overestimate which diminishes asymptotically. In all cases the size of the bias is less than a modelled standard error, so that the implications of the estimates are not significantly different from those of the latent parameters.

## 9.4 THREE-ITEM MULTIPLE-PERSON DICHOTOMOUS TWO-FACET TEST

The five estimators (LOE, FCON, UCON, XCON, PAIR) used for the two item test can be applied to estimating the difference between a pair of items in the context of a three item test. Though closed form solutions exist for certain combinations of generators and for certain estimators, the three item test is generally intractable to direct analysis. There are, however, only a finite number of different possible data sets, and their asymptotic frequency can be determined by means of the latent values of the parameters and the binomial distribution.

For the data set to be estimable, all persons must have a score of either 1 or 2, and each of the three items must have a score in the range from 1 to the number of persons less 1. These constraints are necessary and sufficient for FCON, UCON and XCON. The LOE log-odds estimator, however, becomes a comparison based only on two of the three items. Consequently, since this bias study will be based only on the relative difficulty of the first two of the three items, LOE requires that some persons succeed on the first item and fail on the second, and vice-versa. The LOE estimate is based on the sub-set of outcomes for which it is estimable. Similarly the PAIR estimate is based on the somewhat larger sub-set of outcomes for which it is estimable. Even after applying the technique for missing pair-wise differences suggested in Wright and Masters (1982, 70), this sub-set is still smaller than that for FCON.

Under some circumstances, the XCON algorithm fails to converge in the usual sense of the expected marginal scores being arbitrarily close to the observed marginal scores. This occurs, in general, when the empirical data has a structure such that the log odds of the modelled probabilities, $\log(P_{ni1}/P_{ni0})$, have smaller absolute value than the log odds of the estimable probabilities, $\log(E_{ni1}/E_{ni0})$, based on the currently estimated parameter values. This causes the likelihood function to have a saddle rather than a peak. Though the maximum likelihood condition is not strictly satisfied under such circumstances, XCON can provide estimates corresponding to a minimizing the greatest absolute marginal residual. These are regarded as the XCON estimates for the purposes of this analysis.

As with the two item test, estimates were obtained corresponding to all estimable outcomes for three item test, and weighted according to their probability of occurrence based on the generating parameters. For this analysis, the three items were given difficulties of .7, -.3, and -.4. The difference to be estimated is that between the first two items, which is 1 logit. Each simulation contained an even number of persons, half were given an ability of 0 logits, and half an ability of 1 logit. The items are thus non-uniformly distributed, and the persons

are not centered on the test. This was purposely done to avoid giving any estimator unusually favorable conditions.

TABLE 3
BIAS IN ESTIMATORS FOR A THREE ITEM TEST

| Number of person replications | Generating difference = 1.00 Logits Rasch Estimator | | | | | |
|---|---|---|---|---|---|---|
| | LOE | PAIR | FCON | UCON | XCON | S.E. |
| 3 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.30 |
| 4 | 0.33 | 0.39 | 0.52 | 0.54 | 0.67 | 1.13 |
| 6 | 0.59 | 0.70 | 0.83 | 0.87 | 0.91 | 0.92 |
| 8 | 0.77 | 0.88 | 0.98 | 1.03< | 1.00< | 0.80 |
| 10 | 0.89 | 0.98 | 1.05< | 1.09 | 1.06 | 0.71 |
| 12 | 0.97 | 1.03< | 1.07 | 1.12 | 1.08 | 0.65 |
| 14 | 1.02< | 1.06 | 1.08 | 1.12 | 1.08 | 0.60 |
| 16 | 1.05 | 1.08 | 1.08> | 1.12> | 1.08> | 0.56 |
| 18 | 1.07 | 1.09 | 1.08 | 1.12 | 1.08 | 0.53 |
| 20 | 1.08> | 1.10> | 1.08 | 1.12 | 1.08 | 0.50 |
| 22 | 1.07 | 1.08 | 1.06 | 1.09 | 1.06 | 0.48 |

Table 3 reports the results. The UCON estimates are corrected by a factor of 2/3 (Wright and Douglas 1976). S.E. is a modelled standard error based on the generating parameters. "<" indicates transition from underestimate to overestimate of the generating difference. ">" indicates the maximum overestimate.

The bias in all the estimators is smaller than the modelled standard error, and the same pattern of underestimation followed by overestimation is observed as in the two-item test. The size of the bias, though larger than for the equivalent two-item test, is still much smaller than a modelled standard error based on the generating parameters.

## 9.5 COMPARISON OF ESTIMATORS FOR LARGER AND MANY-FACET TESTS

The extreme cases of two-facet two-item and three-item tests, already presented, indicate that the problem of bias is manageable. In general, for all the estimators, the greater the number of observations, the smaller the probability of extreme vectors and consequently the smaller the bias. Only for UCON is the bias so large that a correction factor is required.

In order to investigate the behavior of UCON in estimating from adverse many-facet data sets. A simulation study was conducted of which the results are reported in Table 4. Uncorrected UCON estimates of all parameters were obtained from 100 simulated data sets of dichotomous observations, generated using parameters uniformly distributed across 2 logits for each facet. The bias reported is the regression coefficient obtained when all

estimators are regressed on their generators. Again 1.00 indicates an unbiased parameter. Greater than 1.00 is an overestimate, and less than 1.00 is an underestimate.

When there is a very small number of parameters and facets, the generating parameters are underestimated, but, in general, the parameters are overestimated. As the data sets become larger, the trend is for the bias to diminish as the number of facets increases, and also as the number of parameters per facet increases. Most importantly for many-facet measurement, the addition of an extra facet, even with only two parameters in it, tends to reduce the overall bias of the estimates. For most data sets of a size encountered in practice, the bias in UCON estimates is negligible, obviating the need to use a more computationally intensive algorithm. In situations when the effect of estimator bias may be meaningful, the XCON algorithm is a practical alternative.

TABLE 4
MEAN BIAS OF UNCORRECTED UCON ESTIMATES

| Number of facets | Number of parameters per facet | Bias of uncorrected UCON Estimates (Unbiased = 1.00) | Bias in one more facet with 2 parameters |
|---|---|---|---|
| 2 | 2 | 0.00 | |
| 3 | 2 | 2.69 | |
| 4 | 2 | 3.14 | |
| 5 | 2 | 2.08 | |
| 6 | 2 | 1.23 | |
| 7 | 2 | 1.15 | |
| 8 | 2 | 1.05 | |
| 2 | 3 | 0.40 | 3.53 |
| 3 | 3 | 2.16 | 1.35 |
| 4 | 3 | 1.20 | 1.12 |
| 5 | 3 | 1.05 | 1.03 |
| 6 | 3 | 1.02 | 1.02 |
| 2 | 4 | 0.74 | 1.92 |
| 3 | 4 | 1.30 | 1.10 |
| 4 | 4 | 1.10 | 1.04 |
| 5 | 4 | 1.03 | 1.01 |
| 2 | 5 | 1.01 | 1.41 |
| 3 | 5 | 1.16 | 1.08 |
| 4 | 5 | 1.02 | 1.03 |
| 2 | 6 | 1.26 | 1.30 |
| 3 | 6 | 1.05 | 1.07 |
| 4 | 6 | 1.01 | 1.02 |
| 2 | 7 | 1.25 | 1.26 |
| 3 | 7 | 1.07 | 1.03 |
| 2 | 8 | 1.19 | 1.16 |

# 10 A COMPARATIVE EXAMPLE OF MANY-FACET MEASUREMENT

The many-facet Rasch model facilitates the analysis of judge-awarded ratings by producing measures as free as possible of the particular characteristics of the judging situation. In order to demonstrate this, a data set which has been analyzed previously using the best available technique is here reconsidered.

## 10.1 GUILFORD'S DATA SET

Guilford (1954, 282) gives a data set containing the ratings of the creative performance of seven scientists (examinees) given by three senior scientists (judges) on five traits (items). This data is given in Table 5. The ratings are presented in a somewhat different manner than in Guilford in order to aid in identification of judge behavior.

His Table has been rearranged to show examinees in descending order of performance, and items in descending order of difficulty, left to right. In Table 5, "<" and ">" indicate the most unexplained ratings according to Guilford's model. "*" indicates the most unexpected ratings according to the common scale model, to be described.

## 10.2 A CONVENTIONAL ANALYSIS FOLLOWING GUILFORD

Guilford (1954, 282-288) develops a "true-score" descriptive model based on ratings as linear measures of performance. His model can be expressed in a manner consistent with our notation:

$$X_{ijn} = X_m + X_n - X_i - X_j - X_{ij} - X_{jn} + \epsilon \qquad (10.1)$$

where

$X_{ijn}$ is the observed rating on item i by judge j of examinee n
$X_m$ is the grand mean of all ratings
$X_n$ is the ability of the examinee
$X_i$ is the difficulty of the item
$X_j$ is the severity of the judge: "Leniency error"
$X_{ij}$ is the judge-item interaction
$X_{jn}$ is the judge-examinee interaction: "Halo error"
($X_{in}$ is the item-examinee interaction, not included in Guilford's model)
$\epsilon$ is random error

All terms are calculated in rating score points.

This model goes further than Guilford. In his Tables he calculates numbers equivalent to $X_n$ and $X_i$ but merely refers to them as "deviations." He draws no conclusions about them, even though $X_n$ is essential in making substantive decisions concerning the seven scientists. The results which Guilford reports are summarized in Table 6. The results are expressed in score points, have been adjusted to the model presented here, and corrected for the rounding errors introduced by Guilford's step-wise procedure, which somewhat cloud the nature of his results. Terms are listed in the order of their main effects.

As can be seen from a summary of Guilford's analysis of variance Tables, in Table 7, item difficulties and examinee abilities are highly significant effects. Judge-examinee interaction, which Guilford terms "relative halo effect" is also highly significant.

He states that "within this context [of interactions] the various rater means and trait means represent the base of *objectivity*" (ibid., 284). Unfortunately, the existence of the highly significant judge-examinee interactions indicates that there is no way of knowing how each judge will react to a putative "next" examinee. This fact alone limits the "objectivity" Guilford discovers here to a phenomenon of this data set. Further, even though the model has attempted to model differences between ratings in every non-trivial way, 15% of the total variance in ratings is still unexplained. In Guilford's own analysis, due to his exclusion of the significant $X_{in}$ terms, 29% of the variance is unaccounted for. The rating an examinee gets still have a large component of "luck." The magnitude of this luck, as can be seen from the error residuals, $\varepsilon$, for Guilford's model in Table 8, can be of the order of 2 rating points. 8 of the 105 residuals are 2.0 score points or larger, identified in Table 8 by a "<" designation.

For later comparison, it is useful to record what conclusions Guilford draws from his analysis. In discussing judge-examinee interaction, shown in Table 6, he states that judge A tends to overvalue examinee 5, judge B tends to overvalue examinee 4 and undervalue examinees 5 and 7, and judge C tends to undervalue examinee 4 and overvalue examinee 5. Thus judge B disagrees with judges A and C about examinees 4 and 5. Were judge-item interaction significant, then judge A would tend to see examinees higher on item a and lower on items c and e than the other judges. In the raw data, judge B correlated negatively with both judges A and C, but, after removing explained effects, all judges correlate positively.

## 10.3 A THREE-FACET RASCH ANALYSIS

From the viewpoint of generalizable (i.e. objective) measures, the most desirable model would be one in which each component of the judging situation were represented by only one parameter. Such a model is

$$(10.2) \qquad \log(P_{nijk}/P_{nijk-1}) = B_n - D_i - C_j - F_k$$

where

$P_{nijk}$ and $P_{nijk-1}$ are the probabilities of ratings of k and k-1 respectively

$B_n$ is the ability of examinee n, with n=1,7
$D_i$ is the difficulty of item i, with i=a,e
$C_j$ is the severity of judge j, with j=A,B,C
$F_k$ is the difficulty of category k relative to category k-1, k=2,9

This model specifies that all judges use the rating scale in the same way. As the judges have the rating scale in common, this model is a common-scale model. It is similar to Guilford's equation 10.1 in that each main effect, examinee ability, item difficulty, and judge severity, is represented by a parameter. On the other hand, the interaction terms in equation 10.1 are omitted, and the rating scale is not assumed to be linear.

The ratings are treated as counts of levels of performance exhibited. What performance level each higher rating scale category represents is estimated from the ratings themselves. The structure of the rating scale is thus defined by the way in which the judges used it, and does not require the categories to be equally spaced along a line.

Estimation of parameters by this model produces the results shown in Tables 9 through 13. Before investigating the meaning of the measures, it is important to examine whether, in fact, the ratings do fit this common-scale model. In Tables 9 through 13, the mean-square fit statistics (MnSq) have an expectation of 1, and are generally not alarming. This analysis itself, in fact, has already satisfactorily explained the variation in the ratings since the overall mean square fit statistics and the distribution of the standardized residuals in the summary line at the end of Table 13 are in accordance with model expectations. Thus the measures given in Table 9 are objective estimates of the creativity of the examinees on a linear scale representing the latent variable, with an accuracy reflected in the size of the standard errors. If another examinee were to be judged, or another judge to award ratings, no meaningful change would be expected in these Tables.

