ED 201 669 TM 810 264

AUTHOR	Cason, Gerald J.; Cason, Carolyn L.
TITLE	Some Promising Early Results from a Rudimentary Latent-Trait Theory of Performance Rating.
PUB DATE	Apr 81
NOTE	39p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Clinical Experience; *Cohort Analysis; *Latent Trait Theory; Mathematical Models; *Medical Students; *Performance Tests
IDENTIFIERS	Inter Rater Reliability

ABSTRACT

A theory is discussed in which observed performance ratings are derived from the distance between a rater reference point and subject performance point located on a postulated equal-interval scale and a postulated s-shaped rater characteristic curve, operationalized as the normal ogive. Least-squares estimates of rater (nR=47, 31, and 29) and subject (nS=29, 30, and 35) points were determined separately on each of three junior medical student cohorts' data. The proposed model fit each of the data sets better than two alternative models (r is greater than 0.70; p is less than 0.01). Test-retest reliabilities for rater parameters were r is less than 0.29 (joint p is less than 0.04). Cross-validating results also supported the theory. (Author/RL)

Some Promising Early Results from a Rudimentary Latent-Trait Theory of Performance Rating

Gerald J. Cason and Carolyn L. Cason
University of Arkansas for Medical Sciences

A theory in which observed performance ratings are derived from the distance between a rater reference point and subject performance point located on a postulated equal-interval scale and a postulated s-shaped rater characteristic curve, operationalized as the normal ogive, is presented. Least-squares estimates of rater ($nR=47$, 31, and 29) and subject ($nS=29$, 30, and 35) points were determined separately on each of three junior medical student cohorts' data. The proposed model fit each of the data sets better than two alternative models ($r>0.70$; $p<0.01$). Test-retest reliabilities for rater parameters were $r<0.29$ (joint $p<0.04$). Cross-validating results also supported the theory.

Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, 1981.

Some Promising Early Results from a Rudimentary Latent-Trait Theory of Performance Rating

Gerald J. Cason and Carolyn L. Cason
University of Arkansas for Medical Sciences

Usually we must rely upon the judgement of human raters to assess, i.e., to measure and evaluate, complex human performance and products. In this context, "measure" means a systematic procedure which assigns numbers (e.g., scores, ratings) the values of which represent how much of some attribute, characteristic, or factor is present. "Evaluation" means the determination of merit or adequacy. We rely upon human judgement to assess performances as varied as (a) conducting a cross-examination in a trial court, (b) diagnosing a patient's medical problem, and (c) landing a high-performance aircraft. Also, human judgement is fundamental to the assessment of such products as (a) an article submitted for publication, (b) the prototype of an implantable mechanical heart, and (c) the design plans for a new mousetrap or orbital shuttlecraft.

The research reported here is concerned with improving ratings-based measures of human performance. Our interest in the problems associated with ratings arose in the context of health professions education. Specifically, we were interested in improving the assessment of student achievement in real or high-fidelity simulated practice settings, that is, assessment of their clinical performance. Clinical performance appears to be almost archetypical of complex performance in a complex setting. We shall explicitly address only the restricted domain of health professionals' clinical performance. Nevertheless, the discussion has direct implications for other areas which share the common elements of reliance upon rater judgement and the assessment of something that is intrinsically complex.

Because the membership of AERA Division I (Professions Education) is quite heterogeneous and at the specific request of two of the reviewers of our paper proposal, we first provide a fairly discursive conceptual, intuitive discussion of factors affecting rating reliability and validity. The rating process is presented in contrast to the objective testing process because the fundamentals of test design and analysis concepts and statistics are fairly broadly understood in the division. Latent-trait theory is then introduced in the same way: first, as it applies to objectively scored tests; then, we present our proposed latent-trait theory of performance rating and a simplified

Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, 1981.

model of it. The balance of the paper presents the specific research objectives, method, results, discussion and conclusions from empirical tests of our rudimentary theory. Briefly, we found what we consider substantial support for our proposed model where it may be appropriately applied.

Problem

One can get reliable and valid ratings-based measures of complex human performance using a very few well trained raters or by averaging across a larger number of less well trained raters if all of them rated all subjects under controlled circumstances. What the current state-of-the-art does not provide is a useful way to extract reliable, valid ratings from the kind of dirty and incomplete data sets ordinarily available. Dirty rating data is produced by lack of control which permits extraneous factors to influence the ratings given. Such things as inadequate rater training, poorly validated rating procedures and forms, variability in conditions under which performance is rated all tend to produce dirty data. Incomplete data sets are those in which not all raters rate all subjects.

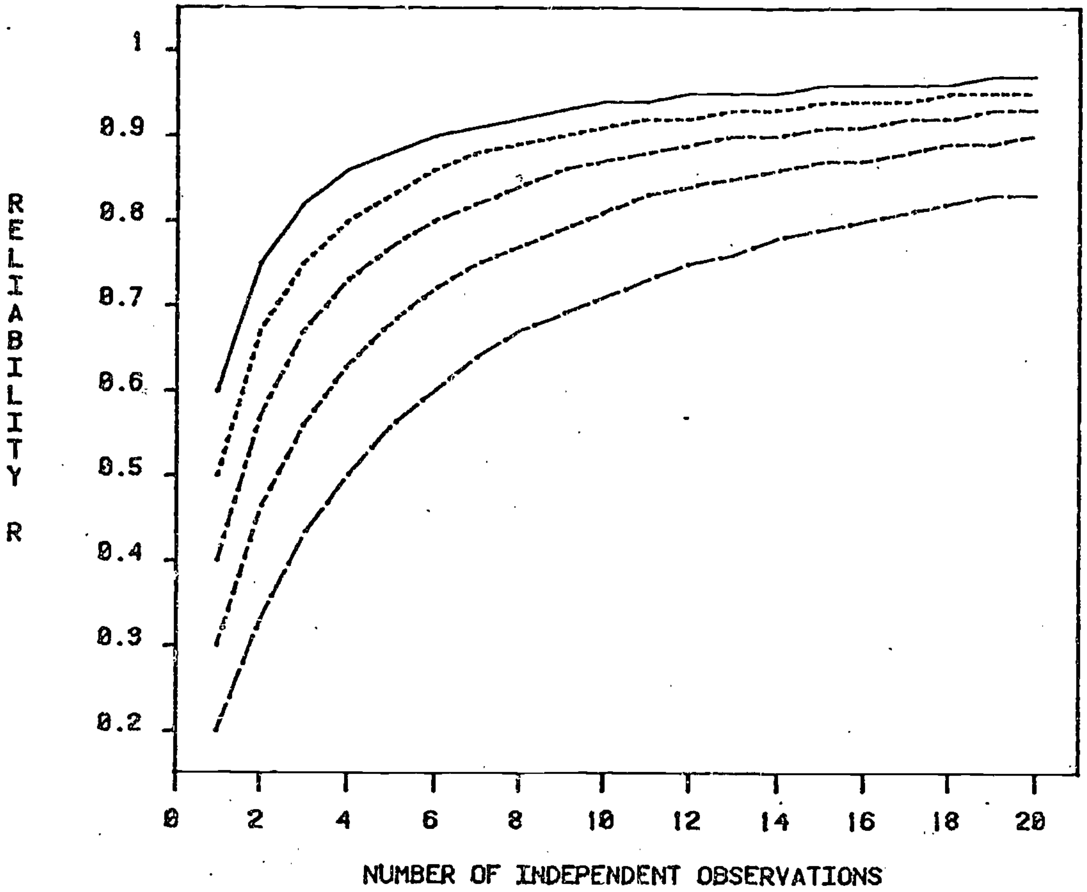
Any significant steps toward the resolution of this problem would have immediate beneficial effects in the practical evaluation of complex performance in ordinary settings and in research in which complex performance is a variable of interest.