There are, however, some aspects of the common-scale results which, if the intention of the analysis were to study judge behavior, would merit closer inspection. In Table 9, examinee 5 has the most unmodelled variation in his ratings. Inspection of Table 7 shows that he was rated high by judges A and C but low by judge B. Examinee 1 has much less variation in his ratings than expected. All of his ratings were in the narrow range of 3 to 6. The supposition that judge B has a different perspective to judges A and C is further supported by Table 13 in which all the most unexpected ratings were awarded by judge B and they comprised 5 out of the 35 he awarded.

The manner in which the judges have used the rating scale is also clearly idiosyncratic. In Table 12, when all three judges are considered together, categories 3,5,7 are used disproportionately often, suggesting a problem in the definition of the rating scale. The reason for this is the manner in which the three individual judges have used the rating scale, which is shown in the last three columns of Table 12. Judges A and C have tended to use the odd-numbered categories, while judge B has tended to use the central categories.

### 10.3.1 A three-facet analysis with judge-scale interaction

Since the previous analysis raises doubts about the judges' use of the rating scale, a second analysis is performed in which each judge is allowed to define his own use of the rating scale. This is a judge-scale model. The model equation is

$$\log(P_{nijk}/P_{nijk-1}) = B_n - D_i - C_j - F_{jk} \quad (10.3)$$

where the terms have the same meaning as in equation 10.2, and

$F_{jk}$ is the difficulty of category k relative to category k-1 for judge j, with the categories renumbered for each judge, $0,K_j$.

To apply this model, it is necessary to renumber the categories for each judge, since all judges did not use all categories. What remains in common is that the categories continue to represent ordered levels of performance, so that each higher category actually awarded by a judge represents the next higher level of performance recognized by that judge. The categories used are renumbered in order of level of performance, and the orders reached in each rating are counted and summed to determine each examinees overall score. The differences between levels of performance, as used by each judge, are what the $F_{jk}$ terms are estimate.

Tables 14 through 18 give the results of this analysis. Table 17 gives the calibrations for each of the three judges' rating scales, in a frame of reference in which the judges are modelled to agree as to the ability of each examinee and the difficulty of each item.

Surprisingly, allowing each judge his own scale definition has not improved the measurement capabilities of the model. In Table 14, the overall fit is worse than in Table 9. Examinee 5 now shows even more unexpected variation in his ratings, while examinee 1 still shows too little variation. The fit of judge B is also less in accord with the model, and, compared with Table 13, one more of his ratings than is listed in Table 18 as quite unexpected.

Nevertheless, as depicted in fig. 10.1, the measures for the examinees are not significantly different for the common-scale and judge-scale models. Thus these measures do have some degree of generalizability. The diagonal appearance of this plot also shows that Guilford's $X_n$ scores are highly correlated with these measures, which is a reminder that, under many circumstances, scores may give the appearance of being linear measures even though they lack fundamental measurement properties.

The common-scale analysis suggests judge B was not in agreement with the other judges as to the definition of the rating scale. The judge-scale analysis reveals, however, that, even after allowing each judge to define his own scale and so removing the constraint of a common scale, the disagreement by judge B is yet more pronounced. This shows that his disagreement can not be explained as a quirk of the rating scale categories. His disagreement is more fundamental - it concerns a substantially different view of the

performance of the examinees. On the other hand, there has been no indication that judges A and C are not in accord.

### 10.3.2 Modelling the judges in two separate groups

In an attempt to determine whether judge B is self-consistent, a separate analysis is done of judge B by himself, shown in Tables 19 through 23, and similarly of judges A and C together, Tables 24 through 28. When the judges are calibrated separately, a judge sub-scale equating step is required in order to compare their severities. This has not been done here, and so measures in the two sets of Tables have different local origins. The logit differences between estimates within the same Table, however, are comparable to the same differences in the corresponding Table in the other set. The analysis of judge B by himself indicates that apart from the unexpectedly low rating he gave to examinee 2 on item e, he is generally self-consistent. Closer inspection of judges A and C defining a joint scale together, indicates exceptionally close agreement. They even agree in having trouble on the variability of examinee 3's performance level, as shown in Table 28.

We now see that judge B rates the examinees from one viewpoint, while judges A and C rate them from another. This can be clearly seen in fig. 10.2. The disagreement with judges A and C stems from the fact that judge B perceives the examinees, except 2 and 5, to be much the same. It is as though judge B is measuring their weight, while judges A and C are measuring their height. A test can be made of the hypothesis that the pairs of measures for each examinee are statistically equivalent, apart from the arbitrary placement of the origins of the scales. This yields a chi-square of homogeneity of 7.5 with 6 degrees of freedom, which has a .28 probability of being exceeded. Thus the differences between the measures given by the judges are not statistically significant, though they are substantially different in terms of any decisions that are to be made based on the measures.

## 10.4 CONCLUSIONS OF THE EXAMPLE ANALYSIS

Treating Guilford's data as an example of a measurement problem rather than a problem of description has revealed both strengths and weaknesses in his data. Guilford assumed the ratings lay along a linear scale, and attempted to explain, and hence eliminate, variance in the ratings by means of many descriptive terms. These terms, particularly the interactions, could have no general meaning beyond the local test situation. Even were the adjusted scores to be regarded as measures, their meaning is limited, as presented by Guilford, because there is no indication of their accuracy, i.e. there are no standard errors. Further the degree to which the adjusted score really summarizes the examinee's performance is unclear, even in the limited context of this set of ratings, since there are no fit statistics.

Comparing the results of the objective measurement process with Guilford's conclusions, Guilford notices that judge B behaves differently from judges A and C, but treats this behavior as specific to certain interactions. In contrast, the measurement analysis shows

that judge B has a different way of using the rating scale, and also a different perception of the examinees than judges A and C. How this is handled for the purposes of measurement, in practical terms, is a the examination board must make, not a quirk of interaction terms.

In spite of the clear anomalies in ratings, the common-scale model provides statistically supportable measures for all the examinees on a linear scale. The avoidance of interaction terms, together with the application of the axioms of objectivity, leads to measures for examinees which can be considered as estimates of their abilities on a general variable of "creative performance," as far as this judging session represents that variable.

TABLE 5
RATINGS OF SEVEN SCIENTISTS (EXAMINEES) BY THREE SENIOR SCIENTISTS (JUDGES) ON FIVE TRAITS (ITEMS)

| Examinee | Hard Item e | Item c | Item b | Item a | Easy Item d |
|---|---|---|---|---|---|
| Judge: | A C B | A C B | A C B | A C B | A C B |
| High 2 | 5 5 2< | 5 5 5 | 7 7 7 | 9 7 8 | 8 7 7 |
| 5 | 5 7 3 | 7 7 3 | 7 7 4 | 9 9 2* | 8 7 2* |
| 7 | 5 7 4 | 5 7 5 | 7 7 3 | 7 7 3 | 5 5 5 |
| 1 | 3 3 3 | 3 5 4 | 5 5 5 | 5 5 6 | 5 7 6 |
| 3 | 1 5 6* | 3 5 3 | 3 5 5 | 3 3 4 | >7 5 6 |
| 4 | 3 1 5* | 1 3 4 | 3 3 6 | 7 5 5 | 3 3 5 |
| Low 6 | 1 3 2 | 3 3 6* | 5 3 4 | 3 3 4 | 5 5 4 |

"<" and ">" are the most unexplained ratings according Guilford's model. "*" are the most unexpected according to the common-scale model. Source: Guilford (1954, 282).

TABLE 6
VALUES OF TERMS FOR GUILFORD'S MODEL

| $X_m$ | grand mean of all the ratings |
|---|---|
| | 4.84 |

Xn the abilities of the examinees

| Ex. 2 | Ex. 5 | Ex. 7 | Ex. 1 | Ex. 3 | Ex. 4 | Ex. 6 |
|---|---|---|---|---|---|---|
| 1.43 | 0.96 | 0.63 | -0.17 | -0.57 | -1.04 | -1.24 |

Xi the difficulties of the items

| Item e | Item c | Item b | Item a | Item d |
|---|---|---|---|---|
| 1.08 | 0.46 | -0.30 | -0.59 | -0.64 |

Xj severities of the judges

| Judge A | Judge C | Judge B |
|---|---|---|
| -0.05 | -0.33 | 0.38 |

Xjn judge-examinee halo error

| Examinee | Judge A | Judge C | Judge B |
|---|---|---|---|
| 2 | -0.49 | 0.40 | 0.09 |
| 5 | -1.35 | -1.27 | 2.62 |
| 7 | -0.29 | -0.80 | 1.09 |
| 1 | 0.51 | 0.00 | -0.51 |
| 3 | 0.91 | 0.00 | -0.91 |
| 4 | 0.45 | 1.13 | -1.58 |
| 6 | 0.25 | 0.53 | -0.78 |

Xij the judge-item interaction

| | Item e | Item c | Item b | Item a | Item d |
|---|---|---|---|---|---|
| Judge A | 0.52 | 0.57 | -0.10 | -0.67 | -0.33 |
| Judge C | -0.33 | -0.29 | 0.19 | 0.19 | 0.24 |
| Judge B | -0.19 | -0.29 | -0.10 | 0.48 | 0.10 |

TABLE 7
GUILFORD'S ANALYSIS OF VARIANCE

| Source | Sum of squares | Degrees of freedom | Variance | F-ratio | P |
|---|---|---|---|---|---|
| $X_j$ | 9.05 | 2 | 4.52 | 3.83 | <.05 |
| $X_i$ | 46.53 | 4 | 11.63 | 9.86 | <.01 |
| $X_n$ | 94.92 | 6 | 15.82 | 13.41 | <.01 |
| $X_{ij}$ | 12.96 | 8 | 1.62 | 1.37 | >.05 |
| $X_{jn}$ | 98.68 | 12 | 8.22 | 6.97 | <.01 |
| $X_{in}$ | 51.47 | 24 | 2.14 | 1.81 | <.05 |
| $\varepsilon$ | 56.64 | 48 | 1.18 | | |
| Total | 370.25 | 104 | | | |

TABLE 8
ERROR RESIDUALS FOR GUILFORD'S MODEL

| Examinee | Item e | | | Item c | | | Item b | | | Item a | | | Item d | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Judge: | A | C | B | A | C | B | A | C | B | A | C | B | A | C | B |
| 2 | -0.2 | -0.5 | -2.9< | -0.8 | -1.0 | -0.6 | -0.2 | 0.7 | 0.8 | 0.9 | 0.4 | 2.1< | 0.2 | 0.4 | 0.7 |
| 5 | -0.6 | 0.3 | 1.1 | 0.8 | -0.2 | 0.4 | -0.6 | -0.5 | 0.8 | 0.5 | 1.2 | -0.9 | -0.2 | -0.8 | -1.3 |
| 7 | 0.8 | 1.1 | 0.9 | 0.2 | 0.6 | 1.2 | 0.8 | 0.3 | -1.4 | -0.1 | -0.0 | -1.1 | -1.8 | -2.0< | 0.5 |
| 1 | 0.4 | -1.3 | -0.9 | -0.2 | 0.2 | -0.6 | 0.4 | -0.1 | -0.2 | -0.5 | -0.4 | 1.1 | -0.2 | 0.7 | 0.7 |
| 3 | -0.8 | 1.1 | 2.1< | 0.6 | 0.6 | -1.6 | -0.8 | 0.3 | -0.2 | -1.7 | -2.0< | -0.9 | 2.6< | -0.0 | 0.7 |
| 4 | 1.2 | -1.3 | 0.9 | -1.4 | 0.2 | -0.8 | -0.8 | -0.1 | 0.6 | 2.3< | 1.6 | -0.1 | -1.4 | -0.4 | -0.5 |
| 6 | -0.8 | 0.3 | -1.1 | 0.6 | -0.2 | 2.2< | 1.2 | -0.5 | -0.4 | -1.7 | -0.8 | -0.1 | 0.6 | 1.2 | -0.5 |

TABLE 9
EXAMINEE MEASURES FOR COMMON-SCALE MODEL

| | Score | Count | Measure Logit | Model Error | Infit MnSq | Outfit MnSq |
|---|---|---|---|---|---|---|
| Examinee 2 | 79 | 15 | 0.66 | 0.18 | 0.6 | 0.6 |
| Examinee 5 | 72 | 15 | 0.44 | 0.18 | 2.0 | 2.0 |
| Examinee 7 | 67 | 15 | 0.29 | 0.17 | 0.9 | 0.9 |
| Examinee 1 | 55 | 15 | -0.07 | 0.17 | 0.2 | 0.2 |
| Examinee 3 | 49 | 15 | -0.26 | 0.18 | 1.2 | 1.3 |
| Examinee 4 | 42 | 15 | -0.48 | 0.18 | 1.3 | 1.4 |
| Examinee 6 | 39 | 15 | -0.58 | 0.19 | 0.7 | 0.8 |
| Mean: | 57.6 | 15.0 | 0.00 | 0.18 | 1.0 | 1.0 |
| S.D.: | 14.3 | 0.0 | 0.44 | 0.00 | 0.5 | 0.5 |

## TABLE 10
## ITEM CALIBRATIONS FOR COMMON-SCALE MODEL

| | Score | Count | Calib. Logit | Model Error | Infit MnSq | Outfit MnSq |
|---|---|---|---|---|---|---|
| Item e | 58 | 21 | 0.52 | 0.16 | 1.3 | 1.3 |
| Item c | 71 | 21 | 0.21 | 0.15 | 0.8 | 0.8 |
| Item b | 87 | 21 | -0.15 | 0.15 | 0.6 | 0.6 |
| Item a | 93 | 21 | -0.28 | 0.15 | 1.4 | 1.4 |
| Item d | 94 | 21 | -0.30 | 0.15 | 0.9 | 0.9 |
| Mean: | 80.6 | 21.0 | 0.00 | 0.15 | 1.0 | 1.0 |
| S.D.: | 14.0 | 0.0 | 0.32 | 0.00 | 0.3 | 0.3 |