Some Factors Affecting Reliability and Validity

No measurement, whether a test score or rating, may be more valid than it is reliable. Reliability sets an upper limit on the potential validity of the measure. Neither individual items nor individual raters are perfectly reliable measurement instruments in the sense of being completely accurate, stable, and consistent. In classical test theory and traditional practice, an individual item's reliability is measured by either its mean correlation with all other items on the test or its correlation with the total test score. Both of these give essentially the same result and are equivalent to the test item's expected correlation with another randomly chosen single item from the same content domain. Depending upon the calculational procedure used, an individual item reliability may be called a correlation of some kind or a discrimination index. Similarly, the reliability of the ratings given by a single judge is equal to the expected value of the correlation between this judge's ratings and the ratings of another independent, randomly chosen qualified judge. Two strategies, separately or in combination, may be used to improve the reliability of either a rating or a test score: use more or use better.

Spearman-Brown's test reliability formula was first developed to provide an estimate of how much the reliability of a test's total score would be changed by adding or deleting test items. Remmers, Shock, and Kelly (1927) demonstrated that pooling (i.e., summing or averaging) ratings across raters (where all or representative subsets of raters rate all subjects) had the same effect as pooling the item scores on an objective test. This means Spearman-Brown's formula is equally applicable to both items on tests and ratings provided by independent raters. Figure 1 depicts the relationship defined by Spearman-Brown's formula between the reliability of the total score, reliability of individual item scores or ratings, and the total number of independent items or ratings pooled (i.e., summed or averaged) together.

FIGURE 1. RELIABILITY AS FUNCTION OF OBSERVATIONS: ITEMS OR RATINGS



Under ordinary "real world" circumstances most ratings are obtained where many or all of the following conditions prevail: (a) raters have had no systematic training in rating based upon the use of standard stimuli and corrective feedback; (b) raters receive no or little information regarding how other raters rate the same subject under equivalent circumstances; (c) the scales used are vaguely defined as are the meanings of the individual point values or scale categories; (d) different raters do not observe the same performance under the same conditions; (e) not all subjects are rated by all raters, frequently none of the subjects are rated by all raters; (f) not all raters rate all subjects, frequently no rater rates all subjects; and (g) subsets of raters are not representative of the rater pool. On the basis of empirical evidence, Symonds (1931) concluded that under these kinds of ordinary circumstances the correlation between independent pairs of raters (i.e., the reliability of a single rater) is typically around r=0.55. The region between the upper two lines in Figure 1 approximates the reliability of pooled ratings as a function of the number of independent ratings under typical conditions (assuming all or representative subsets of raters rate each subject). Clearly one way to improve the overall reliability of either a rating or a test score is to base it upon more ratings or test items.

Alternatively, the reliability of each individual test item or rating may be improved. In testing practice, this is accomplished by selecting only those individual items which have had reliabilities above a specified value when used in earlier administrations of the test. Nunnally (1967) suggests a minimum individual item reliability of between r=0.10 and r=0.20. When this rule is used on the typical classroom objective test, the mean individual item reliability generally falls between r=0.20 and r=0.30. The lower two curves in Figure 1 define the region of expected total test score reliability as a function of (a) typical average item reliability and (b) the number of preselected items the test contains. Selecting the most reliable raters may occasionally be helpful; but, under typical circumstances more is gained from pooling across all available rating data rather than discarding the least reliable and pooling the remainder.

Efforts to improve the reliability of individual rater judgements (and thereby the reliability of the individual rater) are generally directed towards eliminating the conditions (described above) under which ratings tend to be made in real world settings. Frequently, they rely upon techniques such as improving the precision of the definitions of the attribute to be rated and values on the scale. Often this is implemented in the form of a behaviorally anchored rating (BAR) scale (Smith and Kendall, 1963; Landy and Barnes, 1979). However, when BAR scales are used in otherwise typical rating circumstances there is a dearth of data indicating any improvement over non-BAR scales. For example, Davidge, Davis, and Hull (1980; also in Dielman, Hull, and Davis, 1980) report full scale interrater reliabilties for individual house officers (residents) of r=0.61 and for individual attending (faculty) physicians of r=0.41. Davidge, et al.'s results are for the use of a very carefully designed BAR scale for measuring medical students' clinical performance. The reliabilities straddle Symond's (1931) value for reliabilities obtained under typical (non-BARS) rating conditions. We obtained a mean interrater reliability for an individual rater of r=0.50 across attendings and residents at two training sites who used a non-BAR scale inventory to rate the clinical performance of Junior year medical students (Cason and Cason, 1979). In the same paper (Cason and Cason, 1979), we concluded that in most of the published literature on rating health care professionals' clinical performance, the single factor most influencing the reported reliability of the total rating was the number of independent raters across whom it was summed or averaged.

A BAR scale used in conjunction with rigorous rater training can improve rater reliability over the value of r=0.55 reported by Symonds (1931) for typical rating circumstances. Stillman (1980) has achieved interrater reliabilities of r=0.85 and intra-rater reliabilities of r=0.90. Stillman obtained these results using the behaviorally anchored, empirically validated Arizona Clinical Interview Rating Scale in conjunction with rater training. The rater training was based upon use of standard stimuli (video tapes of interviews) and informative feedback to the rater. The program has proved successful in training raters belonging to three distinct groups: physicians, nurse practitioners, and "programmed patients". Stillman's results are directly attributable to her program's success in eliminating many of the conditions found in typical rating settings. While there are obvious practical obstacles to emulating Stillman's approach, her results provide a good benchmark for what can be accomplished (at least in some settings) when sufficient interest, skill, and resources are available.

It has long been acknowledged in both the folklore and research literature relating to rating that raters may vary in their general tendency to be stringent or lenient. This variation can affect reliability. Ebel (1951) has suggested two ways of applying Snedecor's (1946) (intraclass) reliability formula depending on whether variations in rater leniency could affect the stability of subject's mean (across raters) ratings. The first method applies when all raters rate all subjects. When there is variation in rater leniency, the first method yields a higher value than does the second method. This first method ignores any differences between the means of ratings given by different raters in the same way as does an ordinary (Pearson product-moment) correlation coefficient. For example, if rater A assigned

2, 6, 5, 3, 4

to five subjects in succession, and rater B assigned

5, 9, 8, 6, 7

to the same five subjects rated in the same order, the correlation between the ratings is r=1.00. Yet, rater B is systematically more lenient than rater A. When all (or representative subsets of) raters rate all subjects there is no systematic effect of rater leniency on individual subject's mean ratings. By contrast, when subjects are rated by different (non-representative) subsets of raters, the mean of the observed ratings on each subject is a less accurate measure of the subject's performance because some subjects are rated by a more lenient group of raters than are other subjects. The second method for estimating interrater reliability suggested by Ebel, unlike an ordinary correlation coefficient, takes into account differences in rater leniency and thus yields a smaller and more appropriate reliability value.

There is no shortage of evidence that different categories of health professionals vary in their leniency when called upon to rate the same performance under ordinary (i.e., poorly controlled) conditions. For example, ratings of Junior medical students by residents (house staff) have been consistently and widely reported to be more lenient than are those given by faculty (attending) physicians (Printen, Chappell, and Whitney, 1973; O'Donohue and Wergin, 1978; Pierlioni, Clark, and Dudding, 1979; Cason and Cason, 1979; Dielman, Hull, and Davis, 1980). The same studies also indicate the presence of variation in the leniency of raters in the same category.