## TABLE 11
## JUDGE CALIBRATIONS FOR COMMON-SCALE MODEL

| | Score | Count | Calib. Logit | Model Error | Infit MnSq | Outfit MnSq |
|---|---|---|---|---|---|---|
| Judge B | 121 | 35 | 0.24 | 0.12 | 1.5 | 1.5 |
| Judge A | 136 | 35 | 0.04 | 0.12 | 0.8 | 0.9 |
| Judge C | 146 | 35 | -0.10 | 0.12 | 0.7 | 0.7 |
| Mean: | 134.3 | 35.0 | 0.06 | 0.12 | 1.0 | 1.0 |
| S.D.: | 10.3 | 0.0 | 0.14 | 0.00 | 0.3 | 0.4 |

## TABLE 12
## CATEGORY CALIBRATIONS FOR COMMON-SCALE MODEL

| Use of categories by all Judges | | | | | | Use of categories by Judges | | |
|---|---|---|---|---|---|---|---|---|
| Cat | Step | Count | % | Logit | S.E. | Judge A | Judge B | Judge C |
| 1 | 0 | 4 | 4 | | | 3 | | 1 |
| 2 | 1 | 4 | 4 | -0.69 | 0.54 | | 4 | |
| 3 | 2 | 25 | 24 | -2.35 | 0.40 | 10 | 6 | 9 |
| 4 | 3 | 8 | 8 | 0.81 | 0.25 | | 8 | |
| 5 | 4 | 31 | 30 | -1.48 | 0.24 | 11 | 8 | 12 |
| 6 | 5 | 6 | 6 | 1.72 | 0.25 | | 6 | |
| 7 | 6 | 21 | 20 | -0.98 | 0.26 | 7 | 2 | 12 |
| 8 | 7 | 3 | 3 | 2.39 | 0.45 | 2 | 1 | |
| 9 | 8 | 3 | 3 | 0.58 | 0.61 | 2 | | 1 |

## TABLE 13
## MOST UNEXPECTED RATINGS FOR COMMON-SCALE MODEL

| Facets producing rating | | | Cat | Step | Exp. | Resd | StRes |
|---|---|---|---|---|---|---|---|
| Judge B | Examinee 3 | Item e | 6 | 5 | 1.9 | 3.1 | 2 |
| Judge B | Examinee 4 | Item e | 5 | 4 | 1.6 | 2.4 | 2 |
| Judge B | Examinee 5 | Item a | 2 | 1 | 5.0 | -4.0 | -2 |
| Judge B | Examinee 5 | Item d | 2 | 1 | 5.1 | -4.1 | -2 |
| Judge B | Examinee 6 | Item c | 6 | 5 | 1.9 | 3.1 | 2 |
| For all residuals: | | Mean: | 4.8 | 3.8 | 3.8 | 0.0 | 0.0 |
| Count = 105 | | S.D.: | 1.9 | 1.9 | 1.2 | 1.4 | 1.0 |

## TABLE 14
## EXAMINEE MEASURES FOR JUDGE-SCALE MODEL

| | Score | Count | Measure Logit | Model Error | Infit MnSq | Outfit MnSq |
|---|---|---|---|---|---|---|
| Examinee 2 | 48 | 15 | 0.91 | 0.24 | 0.9 | 0.7 |
| Examinee 5 | 37 | 15 | 0.27 | 0.24 | 2.4 | 2.2 |
| Examinee 7 | 36 | 15 | 0.21 | 0.24 | 0.9 | 0.8 |
| Examinee 1 | 32 | 15 | -0.04 | 0.25 | 0.2 | 0.2 |
| Examinee 3 | 29 | 15 | -0.22 | 0.25 | 1.1 | 1.1 |
| Examinee 4 | 26 | 15 | -0.42 | 0.26 | 1.2 | 1.3 |
| Examinee 6 | 22 | 15 | -0.71 | 0.27 | 0.9 | 0.9 |
| Mean: | 32.9 | 15.0 | 0.00 | 0.25 | 1.1 | 1.0 |
| S.D.: | 7.9 | 0.0 | 0.49 | 0.01 | 0.6 | 0.6 |

## TABLE 15
## ITEM CALIBRATIONS FOR JUDGE-SCALE MODEL

| | Score | Count | Calib. Logit | Model Error | Infit MnSq | Outfit MnSq |
|---|---|---|---|---|---|---|
| Item e | 31 | 21 | 0.70 | 0.23 | 1.6 | 1.4 |
| Item c | 40 | 21 | 0.25 | 0.22 | 0.9 | 0.9 |
| Item b | 50 | 21 | -0.19 | 0.21 | 0.7 | 0.6 |
| Item a | 54 | 21 | -0.36 | 0.21 | 1.6 | 1.4 |
| Item d | 55 | 21 | -0.40 | 0.21 | 0.9 | 0.8 |
| Mean: | 46.0 | 21.0 | -0.00 | 0.21 | 1.1 | 1.0 |
| S.D.: | 9.2 | 0.0 | 0.42 | 0.01 | 0.4 | 0.3 |

## TABLE 16
## JUDGE CALIBRATIONS FOR JUDGE-SCALE MODEL

| | Score | Count | Calib. Logit | Model Error | Infit MnSq | Outfit MnSq |
|---|---|---|---|---|---|---|
| Judge B | 86 | 35 | 0.37 | 0.14 | 1.6 | 1.7 |
| Judge A | 71 | 35 | 0.22 | 0.16 | 0.6 | 0.6 |
| Judge C | 73 | 35 | -0.03 | 0.21 | 0.8 | 0.8 |
| Mean: | 76.7 | 35.0 | 0.19 | 0.17 | 1.0 | 1.0 |
| S.D.: | 6.6 | 0.0 | 0.16 | 0.03 | 0.4 | 0.5 |

## TABLE 17
## CATEGORY CALIBRATIONS FOR JUDGE-SCALE MODEL

| Cat | Judge A Step | Count | Logit | S.E. | Judge B Step | Count | Logit | S.E. | Judge C Step | Count | Logit | S.E. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 3 | | | | | | | 0 | 1 | | |
| 2 | | | | | 0 | 4 | | | | | | |
| 3 | 1 | 10 | -1.87 | 0.63 | 1 | 6 | -1.29 | 0.58 | 1 | 9 | -2.68 | 1.03 |
| 4 | | | | | 2 | 8 | -0.91 | 0.43 | | | | |
| 5 | 2 | 11 | -0.46 | 0.39 | 3 | 8 | -0.36 | 0.40 | 2 | 12 | -0.46 | 0.41 |
| 6 | | | | | 4 | 6 | 0.20 | 0.44 | | | | |
| 7 | 3 | 7 | 0.38 | 0.41 | 5 | 2 | 1.27 | 0.66 | 3 | 12 | 0.16 | 0.38 |
| 8 | 4 | 2 | 1.47 | 0.59 | 6 | 1 | 1.09 | 1.06 | | | | |
| 9 | 5 | 2 | 0.48 | 0.79 | | | | | 4 | 1 | 2.97 | 1.03 |

## TABLE 18
## MOST UNEXPECTED RATINGS FOR JUDGE-SCALE MODEL

| Facets producing rating | | | Cat | Step | Exp. | Resd | StRes |
|---|---|---|---|---|---|---|---|
| Judge B | Examinee 2 | Item e | 2 | 0 | 2.8 | -2.8 | -2 |
| Judge B | Examinee 3 | Item e | 6 | 4 | 1.1 | 2.9 | 2 |
| Judge B | Examinee 4 | Item e | 5 | 3 | 0.9 | 2.1 | 2 |
| Judge B | Examinee 5 | Item a | 2 | 0 | 3.4 | -3.4 | -2 |
| Judge B | Examinee 5 | Item d | 2 | 0 | 3.5 | -3.5 | -2 |
| Judge B | Examinee 6 | Item c | 6 | 4 | 1.1 | 2.9 | 2 |
| For all residuals | | Mean: | 4.8 | 2.2 | 2.2 | -0.0 | 0.0 |
| Count = 105 | | S.D.: | 1.9 | 1.3 | 0.7 | 1.1 | 1.0 |

## TABLE 19
## EXAMINEE MEASURES FOR JUDGE B ALONE

| | Score | Count | Measure Logit | Model Error | Infit MnSq | Outfit MnSq |
|---|---|---|---|---|---|---|
| Examinee 2 | 19 | 5 | 1.02 | 0.38 | 1.9 | 1.9 |
| Examinee 4 | 15 | 5 | 0.45 | 0.38 | 0.3 | 0.3 |
| Examinee 1 | 14 | 5 | 0.30 | 0.38 | 0.4 | 0.4 |
| Examinee 3 | 14 | 5 | 0.30 | 0.38 | 1.3 | 1.3 |
| Examinee 6 | 10 | 5 | -0.30 | 0.40 | 1.0 | 1.1 |
| Examinee 7 | 10 | 5 | -0.30 | 0.40 | 0.9 | 0.9 |
| Examinee 5 | 4 | 5 | -1.48 | 0.53 | 0.9 | 0.9 |
| Mean: | 12.3 | 5.0 | 0.00 | 0.41 | 1.0 | 1.0 |
| S.D.: | 4.4 | 0.0 | 0.74 | 0.05 | 0.5 | 0.5 |

## TABLE 20
## ITEM CALIBRATIONS FOR JUDGE B ALONE

| | Score | Count | Calib. Logit | Model Error | Infit MnSq | Outfit MnSq |
|---|---|---|---|---|---|---|
| Item e | 11 | 7 | 0.73 | 0.36 | 2.2 | 2.0 |
| Item c | 16 | 7 | 0.12 | 0.34 | 1.2 | 1.1 |
| Item a | 18 | 7 | -0.10 | 0.33 | 0.9 | 0.9 |
| Item b | 20 | 7 | -0.32 | 0.33 | 0.4 | 0.5 |
| Item d | 21 | 7 | -0.43 | 0.33 | 0.3 | 0.4 |
| Mean: | 17.2 | 7.0 | 0.00 | 0.34 | 1.0 | 1.0 |
| S.D.: | 3.5 | 0.0 | 0.41 | 0.01 | 0.7 | 0.6 |

## TABLE 21
## JUDGE CALIBRATION FOR JUDGE B ALONE

| | Score | Count | Calib. Logit | Model Error | Infit MnSq | Outfit MnSq |
|---|---|---|---|---|---|---|
| Judge B | 86 | 35 | 0.40 | 0.15 | 1.0 | 1.0 |

## TABLE 22
## CATEGORY CALIBRATIONS FOR JUDGE B ALONE

| Cat | Step | Count | Logit | S.E. |
|---|---|---|---|---|
| 1 | | | | |
| 2 | 0 | 4 | | |
| 3 | 1 | 6 | -1.63 | 0.62 |
| 4 | 2 | 8 | -0.98 | 0.47 |
| 5 | 3 | 8 | -0.28 | 0.41 |
| 6 | 4 | 6 | 0.33 | 0.45 |
| 7 | 5 | 2 | 1.38 | 0.66 |
| 8 | 6 | 1 | 1.18 | 1.06 |
| 9 | | | | |

## TABLE 23
## MOST UNEXPECTED RATINGS FOR JUDGE B ALONE

| Facets producing rating | | Cat | Step | Exp. | Resd | StRes |
|---|---|---|---|---|---|---|
| Judge B Examinee 2 Item e | | 2 | 0 | 2.8 | -2.8 | -2 |
| For all residuals | Mean: | 4.5 | 2.5 | 2.5 | -0.0 | 0.0 |
| Count = 35 | S.D.: | 1.5 | 1.5 | 1.0 | 1.1 | 1.0 |

## TABLE 24
## EXAMINEE MEASURES FOR JUDGES A AND C TOGETHER

| | Score | Count | Measure Logit | Model Error | Infit MnSq | Outfit MnSq |
|---|---|---|---|---|---|---|
| Examinee 5 | 34 | 10 | 3.60 | 0.42 | 0.5 | 0.5 |
| Examinee 2 | 29 | 10 | 2.62 | 0.48 | 1.0 | 0.8 |
| Examinee 7 | 26 | 10 | 1.80 | 0.56 | 1.0 | 1.1 |
| Examinee 1 | 18 | 10 | -0.74 | 0.56 | 0.4 | 0.4 |
| Examinee 3 | 15 | 10 | -1.69 | 0.56 | 2.0 | 2.1 |
| Examinee 6 | 12 | 10 | -2.64 | 0.56 | 0.8 | 0.8 |
| Examinee 4 | 11 | 10 | -2.96 | 0.57 | 1.2 | 1.2 |
| Mean: | 20.7 | 10.0 | 0.00 | 0.53 | 1.0 | 1.0 |
| S.D.: | 8.3 | 0.0 | 2.45 | 0.05 | 0.5 | 0.5 |