Exemplary programs such as Stillman's can sometimes reduce variations in rater leniency to the point where it is no longer of practical importance as a source of inaccuracy in ratings (Stillman, Brown, Redfield, and Sabers, 1977; Sabers, 1981: personal communication). Nevertheless, when Meskauskas and Norcini (1980) discuss the problem of variability in rater leniency, in both standards setting and rating performance, they suggest the need to go beyond the things found in programs such as Stillman's. Meskauskas and Norcini suggest that in both standards setting and performance rating judges' ratings should be "handicapped" (i.e., corrected or adjusted) for variation in the judges' leniency by applying methods presented by Stanley (1961). Meskauskas and Norcini appear to be implying that it is at least difficult if not impossible to reduce rater leniency variation below the level of practical concern entirely through the use of BAR scales in conjunction with rater training.

Stanley's (1961) methods allow one to both determine the extent of variation in rater leniency and develop correction formulas for each rater. Stanley's analysis-of-variance related procedures allow the determination of the separate contribution of rater leniency and subject performance to the variation in the observed rating data. However, Stanley's procedures may be applied only when all raters have rated all subjects, i.e., to data sets with no missing data. But, as Stanley points out (and as was implied above in discussing Ebel's procedures) if all raters have rated all subjects, there is no need for adjusting the ratings. When all raters have rated all subjects, the mean or sum of the raw ratings on any subject is as valid and reliable as can be produced by any adjustment for rater leniency. Although correction formulas for raters developed at one time (when all raters rated all subjects) might be used later when subjects were rated by only (potentially non-representative) subsets of raters, this would be defensible only after it had been demonstrated that individual raters' relative leniency remained stable over time.

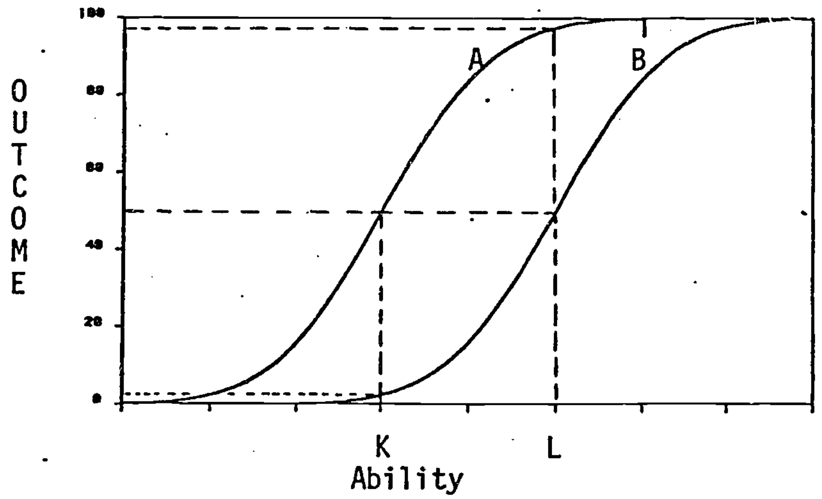
In summary, if one desires to obtain a highly reliable and valid assessment of a complex human performance based upon ratings from human judges, the current state-of-the-art, as suggested in the literature reviewed above, indicates that a model assessment program would include: (a) carefully trained raters; (b) empirically validated, behaviorally anchored scales; (c) controlled, uniform conditions under which performance is observed and rated; (d) multiple raters for each subject; (e) all raters (or representative subsets of raters) rate all subjects; and (f) use of the mean rating (across raters) obtained by a subject as the best available measure of the subject's true performance. In actual settings most of these conditions are hard to satisfy. Having more raters per subject (d) can be used to offset shortcomings in conditions "a" through "c" but only if condition "e" is satisfied. Otherwise variations in rater leniency will lower the reliability and validity of the outcome. However in practice, condition "e" is frequently not satisfied.

Although the theory we set forth below was neither derived from nor motivated by the applications of latent-trait theory to objective testing, we have discovered, with the benefit of hindsight, that our theory is most easily grasped by someone already familiar with the general schema of latent-trait theory as applied to objective testing. Consonant with the expository strategy used above, we have chosen to begin with the more familiar ground of testing, then go on to our theory of performance rating.