## TABLE 25
## ITEM CALIBRATIONS FOR JUDGES A AND C TOGETHER

| | Score | Count | Calib. Logit | Model Error | Infit MnSq | Outfit MnSq |
|---|---|---|---|---|---|---|
| Item e | 20 | 14 | 1.95 | 0.49 | 1.1 | 1.1 |
| Item c | 24 | 14 | 1.01 | 0.48 | 0.7 | 0.7 |
| Item b | 30 | 14 | -0.33 | 0.45 | 0.3 | 0.3 |
| Item d | 34 | 14 | -1.07 | 0.41 | 1.1 | 1.3 |
| Item a | 37 | 14 | -1.56 | 0.40 | 1.5 | 1.6 |
| Mean: | 29.0 | 14.0 | 0.00 | 0.45 | 0.9 | 1.0 |
| S.D.: | 6.3 | 0.0 | 1.30 | 0.04 | 0.4 | 0.4 |

## TABLE 26
## JUDGE CALIBRATIONS FOR JUDGES A AND C TOGETHER

| | Score | Count | Calib. Logit | Model Error | Infit MnSq | Outfit MnSq |
|---|---|---|---|---|---|---|
| Judge A | 71 | 35 | 0.98 | 0.28 | 1.1 | 1.1 |
| Judge C | 74 | 35 | 0.75 | 0.28 | 0.8 | 0.9 |
| Mean: | 72.5 | 35.0 | 0.87 | 0.28 | 1.0 | 1.0 |
| S.D.: | 1.5 | 0.0 | 0.12 | 0.00 | 0.1 | 0.1 |

## TABLE 27
## CATEGORY CALIBRATIONS FOR JUDGES A AND C TOGETHER

| Cat | Step | Count | Logit | S.E. |
|---|---|---|---|---|
| 1 | 0 | 4 | | |
| 2 | | | | |
| 3 | 1 | 19 | -5.72 | 0.60 |
| 4 | | | | |
| 5 | 2 | 23 | -2.49 | 0.38 |
| 6 | | | | |
| 7 | 3 | 19 | 0.48 | 0.41 |
| 8 | 4 | 2 | 4.73 | 0.60 |
| 9 | 5 | 3 | 3.01 | 0.73 |

TABLE 28
MOST UNEXPECTED RATINGS FOR JUDGES A AND C TOGETHER

| | | | Cat | Step | Exp. | Resd | StRes |
|---|---|---|---|---|---|---|---|
| Judge A | Examinee 3 | Item d | 7 | 3 | 1.8 | 1.2 | 2 |
| Judge A | Examinee 4 | Item a | 7 | 3 | 1.5 | 1.5 | 2 |
| Judge C | Examinee 3 | Item e | 5 | 2 | 0.9 | 1.1 | 2 |
| For all residuals: | | Table Mean: | 5.0 | 2.1 | 2.1 | -0.1 | 0.0 |
| Count = 70 | | Table S.D.: | 2.0 | 1.1 | 0.9 | 0.9 | 1.0 |

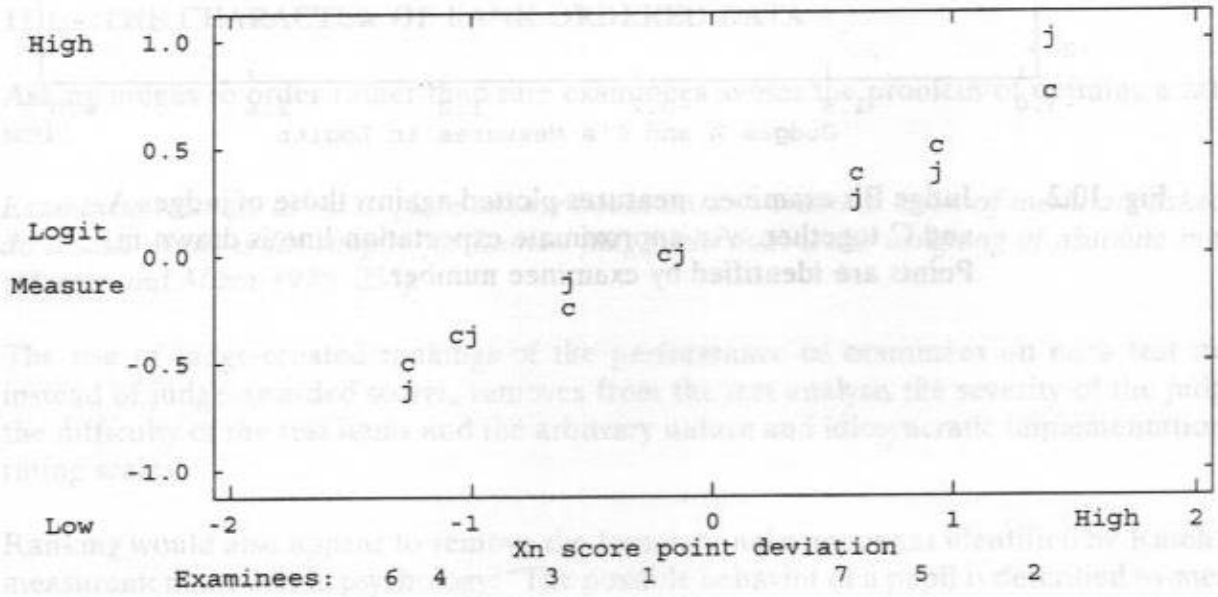


Fig. 10.1 Plot of examinee measures for models including all judges. Common scale, c, and judge-scale, j, measures against Guilford's Xn ability score.

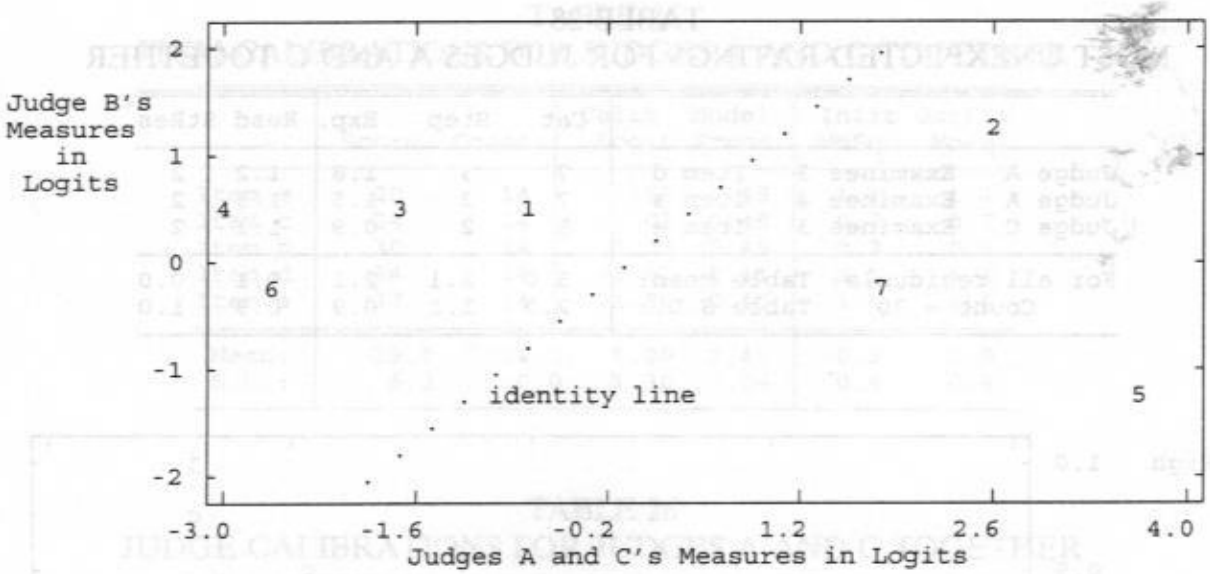


Fig. 10.2 Judge B's examinee measures plotted against those of judges A and C together. An approximate expectation line is drawn in. Points are identified by examinee number.

# 11 RANK ORDERING AS ONE-FACET MEASUREMENT

A special case of the many-facet model is that in which the observations are explicit manifestations of only one facet. Such a situation occurs when objects (e.g. examinees) are rank ordered by judges, rather than awarded ratings. A precise measurement model for rank ordered data is first constructed, and then approximated by a one-facet version of the many-facet model.

## 11.1 THE CHARACTER OF RANK ORDERED DATA

Asking judges to order rather than rate examinees avoids the problem of defining a rating scale.

*Examiners who are asked to place answer books in rank order, or order of merit, are asked to do a task which is far simpler for human judgement than is the assigning of absolute marks (Harper and Misra 1976, 255).*

The use of judge-created rankings of the performance of examinees on each test item, instead of judge-awarded scores, removes from the test analysis the severity of the judges, the difficulty of the test items and the arbitrary nature and idiosyncratic implementation of rating scales.

Ranking would also appear to remove the foundational component identified by Rasch for measurement models in psychology: "The possible behavior of a pupil is described by means of a probability that he solves the task" (Rasch 1980, 11). A ranking of examinees contains no information about the difficulty of the particular item on which they were judged. It does, however, contain information about their relative levels of success.

The stochastic element still remains. It has been observed that judges differ considerably in the rankings they assign:
*The [examinee's performance with] the highest degree of agreement still covered nearly one-third of the range of ranks, while the average [range of ranking a performance] included nearly two-thirds of the available ranks (Harper and Misra 1976, 14).*

It is this variation in the rankings across judges which provides the stochastic element necessary for Rasch measurement. In this we are following Thurstone, who also perceives that the probabilistic element opens the door to measurement.

*It is possible to use the proportion of readers or judges who agree about the rank order of any two statements as a basis for actual measurement (Thurstone and Chave 1929, 18).*

Though Thurstone then presents a method of building a linear scale, based on his Law of Comparative Judgement (Thurstone 1927a, 1927b), his scale lacks a well-defined unit of measurement and other useful statistical properties of the model presented here.

Two measurement models for rankings are now discussed. The first model is derived from the conceptualization of a rank order as a composite of paired comparisons between the objects being ranked. The second model considers a rank order as a rating scale on which, typically, only one object is classified in each rating scale category.

## 11.2 THE OBJECTIVE MEASUREMENT OF PAIRED OBJECTS

As a pre-cursor to the investigation of more comprehensive orderings, consider the ordering of just a pair of objects. A comparison of the performance of two objects (e.g. $O_m$ and $O_n$) across numerous replications of a given agent (e.g. a test item), or of a combination of agents, yields counts of the three possible outcomes:

1) $F_{mn}$, the frequency with which $O_m$ out-performs $O_n$
2) $F_{nm}$ the frequency with which $O_n$ out-performs $O_m$
3) the frequency with which they perform at the same level

For the purposes of this discussion, the discrimination of performance is assumed to be so fine that identical performance levels never occur. (This constraint will be removed later). Thus, following the argument of chapter 5, a comparison of the performance levels of these objects is $F_{mn}/F_{nm}$, which becomes, in the stochastic limit, $P_{mn}/P_{nm}$, where $P_{mn}$ is the probability that $O_m$ out-performs $O_n$ and $P_{nm}$ is the probability that $O_n$ out-performs $O_m$.

We can also define object $O_0$, whose performance level is at the local origin of the measurement scale. Then we can compare the performance of $O_m$ with $O_0$ yielding $P_{m0}/P_{0m}$, and also $O_n$ with $O_0$ yielding $P_{n0}/P_{0n}$.

Rasch pronounced the general dictum:
*If a relationship between two or more variables is to be considered really important, as more than an ad hoc description of a very limited set of data - if a more or less general interdependence may be considered in force - the relationship should be found in several sets of data which differ materially in some relevant respects (Rasch 1980, 9).*

In our case, this implies that the results of a direct comparison of $O_m$ and $O_n$ should lead to the same conclusion as a comparison of $O_m$ with $O_n$ via $O_0$. The requirement for generalizability thus leads to

$$\frac{P_{mn}}{P_{nm}} = \frac{P_{m0}}{P_{0m}} \Big/ \frac{P_{n0}}{P_{0n}} \qquad (11.1)$$

But $P_{m0}/P_{0m}$ is the performance of $O_m$ relative to a measure defined to be at the origin of the scale, and so is a constant, say, $A_m$. Similarly $P_{n0}/P_{0n}$ is a constant, say, $A_n$. Then taking logarithms,

$$\log(P_{mn}/P_{nm}) = \log(A_m) - \log(A_n) \quad (11.2)$$

or, reparameterizing, this becomes the measurement model for paired objects,

$$\log(P_{mn}/P_{nm}) = B_m - B_n \quad (11.3)$$

where

$P_{mn}$ is the probability that object $O_m$ out-performs object $O_n$
$P_{nm}$ is the probability that object $O_n$ out-performs object $O_m$
$B_m$ is the measure of object $O_m$
$B_n$ is the measure of object $O_n$

But $P_{mn} = 1 - P_{nm}$, since tied ranks are excluded, so that,

$$P_{mn} = \exp(B_m)/(\exp(B_m) + \exp(B_n)) \quad (11.4)$$

which is an exponential form of the model for paired comparisons. This model was previously proposed by Bradley-Terry as a descriptive rather than measurement model.

## 11.3 EXTENDING THE PAIRED-COMPARISON MEASUREMENT MODEL TO RANKINGS

If a ranking is of only two objects, then the measurement model for paired objects applies directly. Thus the probability, Rab of observing object Oa ranked higher than object Ob is given by

$$R_{ab} = P_{ab} = \frac{P_{ab}}{P_{ab} + P_{ba}} \quad (11.5)$$

where the denominator contains $2! = 2$ terms, representing all the possible valid numerators for ordering two objects.

The ranking of three objects, $O_a$, $O_b$, $O_c$, can be regarded as a set of three paired rankings, but with the constraint that if $O_a$ is ranked higher than $O_b$, and $O_b$ is ranked higher than $O_c$, then $O_a$ must be ranked higher than $O_c$. The probabilities of their eight theoretically possible paired relationships are shown in Table 29.

TABLE 29
THE COMPARISON OF THREE OBJECTS

| Probability of independent pairings | Representation as rank order |
|---|---|
| $P_{ab}*P_{ac}*P_{bc}$ | $R(O_a,O_b,O_c)$ |
| $P_{ab}*P_{ac}*P_{cb}$ | $R(O_a,O_c,O_b)$ |
| $P_{ab}*P_{ca}*P_{cb}$ | $R(O_c,O_a,O_b)$ |
| $P_{ba}*P_{ac}*P_{bc}$ | $R(O_b,O_a,O_c)$ |
| $P_{ba}*P_{ca}*P_{bc}$ | $R(O_b,O_c,O_a)$ |
| $P_{ba}*P_{ca}*P_{cb}$ | $R(O_c,O_b,O_a)$ |
| $P_{ab}*P_{ca}*P_{bc}$ | inconsistent |
| $P_{ba}*P_{ac}*P_{cb}$ | inconsistent |

The contents of R( ) represent the ordering of the objects.

The effect of the constraint on the pairings imposed by ranking is the determination that two of the possible paired combinations of objects are inconsistent and can never be observed. Apart from this constraint, the probability of observing any particular rank ordering is assumed to depend only on the paired comparison of the objects and not to involve any other characteristics of the sample of objects. This is equivalent to the "local independence" axiom of other Rasch models. Thus the comparison of the objects manifested in the ranking is as "sample-free" as possible. If, for any particular data set of rankings, this is not the case, then the data set cannot be expected to fit the measurement model presented here. Fit statistics can diagnose this eventuality.

Considering the possible rankings, if $R_{ab}$ is the probability that $O_a$ is ranked higher than $O_b$ in the rank ordered data, then,

$$\frac{R_{ab}}{R_{ba}} = \frac{\text{Probability of observing } O_a \text{ higher than } O_b}{\text{Probability of observing } O_b \text{ higher than } O_a} \qquad (11.6)$$

But, tied rankings are not allowed, so that $R_{ab} = 1 - R_{ba}$, giving

$$R_{ab} = \frac{\text{Probability of observing } O_a \text{ ranked higher than } O_b}{\text{Probability of all possible rankings}} \qquad (11.7)$$

$R_{bc}$ and $R_{ac}$ are similarly obtained. These probabilities are not independent as the following identity makes clear:

$$R_{abc} \equiv R_{ab} \cap R_{bc} \equiv R_{ab} \cap R_{bc} \cap R_{ac} \tag{11.8}$$

where

$R_{abc}$ is the probability of observing the ranking $R(O_a,O_b,O_c)$
$\cap$ is the intersection of the sample spaces

Thus, in general,

$$R_{abc} < R_{ab} * R_{bc} \tag{11.9}$$

The probability of an ordering of three objects follows from equation 11.7 and is given by

$$R_{abc} = \frac{\text{Probability of } R(O_a,O_b,O_c)}{\text{Probability of all possible rankings}} = \frac{P_{ab} * P_{ac} * P_{bc}}{\text{Probability of all possible rankings}} \tag{11.10}$$

where the denominator contains all 3! = 6 rankings listed in Table 29.

This can be expressed in a more general way for any three objects, by numbering the objects arbitrarily, O1, O2, O3. Then the probability of their empirical rank ordering, whatever it is, is given by

$$R(\{3\}) = \frac{\prod_{j=1}^{3} \prod_{k=j+1}^{3} (X_{jk} * P_{jk} + X_{kj} * P_{kj})}{\Sigma R(\{3\})} \tag{11.11}$$

where

$R(\{3\})$ is the probability of a particular ranking of 3 objects
$X_{jk} = 1$ if $O_j$ is ranked higher than $O_k$,
$= 0$ otherwise
$X_{kj} = 1 - X_{jk}$
$\Sigma R(\{3\})$ is the sum of all possible numerators and contains one term for every permutation of 3 objects, i.e. 3! = 6 terms.

## 11.4 RANK ORDERING OF N OBJECTS

The principles employed in ranking 3 objects can be extended to the ordering of n objects. For convenience of generalization, let us arbitrarily number the objects O1, O2,..,On with corresponding parameters B1, B2,..,$B_n$. Then, for some rank ordering of the objects, $R(\{n\})$,

$$R(\{n\}) = \frac{\prod_{j=1}^{n} \prod_{k=j+1}^{n} (X_{jk} * P_{jk} + X_{kj} * P_{kj})}{\Sigma R(\{n\})} \qquad (11.12)$$

with the same conventions as before. In particular, $\Sigma R(\{n\})$ is a sum including one term for each of the possible numerators, identical to that numerator. The number of possible numerators is the number of ways of permuting n objects, that is n!.

Alternative forms of the measurement models defining the relationship between objects $O_j$ and $O_k$ are, by equation 11.4,

$$P_{jk} = \exp(B_j) \; / \; (\exp(B_j) + \exp(B_k)) \qquad (11.13)$$

and

$$P_{kj} = \exp(B_k) \; / \; (\exp(B_j) + \exp(B_k)) \qquad (11.14)$$

so that the probability of a rank ordering of n objects in terms of the parameters of the objects is

$$R(\{n\}) = \frac{\prod_{j=1}^{n} \prod_{k=j+1}^{n} \frac{X_{jk}*\exp(B_j) + X_{kj}*\exp(B_k)}{\exp(B_j)+\exp(B_k)}}{\Sigma R(\{n\})} \qquad (11.15)$$

## 11.5 INDEPENDENT RANK ORDERINGS OF N OBJECTS

If independent rank orderings of the same n objects have been compiled by T judges, then the likelihood of the data set, $\cap\{n\}$, becomes

$$\cap\{n\} = \prod_{r=1}^{T} \frac{\prod_{j=1}^{n} \prod_{k=j+1}^{n} \frac{X_{rjk}*\exp(B_j) + X_{rkj}*\exp(B_k)}{\exp(B_j)+\exp(B_k)}}{\Sigma R(\{n\})} \qquad (11.16)$$

If all objects do not participate in every rank ordering, the overall likelihood becomes the product of the likelihood of homogeneous subgroups in which the same set of objects has been ranked by one or more judges. Thus if n objects have been ranked by T judges, and m objects (including some of the n objects) have been ranked by S judges, then the likelihood of the joint data set, $\cap\{mUn\}$, is

$$\cap\{mUn\} = \cap\{n\} * \cap\{n\} \qquad (11.17)$$

which is, explicitly,

$$\prod_{j=1}^{n} \prod_{k=j+1}^{n} X_{rjk}*\exp(B_j) + X_{rkj}*\exp(B_k)$$

$$\text{(11.18)}\qquad \cap\{mUn\} = \left( \prod_{r=1}^{T} \frac{\prod_{j=1}^{n} \prod_{k=j+1}^{n} \exp(B_j)+\exp(B_k)}{\Sigma R(\{n\})} \right) * \left( \prod_{r=1}^{S} \frac{\prod_{j=1}^{m} \prod_{k=j+1}^{m} \frac{X_{rjk}*\exp(B_j) + X_{rkj}*\exp(B_k)}{\exp(B_j)+\exp(B_k)}}{\Sigma R(\{m\})} \right)$$

The following derivation can then be adapted to this formulation of the data, but, for clarity, we return to the consideration of a homogeneous data set.

The factor

$$\text{(11.19)}\qquad \prod_{j=1}^{n} \prod_{k=j+1}^{n} \frac{1}{\exp(B_j)+\exp(B_k)}$$

is common to every term in the numerator and denominator of equation 11.16, and so can be cancelled out. Thus equation 11.16 becomes

$$\text{(11.20)}\qquad \cap\{n\} = \prod_{r=1}^{T} \frac{\prod_{j=1}^{n} \prod_{k=j+1}^{n} (X_{rjk}*\exp(B_j) + X_{rkj}*\exp(B_k))}{\sum_{s=1}^{n!} \prod_{j=1}^{n} \prod_{k=j+1}^{n} (X_{sjk}*\exp(B_j) + X_{skj}*\exp(B_k))}$$

The denominator includes all the possible numerators corresponding to all valid rankings and so consists of n! terms corresponding to the n! ways of ordering n objects.

Taking logarithms, the log-likelihood of a set of T rank orderings of n objects is, and defining $\log(\cap\{n\})$ to be ¥,

$$\text{(11.21)}\qquad ¥ = \sum_{r=1}^{T} \sum_{j=1}^{n} \sum_{k=j+1}^{n} (\log (X_{rjk}*\exp(B_j) + X_{rkj}*\exp(B_k))) - T*\log\left( \sum_{s=1}^{n!} \prod_{j=1}^{n} \prod_{k=j+1}^{n} (X_{sjk}*\exp(B_j) + X_{skj}*\exp(B_k)) \right)$$

## 11.6 ESTIMATION EQUATIONS FOR RANK ORDERED OBJECTS

The Newton-Raphson estimation equations for the parameters can be obtained using first and second derivatives of the log-likelihood function, equation 11.21. Thus, to estimate $B_m$, partially differentiate equation 11.21 with respect to $B_m$,

$$\text{(11.22)}\qquad \frac{\delta ¥}{} = \sum_{r=1}^{T} \sum_{m=1}^{n} X_{rmj}$$

$$\delta B_m \qquad r=1 \;\; j=1,\sim m$$

$$- T * \sum^{n!} \left( \sum^{n} X_{rml} \right) \frac{\prod_{j=1}^{n} \prod_{k=j+1}^{n} \left( X_{rjk} * \exp(B_j) + X_{rkj} * \exp(B_k) \right)}{\sum_{s=1}^{n!} \prod_{j=1}^{n} \prod_{k=j+1}^{n} \left( X_{sjk} * \exp(B_j) + X_{skj} * \exp(B_k) \right)}$$

The first term of equation 11.22 represents the observed score which is the count of the number of objects higher than which Om is ranked in all the observed orderings. The second term represents the expected score and is the sum, across all possible orderings, of the number of objects higher than which Om is ranked in each ordering, multiplied by the probability of that ordering, all multiplied by the number of orderings in the observed data.

Differentiating the log-likelihood again with respect to $B_m$,

$$\frac{\delta^2 \Psi}{\delta B_m^2} = T * \left( \frac{\sum_{r=1}^{n!} \left( \sum^{n} X_{rml} \right) \prod_{j=1}^{n} \prod_{k=j+1}^{n} \left( X_{rjk} * \exp(B_j) + X_{rkj} * \exp(B_k) \right)}{\sum_{s=1}^{n!} \prod_{j=1}^{n} \prod_{k=j+1}^{n} \left( X_{sjk} * \exp(B_j) + X_{skj} * \exp(B_k) \right)} \right)^2$$

$$- T * \left( \frac{\sum_{r=1}^{n!} \left( \sum^{n} X_{rml} \right)^2 \prod_{j=1}^{n} \prod_{k=j+1}^{n} \left( X_{rjk} * \exp(B_j) + X_{rkj} * \exp(B_k) \right)}{\sum_{s=1}^{n!} \prod_{j=1}^{n} \prod_{k=j+1}^{n} \left( X_{sjk} * \exp(B_j) + X_{skj} * \exp(B_k) \right)} \right) \qquad (11.23)$$

These provide the specific forms of the terms of the general form of the Newton-Raphson estimation equation for $B_m'$, the improved estimate of $B_m$, the measure corresponding to object $O_m$,

$$B_m' = B_m - \frac{\delta \Psi}{\delta B_m} \Big/ \frac{\delta^2 \Psi}{\delta B_m^2} \qquad (11.24)$$

When the iterative process has converged, an asymptotic standard error of the estimate, $B_m$, is given by

$$S.E.(B_m) = 1 \Big/ \sqrt{\left( - \frac{\delta^2 \Psi}{\delta B_m^2} \right)} \qquad (11.25)$$

Rasch model fit statistics, both information-weighted and outlier-sensitive, can also be calculated (Wright and Masters 1982, 100).

## 11.7 ESTIMABILITY OF RANK ORDERED DATA

For estimability of all parameters in one frame of reference, it is required that the orderings of the objects overlap in such a way that every object can be compared to every other object, either directly or indirectly, in terms of both relative successes and relative failures. If, for instance, one object is always ranked highest, then its parameter is inestimable. A more subtle example of inestimability is a set of orderings in which the objects form two groups, the high group and the low group, and no object in the high group is ever ranked below any object in the low group.

## 11.8 TIED RANKINGS

In some judging situations, two or more objects may be given the same ranking. If two objects $O_j$ and $O_k$ are given the same ranking, then this is equivalent to the statement that orderings $(O_j, O_k)$ and $(O_k, O_j)$ are equally probable as representations of the ordering of the objects on the latent variable. Consequently, if the tied ranking in the empirical data is replaced with two orderings $(O_j, O_k)$ and $(O_k, O_j)$, each of which is given a weighting of one-half, then the sum is equivalent to the tied ordering. Thus, if $O_j$ and $O_k$ are tied in an ordering, then $X_{jk} = 0.5$ and $X_{kj} = 0.5$ for the purposes of determining empirical scores. Considered in this way, the admissibility of tied rankings does not add any more orderings into the scheme of all possible rank orderings described above.

## 11.9 PAIRED-COMPARISON RANK ORDERING AS A MANY-FACET MODEL

The rank-order measurement model requires in the construction of its denominators the calculation of one term for each possible ordering. This calculation may need to be repeated for each Newton-Raphson iteration. If the number of objects in the ordering is large, problems relating to computation time and estimation precision arise. Consequently, a useful way to apply the rank-order measurement model is to approximate it by a many-facet model in which one facet enters twice.

If the data set consisted of independent paired comparisons, rather than orderings, then the measurement model underlying paired comparisons would be that given in equation 11.3,

$$\log(P_{mn}/P_{nm}) = B_m - B_n \qquad (11.26)$$

As already discussed, ordering the objects prevents independence. Nevertheless a useful approximation to the estimate of the rank-order measurement model could be

$$\log(P_{mn}/(1-P_{mn})) \approx B_m - B_n \qquad (11.27)$$

where $P_{mn}$ is the probability that $O_m$ is ranked higher than $O_n$ and this probability is treated as independent of the constraint of rank ordering.

This has the form of a two-facet measurement model. Then, if the rank orders are recoded in the data set in a form such that $X_{rmn} = 1$ if $O_m$ is ranked higher than $O_n$ in ordering r,

and $X_{rmn} = 0$ if $O_m$ is ranked lower than $O_n$ in ordering r, then the estimation equations become computationally equivalent to those for a two-facet model, namely,

$$B_n' = B_n - \frac{\sum_{n}^{(\Omega|m)} X_{rmn} - \sum_{n}^{(\Omega|m)} P_{mn}}{\sum_{n}^{(\Omega|m)} (P_{mn} - P_{mn}^2)} \qquad (11.28)$$

where $(\Omega|m)$ is the set of all paired comparisons, $X_{rmn}$, in which $O_m$ takes part, with $O_n$ representing the other object in each paired comparison.

In equation 11.28, the term, $\Sigma X_{rmn}$, is the same as the leading term in equation 11.22.

The effect of the constraint of rank ordering, however, was to invalidate inconsistent paired comparisons and so limit the combinations that could be observed. Thus the sample space of rank orders is smaller than the sample space of paired combinations. consequently, the probability of any particular rank ordering being observed is greater than would be the probability of the same ordering if it were to occur through the accumulation of independent paired comparisons. Thus, the term $P_{mn}$ in equation 11.28 underestimates the true probability, so that the one-facet version of the many-facet model overestimates the differences between the measures of rank-ordered objects. The standard error is also overestimated.

With orderings of only two objects, the many-facet model and the rank-ordered model are identical. As the number of objects increases, the bias introduced by ignoring the rank-order constraint also increases. With seven objects, the difference between two measures estimated without the ranking constraint is about twice that of constrained measures. A simulation study indicates that the bias is of the order of 1.15 to the power of the number of objects ranked. The amount of bias also depends on the distribution of the measures of the objects. With this bias correction, the many-facet model provides approximate estimates with considerably less computational effort than the rank-order model. Consequently, the many-facet model can be used to provide convenient estimates for exploratory analysis, or good starting values for the rank-order model.

## 11.10 RANK ORDERING CONSIDERED AS A RATING SCALE MODEL

Andrich (1978) proposed a Rasch model for rating scales with several properties relevant to rank ordering.

1) The categories of the scale are ordered qualitatively.
2) Each higher category of the scale is qualitatively different to its predecessor.
3) The categories exhaust the possible levels of performance.

4) Classification of an object in a category implies that the object supersedes all lower categories.
5) The classification process allows the classification of any object in any category, e.g., a rater always has the full range of categories available when classifying an object.

Ranking can also be conducted in the same way. The judge can be instructed to classify the objects into as many different hierarchically-ordered levels of quality as the judge discriminates in the objects. The outcome will be a judge-defined rating scale for these particular objects, that is also a ranking of the objects. The rankings across can then be treated as rating scale data and analyzed accordingly.

# 12 AN EXAMPLE OF THE MEASUREMENT OF RANK ORDERED OBJECTS

Statistical analysis of rank-ordered data is generally limited to the application of non-parametric tests, which, perhaps by means of some transformation, include "the assumption of normality" in the distribution of the underlying variable (Brownlee 1965, 241). The measurement model derived in chapter 11 makes no assumption about the distribution of the parameters. In this chapter a demonstration of the properties of the measurement model is performed by means of a small data set which is clearly bi-modal.

| Calling the game | Working with analyst | Broadcasting ability |
|---|---|---|
| 1. Vin Scully | 1. Bob Costas | 1. Vin Scully |
| 2. Bob Costas | 2. Al Michaels | 2. Al Michaels |
| 3. Al Michaels | 3. Vin Scully | 3. Bob Costas |
| 4. Skip Caray | 4. Skip Caray | 4. Skip Caray |
| 5. Harry Caray | 5. Steve Zabriskie | 5. Harry Caray |
| 6. Steve Zabriskie | 6. Ralph Kiner | 6. Steve Zabriskie |
| 7. Ralph Kiner | 7. Harry Caray | 7. Ralph Kiner |
| **Quality of anecdotes** | **Knowledge of baseball** | **Enthusiasm level** |
| 1. Vin Scully | 1. Vin Scully | 1. Harry Caray |
| 2. Bob Costas | 2. Ralph Kiner | 2. Al Michaels |
| 3. Al Michaels | 3. Bob Costas | 3. Bob Costas |
| 4. Skip Caray | 4. Al Michaels | 4. Vin Scully |
| 5. Ralph Kiner | 5. Harry Caray | 5. Steve Zabriskie |
| 6. Harry Caray | 6. Skip Caray | 6. Skip Caray |
| 7. Steve Zabriskie | 7. Steve Zabriskie | 7. Ralph Kiner |

Fig. 12.1 Rank Orderings of Baseball Announcers (Polskin 1988).

Polskin (1988) publishes rank orderings of seven baseball announcers, reproduced here in fig. 12.1. These are individuals who reported baseball during the 1987 season for the U.S. TV networks or major independent TV stations. Those whose responses were used in compiling these rank orderings were themselves experienced baseball announcers.

The initial methodology was that the 44 expert respondents were asked to grade each of the seven announcers on the six items of performance with a rating in the range of 1 to 5. These ratings were not published. They were, however, the basis of the rank orderings on

the six items shown in fig. 12.1. These six items will be considered to be independent manifestations of a latent variable which can be termed "Quality of Announcing."

## 12.1 A PAIRED-COMPARISON RASCH ANALYSIS OF THE RANK ORDERINGS

The paired-comparison rank-order measurement model can be used to answer such questions as "How much better is one announcer than another?", "Which announcers have the most consistent quality level?" and "Do these items represent one underlying 'Quality of announcing' variable?"

TABLE 30
ABILITIES OF BASEBALL ANNOUNCERS

| Ability Order | Sum of Rankings | Measure (Logits) | S.E. | Mean Square Fit | Announcer | Polskin's "Overall" Rank |
|---|---|---|---|---|---|---|
| 1 | 11 | 0.98 | 0.41 | 1.51 | Vin Scully | 1 |
| 2 | 14 | 0.67 | 0.35 | 0.40 | Bob Costas | 2 |
| 3 | 16 | 0.50 | 0.32 | 0.34 | Al Michaels | 3 |
| 4 | 28 | -0.26 | 0.28 | 0.43 | Skip Caray | 5 |
| 5 | 29 | -0.33 | 0.29 | 1.73 | Harry Caray | 4 |
| 6 | 34 | -0.69 | 0.33 | 2.12 | Ralph Kiner | 7 |
| 7 | 36 | -0.87 | 0.37 | 0.54 | Steve Zabriskie | 6 |
| | Mean: | 0.00 | | 1.01 | | |

In answer to the question, "How much better is one announcer than another?", Table 30 lists the estimates of the measures obtained for this data set, using the estimation equations presented in Chapter 11. The distinction between "information weighted" fit statistics and "outlier sensitive" fit statistics does not exist here because the variance term for each estimate is uniform across rank orderings. The right-hand column is an overall ranking of the announcers presented by Polskin.

The relationship between the sum of each announcer's ranks and his measure is close to linear as can be seen in fig. 12.2, which plots each announcer's measure against his rank. According to these estimates, the top two announcers are significantly better than the bottom two announcers.

Comparison of these results with Polskin's "Overall" ranking, in the last column of Table 30, shows that the order of the announcers, when ranked on measures, differs from that published, which was based on the unpublished rating scale responses. According to the Polskin, Vin Scully had the highest average rating of 4.4 and Ralph Kiner the lowest of 3.1. The ordering by measure and the "Overall" orderings do not differ in meaning, because the measures obtained for Skip Caray and Harry Caray, and also for Ralph Kiner and Steve Zabriskie, are not significantly different. This result, however, does show how misleading is the equal-interval appearance of a sequentially numbered rank-ordered list.

From Table 30, the announcer with the most consistent level of performance across all items is Al Michaels with a mean-square fit statistic of 0.34, and the most inconsistently performing one is Ralph Kiner with a fit statistic of 2.12. How difficult misfit is to determine, by eye, from lists of rank orderings, is indicated by the quite different conclusions reached by Polskin. According to the analysis given in the text of his article, Polskin had the impression that Harry Caray was the least consistently rated announcer (due to his first place on "Enthusiasm") and Al Michaels was also somewhat inconsistent (not due to his fourth place on "Knowledge," but his third place on "Quality").

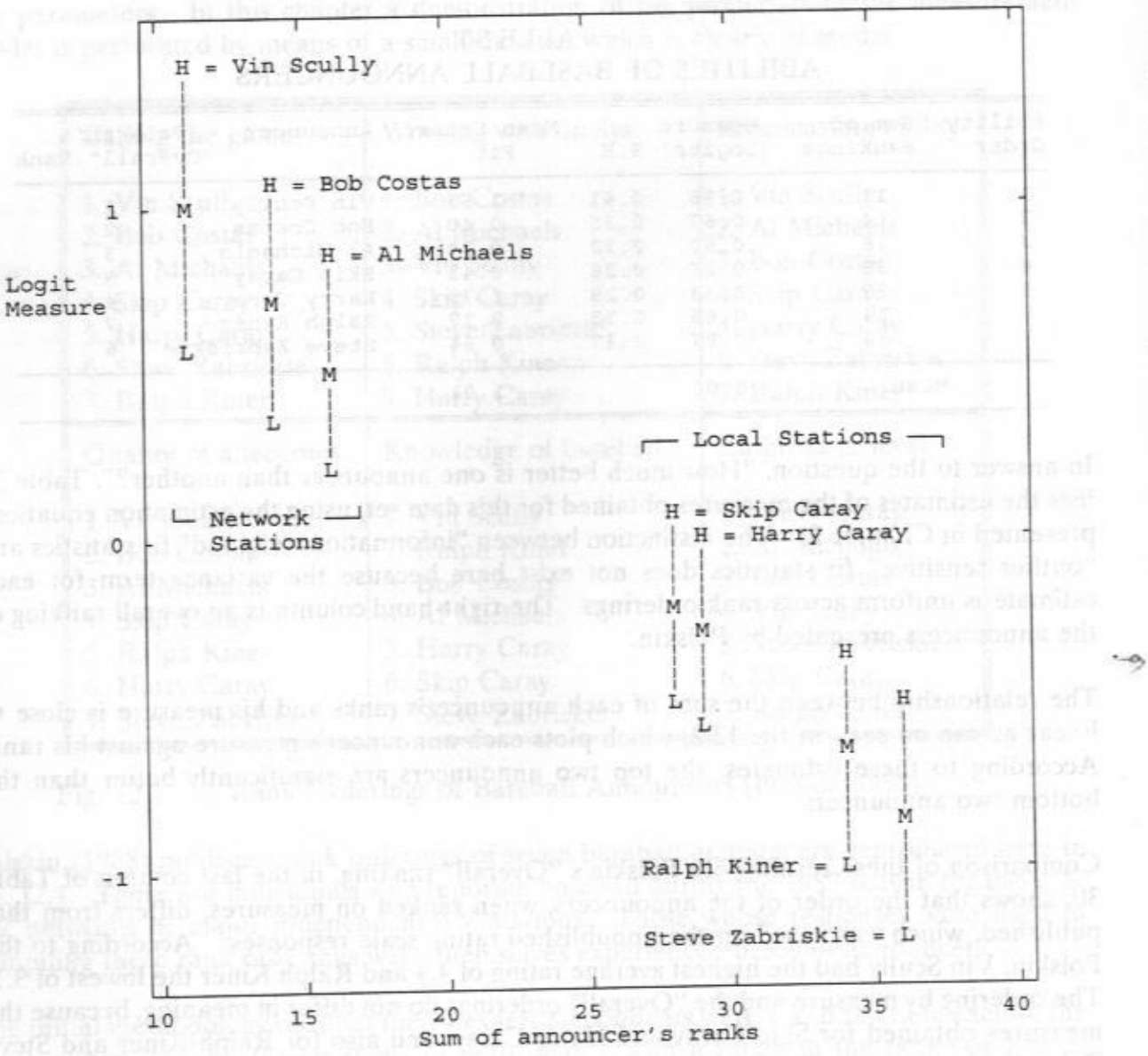


Fig. 12.2 Announcers' measures plotted against sum of rankings.
M = estimated measure, H = M + standard error, L = M - standard error

TABLE 31
MOST UNEXPECTED RANKINGS OF ANNOUNCERS

| Ordering | Announcer | Rank | Expected | Difference | S.E. | Z-Score |
|---|---|---|---|---|---|---|
| Knowledge | Ralph Kiner | 2 | 5.7 | 3.67 | 1.23 | 2.97 |
| Enthusiasm | Harry Caray | 1 | 4.8 | 3.83 | 1.42 | 2.71 |
| Working | Bob Costas | 1 | 2.3 | 1.33 | 1.18 | 1.13 |
| Working | Vin Scully | 3 | 1.8 | -1.17 | 0.99 | -1.18 |
| Working | Harry Caray | 7 | 4.8 | -2.17 | 1.42 | -1.53 |
| Enthusiasm | Vin Scully | 4 | 1.8 | -2.17 | 0.99 | -2.19 |
| Mean for all ranks: | | | | 0.0 | | 0.00 |
| Variance for all ranks: | | | | | | 1.00 |

Table 31 shows those rankings which were the least expected. This Table is an aid to the diagnosis of aberrations in the measuring process and assists in identifying the sources of misfit. It supports one of Polskin's conclusions that Vin Scully was perceived to have performed somewhat inconsistently, partially due to his fourth place on "Enthusiasm." On the other hand, Polskin failed to comment on the fact that Ralph Kiner's rankings have the most inconsistency, and further that the most unexpected ranking of all was that of Ralph Kiner on "Knowledge." Again, a caveat is that Polskin had access to unpublished information. Nevertheless, it is clear that the published rank orders are intended to faithfully represent that information.

TABLE 32
FIT STATISTICS FOR RANK ORDERINGS

| Information-weighted Mean-Square fit | Outlier-sensitive Mean-Square fit | Item of Performance |
|---|---|---|
| 0.29 | 0.32 | Calling the game |
| 0.91 | 0.91 | Working with analyst |
| 0.35 | 0.39 | Broadcasting ability |
| 0.38 | 0.41 | Quality of anecdotes |
| 1.77 | 1.80 | Knowledge of baseball |
| 2.29 | 2.22 | Enthusiasm level |

A fundamental question, not discussed by Polskin, is the uni-dimensionality of the "quality of announcing" variable. Are the six items, on which the announcers have been ordered, independent representations of the same variable? Table 32 summarizes the degree of fit within rank orderings. Since ordering provides no information on, say, how difficult it is to "call the game" relative to the items used to create the other orderings, no difficulty calibrations are shown. On the other hand, there is a numerical difference between the two mean-square fit statistics, though this has no substantive meaning in this data set.

The items are generally acting in a coherent manner in defining the variable. It may well be that "Calling the game" and "Broadcasting ability" are somewhat synonymous and not

independent items, leading to a redundancy in the data, and hence mean-square fit statistics substantially less than their expected value of 1. On the other hand, hypotheses could be presented as to why "Quality of anecdotes" and "Knowledge of baseball" which would appear to be closely related have behaved differently. Only "Enthusiasm" has strong indications of what may be a separate variable, and this is because, according to Polskin, it is easier to announce when you are doing it for the "home-team fans," as Harry Caray does.

## 12.2 CONCLUSION TO THE PAIRED-COMPARISON RANK ORDER ANALYSIS

The application of the principles of fundamental measurement to rank ordered data has provided the means to convert entirely local rankings into generalizable measures of the latent abilities. Moreover, fit statistics for each announcer and for each ordering enable a determination of the success of the ranking process as a measurement operation.

Rasch analysis has yielded information consistent with the published analysis. Yet it has gone beyond that analysis in providing measures which enable the determination of whether the difference between announcers is significant, or only an illusion of the manner of presentation. Further, this analysis has brought to the analyst's attention those aspects of rank-ordered performance which are truly remarkable.

Rasch analysis has also permitted the confirmation or clarification of whether the rank orderings are independent but alternative expressions of performance on the same unidimensional latent variable.

## 12.3 A RATING-SCALE RASCH ANALYSIS OF THE RANK ORDERINGS

Treating rankings as ordered rating scale categories enables measures to be estimated with the many-facet Rasch model derived earlier. The measurement model becomes

$$\log(P_{jk}/P_{jk-1}) = -B_m + F_{jk} \qquad (12.1)$$

where

$P_{jk}$ = the probability of a ranking of k by judge j, where k=1,7
$P_{jk-1}$ = the probability of a higher ranking of k-1 by judge j
$B_m$ = the ability of sportscaster m
$F_{jk}$ = the extra ability implicit in a ranking of k-1 beyond that in a ranking of k, according to judge j.

When tied rankings are not observed, and all judges rate all objects, then the $\{F_{jk}\}$ are identical across judges, and may be parameterized $\{F_k\}$.

## TABLE 33
## RATING SCALE MEASURES FOR RANK-ORDER DATA

| Obsvd Score | Obsvd Count | Obsvd Average | Fair Avrge | Measure Logit | Model Error | Infit MnSq | Infit Std | Outfit MnSq | Outfit Std | N Sportscasters |
|---|---|---|---|---|---|---|---|---|---|---|
| 11 | 6 | 1.8 | 1.9 | 1.25 | 0.42 | 1.6 | 0 | 1.6 | 0 | 1 Vin Scully |
| 14 | 6 | 2.3 | 2.4 | 0.83 | 0.34 | 0.4 | -1 | 0.4 | -1 | 2 Bob Costas |
| 16 | 6 | 2.7 | 2.7 | 0.62 | 0.31 | 0.3 | -1 | 0.3 | -1 | 3 Al Michaels |
| 28 | 6 | 4.7 | 4.7 | -0.33 | 0.28 | 0.4 | -1 | 0.4 | -1 | 4 Skip Caray |
| 29 | 6 | 4.8 | 4.9 | -0.41 | 0.28 | 1.6 | 1 | 1.6 | 1 | 5 Harry Caray |
| 34 | 6 | 5.7 | 5.7 | -0.86 | 0.33 | 2.1 | 1 | 2.1 | 1 | 6 Ralph Kiner |
| 36 | 6 | 6.0 | 6.0 | -1.10 | 0.37 | 0.6 | 0 | 0.6 | 0 | 7 Steve Zabriskie |
| 24.0 | 6.0 | 4.0 | 4.0 | -0.00 | 0.33 | 1.0 | -0.3 | 1.0 | -0.3 | Mean (Count: 7) |
| 9.4 | 0.0 | 1.6 | 1.6 | 0.83 | 0.05 | 0.7 | 1.3 | 0.7 | 1.3 | S.D. |

RMSE 0.34 Adj S.D. 0.76 Separation 2.26 Reliability 0.84
Fixed (all same) chi-square: 37.39 d.f.: 6 significance: .00
Random (normal) chi-square: 5.84 d.f.: 5 significance: .32

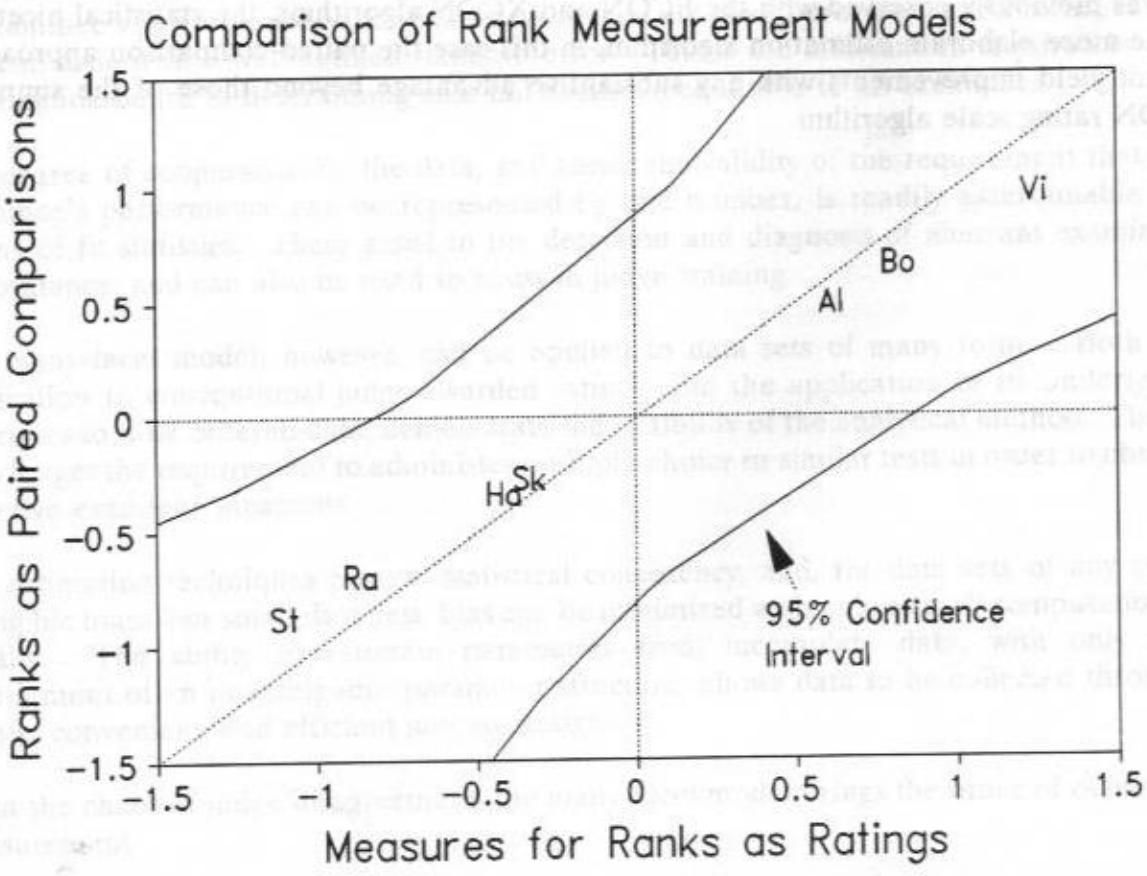


Fig. 12.3 Comparison of announcers' measures from paired-comparison and rating scale analyses. The points do not lie exactly on the identity line, indicating that the methods vary slightly in measure discrimination, but the measures are substantively identical. Vi = Vin Scully etc.

For the Polskin data, a rating scale analysis using the UCON algorithm implemented in the *Facets* program yields Table 33. The relationship between the two analyzes is shown in fig. 12.3. All points lie well within the confidence bands, indicating that they are substantively identical. The slight skewness off the identity line indicates a slight difference in test discrimination between the two methods.

## 12.4 CONCLUSION TO RANK ORDER COMPARISON

Analysis of rank orders as rating scales is straight-forward to implement with generally available Rasch analysis software, e.g, *Facets*. Rating scale analysis yields substantively identical results to the paired-comparison approach, but without the latter's significant computational demands. With the rating scale approach, incomplete rankings, i.e., rankings of not all objects, and tied rankings place no extra demands either on data collection or analysis beyond those required of all rating scale data for unambiguous measure estimation.

As was previously observed with the FCON and XCON algorithms, the statistical niceties of the more elaborate estimation algorithm, in this case the paired-comparison approach, do not yield improvements with any substantive advantage beyond those of the simpler UCON rating scale algorithm.

# 13 CONCLUSION

The many-facet Rasch model extends the possibility of objective measurement to examinations which include subjective judgments. Its development enables the benefits of "sample-free," "test-free," and "judge-free" measurement to be realized in this hitherto intractable area. The use of the many-facet model permits greater freedom from judge bias and greater generalizability of the resulting examinee measures than has previously been available.

The many-facet model has a strong theoretical structure, but imposes a minimum of restrictions on the data to be analyzed. If the empirical data cooperates in the construction of a uni-dimensional variable, of the type required in order to summarize into one measure an examinee's performance on an examination, then the model provides such a measure on a linear scale with a well-defined standard error. This is the information required by an examination board in determining pass-fail decisions equitable to all examinees.

The degree of cooperation by the data, and hence the validity of the requirement that an examinee's performance can be represented by one number, is readily ascertainable by means of fit statistics. These assist in the detection and diagnosis of aberrant examinee performance, and can also be used to assist in judge training.

The many-facet model, however, can be applied to data sets of many forms. Both its application to conventional judge-awarded ratings, and the application of its underlying principles to rank ordered data, demonstrate the flexibility of the analytical method. There is no longer the requirement to administer multiple-choice or similar tests in order to obtain objective examinee measures.

The estimation techniques possess statistical consistency, and, for data sets of any size, negligible bias. For small data sets, bias can be minimized with only a small computational penalty. The ability to estimate parameters from incomplete data, with only the requirement of an unambiguous parameter structure, allows data to be collected through simple, convenient, and efficient judging designs.

From the chaos of judge disagreement, the many-facet model brings the order of objective measurement.

# REFERENCES


Andersen, E. B. 1973. Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology* 26: 31-44.

Andersen, E. B. 1977. Sufficient statistics and latent trait models. *Psychometrika* 42: 69-81.

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

Andrich, D. 1982. An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika* 47: 105-113.

Bannister, B. D., Kinicki, A. J., Denisi, A. S., and Hom, P. W. 1987. A new method for the statistical control of rating error in performance ratings. *Educational and Psychological Measurement* 47: 583-596.

Barnes-Farrell, J. L., and Weiss, H. M. 1984. Effects of standard extremity on mixed standard scale performance ratings. *Personnel Psychology* 37: 301-316.

Barrett, R. S. 1966. *Performance rating*. Chicago, Il.: Science Research Associates.

Borman, W. C. 1978. Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology* 63: 134-144.

Braun, H. I. 1986. *Calibration of essay readers (Final Report): Program Statistics Research Tech. Rep. No. 86-88.* Princeton, N.J.: Educational Testing Service.

Braun, H. I. 1988. Understanding scoring reliability: experiments in calibrating essay readers. *Journal of Educational Statistics* 13: 1-18.

Brogden, H. E. 1977. The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika* 42: 631-634.

Browne, M. W. 1987. Is nothing sacred? Constants of science submit to revision. *New York Times*, 24 February.

Brownlee, K. A. *Statistical theory and methodology*. 2d Ed. New York: John Wiley and Sons.

Campbell, D. T., and Fiske, D. W. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56: 81-105.

Campbell, N. R. 1920. *Physics: the elements*. Cambridge: Cambridge University Press.

Cason, G. J., and Cason, C. L. 1984. A deterministic theory of clinical performance rating. *Evaluation and the health professions* 7: 221-247.

Choppin, B. H. 1968. Item banking using sample-free calibration. *Nature* 219: 870-872. Reprinted in *Evaluation in Education* 1985, 9: 81-85.

Choppin, B. H. 1982. The use of latent trait models in the measurement of cognitive abilities and skills. Chapter 3 of *The improvement of measurement in education and psychology*, ed. D. Spearitt. Hawthorn, Victoria, Australia: ACER. Reprinted in *Evaluation in Education*, 1985, 9: 13-28.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37-46.

Cohen, L. 1979. Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology* 32: 113-120.

Constable, E., and Andrich, A. 1984. Inter-judge reliability: is complete agreement among judges the ideal? Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans. Photocopied.

Crocker, L., and Algina, J. 1986. *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. 1972. *The dependability of behavioral measurements*. New York: John Wiley and Sons.

de Gruiter, D. N. M. 1980. The essay examination. In *Psychometrics for Educational Debates*, ed. L. J. T. van der Kamp et al. New York: John Wiley and Sons.

de Gruiter, D. N. M. 1984. Two simple models for rater effects. *Applied Psychological Measurement* 8: 213-218.

Dillon, W. R., and Mulani, N. 1984. A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research* 19: 438-458.

Dorsey, N. E. 1944. The velocity of light. *Transactions of the American Philosophical Society* 34: 1-110.

Ebel, R. L. 1951. Estimation of the reliability of ratings. *Psychometrika* 16: 407-424.

Edgeworth, F. Y. 1890. The element of chance in competitive examinations. *Journal of the Royal Statistical Society* 53: 460-475 and 644-663.

Eisenhart, C. 1969. Realistic evaluation of the precision and accuracy of instrument calibration systems. In *Precision Measurement and Calibration*, ed. H. Ku. Washington D.C.: National Bureau of Standards.

Feyerabend, P. 1975. *Against method*. Bristol, England: NLB.

Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. *Proceedings of the Royal Society* 222: 309-368.

Fisher, R. A. 1925. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22: 700-725. Reprinted in *Collected Papers of R.A. Fisher*, ed. J. H. Bennett. Adelaide: The University of Adelaide.

Fleiss, J. L. 1981. Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement* 5: 105-112.

Gruenfeld, E. F. 1981. *Performance appraisal: promise and peril*. Ithaca, New York: Cornell University.

Guilford, J. P. 1954. *Psychometric methods*. 2d Ed. New York: Mc-Graw Hill.

Guttman, L. 1950. The basis for scalogram analysis. In *Measurement and Prediction*, Volume IV of *The American Soldier*, ed. S. Stouffer et al. New York, Wiley.

Haberman, S. J. 1977. Maximum likelihood estimates in exponential response models. *The Annals of Statistics* 5: 815-841.

Harper, A. E. Jr., and Misra, V. S. 1976. *Research on examinations in India*. New Delhi, India: National Council of Educational Research and Training.

Hartog, P., and Rhodes, E. C. 1936. *The marks of examiners*. London: Macmillan and Co.

Harwell, M. R., Baker, F. B., and Zwarts, M. 1988. Item parameter estimation via marginal maximum likelihood and an EM algorithm: a didactic. *Journal of Educational Statistics* 13: 243-271.

Hempel, C. 1952. Fundamentals of concept formation in empirical science. In *International Encyclopedia of Unified Science*, ed. O. Neurath, R. Carnap, and C. Morris. Chicago: University of Chicago Press.

Hubert, L., and Golledge, R. G. 1983. Rater agreement for complex assessments. *British Journal of Mathematical and Statistical Psychology* 36: 207-216.

Jansen, P. G. W., van den Wollenberg, A. L., and Wierda, F. W. 1988. Correcting unconditional parameter estimates in the Rasch model for inconsistency. *Applied Psychological Measurement* 12: 297-306.

Jensen, S. T., Johansen, S., and Lauritzen, S. L. 1988. *An algorithm for maximizing a likelihood function*. Preprint No.3. Copenhagen: Institute of Mathematical Statistics, University of Copenhagen.

Kavanaugh, M. J., McKinney A. C., and Wolins, L. 1971. Issues in managerial performance: multitrait-multimethod analysis of ratings. *Psychological Bulletin* 75: 34-49.

Kelderman, H. 1986. *Common item equating using the log-linear Rasch model*. Twente, the Netherlands: University of Twente, Department of Education.

Kelderman, H., and Steen, R. 1988. *LOGIMO computer program for log-linear item response theory modelling*. Twente, the Netherlands: University of Twente, Department of Education.

Kelly, F. J. 1914. *Teachers' marks*. New York: Teachers College.

Kinston, W. 1985. Measurement and the structure of scientific analysis. *Systems Research* 2(2): 95-104.

Kruskal, W. H. 1960. Some remarks on wild observations. *Technometrics*, 2: 1. Reprinted in *Precision Measurement and Calibration*, ed. H. H. Ku. 1969. Washington D.C.: National Bureau of Standards.

Lagueux, B. J., and Amols, H. I. 1986. Making your science fair fairer. *Science Teacher* 53(2): 24-28.

Landy, F. J., and Farr, J. L. 1983. *The measurement of work performance*. New York: Academic Press.

Linacre, J. M. 1987. *An extension of the Rasch model to multi-facet situations*. Chicago: University of Chicago, Department of Education.

Linacre, J. M. 1988. *Facets: a computer program for many-facet Rasch measurement*. Chicago: MESA Press.

Luce, R. D., and Perry, A. D. 1949. A method of matrix analysis of group structure. *Psychometrika* 14: 95-116.

Luce, R. D., and Tukey, J. W. 1964. Simultaneous conjoint measurement: a new type of fundamental measurement. *Journal of Mathematical Psychology* 1: 1-27.

Lunz, M. E., Wright, B. D., & Linacre, J. M. 1990. Measuring the impact of judge severity on examination scores. *Applied Measurement in Education* 3: 331-345

Masters, G. N. 1982. A Rasch model for partial credit scoring. *Psychometrika* 47: 149-174.

McCrea, W. H. 1983. Introductory remarks. *Phil. Trans. R. Soc. Lond. A.* 310: 211-213. Reprinted in *The constants of physics*, ed. W. H. McCrea and M. J. Rees. London: The Royal Society.

Mount, M. K. 1984. Psychometric properties of subordinate ratings of managerial performance. *Personnel Psychology* 37: 687-702.

Muller, Hans. 1987. A Rasch model for continuous ratings. *Psychometrika* 52: 165-181.

Neyman, J., and Scott E. L. 1948. Consistent estimates based on partially consistent observations. *Econometrica* 16: 1-32.

Paul, S. R. 1981. Bayesian methods for calibration of examiners. *British Journal of Mathematical and Statistical Psychology* 34: 213-223.

Pedler, P. J. 1987. *Accounting for psychometric dependence with a class of latent trait models*. Ph.D. diss. Perth: University of Western Australia.

Polskin, H. 1988. The best sportscasters in baseball. *TV Guide*, 30 July.

Pontius, P. E., and Cameron, J. M. 1969. Realistic uncertainties and the mass measurement process. In *Precision Measurement and Calibration*, ed. H. H. Ku. Washington D.C.: National Bureau of Standards.

Rao, C. R. 1952. *Advanced statistical methods in biometric research*. New York: John Wiley and Sons.

Rasch, G. 1960, 1980. *Probabilistic models for some intelligence and attainment tests*. Copenhagen, and Chicago: University of Chicago Press.

Rasch, G. 1961. On general laws and meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press.

Rasch, G. 1964. *Objective comparisons*. Paper presented at UNESCO Seminar, Voksenasen, Oslo. Photocopied.

Rasch, G. 1966. *An individual-centered approach to item analysis with two categories of answers*. Unpublished paper. Photocopied.

Rasch, G. 1968. A mathematical theory of objectivity and its consequences for model construction. In *Report from the European Meeting on Statistics, Econometrics and Management Sciences*. Amsterdam. Photocopied.

Rasch, G. 1977. On specific objectivity: an attempt at formalizing the quest for generality and validity of scientific statements. *Danish Yearbook of Philosophy* 15: 58-94.

Roskam, E. E., and Jansen, P. G. W. 1984. A new derivation of the Rasch model. *Trends in Mathematical Psychology* ed. E. Degreef, and J. van Buggenhaut. North-Holland, the Netherlands: Elsevier Science Publishers.

Ruch, G. M. 1929. *The objective or new-type examination.* Chicago: Scott, Foresman.

Ruggles, A. M. 1911. *Grades and grading*. New York: Teacher's College. Reported in F.J. Kelly, *Teacher's marks*. New York: Teacher's College. 1914.

Saal, F. E., Downey, R. G., and Lahey, M. E. 1980. Rating the ratings: assessing the psychometric quality of rating data. *Psychological Bulletin* 88: 413-428.

Siegel, A. I. 1984. Review of performance measurement and theory, ed. F. Landy, S. Zedeck, and J. Cleveland. 1983. Hillsdale NJ: Lawrence Erlbaum Associates. *Personnel Psychology* 37: 761-763.

Stanley, J. C., and Hopkins, K. D. 1972. *Educational and psychological measurement and evaluation*. Englewood Cliffs, N.J.: Prentice-Hall.

Swaminathan, H., and Gifford, J. A. 1982. Bayesian estimation in the Rasch model. *Journal of Educational Statistics* 7: 175-192.

Thiele, T. N. 1903. *Theory of observations*. London: Charles and Edwin Layton.

Thurstone, L. L. 1927a. A law of comparative judgment. *Psychological Review* 3: 273-286.

Thurstone, L. L. 1927b. Psychophysical analysis. *American Journal of Psychology* July, 368-389.

Thurstone, L.L., and Chave, E. J. 1929. *The measurement of attitude*. Chicago: University of Chicago Press.

Vassiloglou, M., and French, S. 1982. Arrow's theorem and examination assessment. *British Journal of Mathematical and Statistical Psychology* 35: 183-192.

Verhelst, N., and Molenaar, I.W. 1988 Logit based parameter estimation in the Rasch model. *Statistica Neerlandica* 42:4, 273-295.

Wall, K.-D. 1980. A non-parametric approach to reliability estimation of essay examinations. In *Psychometrics for Educational Debates*, ed. L. J. T. van der Kamp et al. New York: John Wiley and Sons.

Whisler, T. L., and Harper, S. F. 1962. *Performance appraisal: research and practice*. New York: Holt, Rinehart and Winston.

Williams, V. D. 1979. *Inter-rater agreement as related to ratee-item distance and homogeneity of two rating scales*. Ph.D. diss. Chicago: University of Chicago.

Wilson, M. 1991. Unobserved categories. *Rasch Measurement Transactions* 5:1, 128.

Windmeijer, F. A. G. 1990. The asymptotic distribution of the sum of weighted squared residuals in binary choice models. *Statistica Neerlandica* 44: 69-78.

Wright, B. D. 1968. Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service.

Wright, B. D. 1986. *Bayes' answer to perfection*. MESA memorandum. Chicago: University of Chicago, Department of Education.

Wright, B. D. 1986. *Why you need fundamental measurement*. MESA Memorandum No. 37. Chicago: University of Chicago, Department of Education.

Wright, B. D. 1988. The efficacy of unconditional maximum likelihood bias correction: comment on Jansen, van den Wollenberg, and Wierda. *Applied Psychological Measurement* 12: 315-318.

Wright, B. D., and Bell, S.R. 1984. Item banks: what, why, how. *Journal of Educational Measurement* 21: 331-345.

Wright, B. D., and Douglas, G. A. 1976. *Better procedures for sample-free item analysis*. MESA Memorandum No. 20. Chicago: University of Chicago, Department of Education.

Wright, B. D., and Douglas, G. A. 1977. Best procedures for sample-free item analysis. *Applied Psychological Measurement* 1: 281-294.

Wright, B. D., and Douglas, G. A. 1986. *The rating scale model for objective measurement*. MESA Memorandum No. 35. Chicago: University of Chicago, Department of Education.

Wright, B. D., and Linacre, J. M. 1987. Rasch models derived from objectivity. *Rasch SIG Newsletter* 1(1): 2-3.

Wright, B. D., and Masters, G. N. 1982. *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Wright, B. D., and Panchapakesan, N. A. 1969. A procedure for sample-free item analysis. *Educational and Psychological Measurement* 29: 23-48

Wright, B. D., and Stone, M. H. 1979. *Best test design: Rasch measurement*. Chicago: MESA Press.

Zwick, R. 1986. *Another look at inter-rater agreement*. Princeton: Educational Testing Service.

# INDEX