Latent-Trait Theory

Latent-trait test score theory (Lord, 1952; 1953; Baker, 1977; Hambleton, Swaminathan, Cook, Eignor, and Gifford, 1978) proposes to account for the score on an individual test item of an individual person. In the theory's simplest form, the probability that the person will answer an item correctly is determined by two factors: the person's true ability and the item's intrinsic difficulty. Item difficulty and person ability are both assumed to reflect the operation of some underlying (i.e., not directly observable, therefore latent) trait, attribute, or factor; for example, the attribute of knowledge. A person with much knowledge would be located high on the latent knowledge scale. Similarly, an item requiring great knowledge to be correctly answered would be located high on the knowledge scale. The probability that a person of a given ability will correctly answer an item of a given difficulty is defined by an "s-shaped" item characteristic curve. Figure 2 gives hypothetical characteristic curves for items A and B. By convention, item A is said to have difficulty K or to be located at point K on the latent scale. A person with ability K (i.e., located at point K) has a 0.50 probability of correctly answering item A. The item characteristic ("s-shaped") curve defines the exact relationship between a person's ability (at any point on the latent scale) and that person's probability of correctly answering that item. Consider Figure 2: a person of ability K has a near zero

Figure 2. Characteristic Curves



probability of answering item B. While a person with ability L has near a 1.0 probability of answering item A correctly, this person's probability of answering item B correctly is only 0.50.

Probably the greatest number of latent-trait theory applications have been based upon the Rasch (1966) measurement model. This may be largely attributed to the work of Ben Wright and his colleagues (Wright, 1968; Wright and Stone, 1979; Mead, Wright, and Bell, 1979) such as their development of techniques, including computer programs, which make Rasch analysis easier; as well as, their zealous advocacy of Rasch measurement techniques. The defining characteristics of the Rasch model are (a) only one parameter, location on the latent-trait, is used to characterize each person or item; and, (b) the "s-shaped" item characteristic curve is operationally defined by the logistic function. Other models of latent-trait test theory include additional factors (e.g., item discrimination, a guessing factor, and so forth) in their characterization of test items and people and/or define the characteristic curve using a different mathematical function, e.g., the normal ogive.

Irrespective of what particular model of latent-trait theory is used, the usefulness of the model rests upon the (testable) assumption of parameter invariance. In contrast to conventional test item statistics (e.g., difficulty index and discrimination index) and norm-referenced test scores, the parameter values for item difficulty and person ability are independent of the context of both the particular group of people who took the test and the particular set of items in the test. This may be most clearly explained by analogy to the physical measurement of temperature in the days when chemists (or alchemists) made their own thermometers.

In Figure 3, the horizontal lines T1, T2, and T3 are thermometers. The letters "A" through "O" represent specific observed melting and boiling points for various materials, e.g., alcohol, water, paraffin, lead, and so forth. Note that T1 and T2 share points "B" and "D". T2 and T3 share points "I" and "M". But T1 and T3 share no observed points in common. No matter how the individual thermometers were originally graduated or where their arbitrary zero points were placed, the relative positions (ordering) and distances between observed melting and boiling points would remain the same. Thus, the observations that are in common to two thermometers can be used to calibrate the measurements on one thermometer against the other. Because T1 and T3 are linked through common observed points on T2, the information on all these instruments can be placed on a single temperature scale running from "A" to "O". The location (parameter) of a melting point of one material is invariant with respect to the relative locations (ordering and distances) of other melting points.

Figure 3. Invariance of Parameter Locations: Ordering and Relative Inter-point Distances

```
T1:  A...B.........D.........G

T2:      B.....C...D..E.............I....J.......M.N.

T3:                     F.......H...I.........L..M....O


     A...B.....C...D..E.F....G..H...I....J..K.L..M.N..O
```

Latent Attribute

Latent-trait analysis of the responses of a group of people to a group of items on a test produces estimates of their locations (i.e., true ability of persons, intrinsic difficulty of items) on an underlying trait. Figure 3 can be used to represent different objective tests (i.e., T1, T2, and T3) with the letters being either items or people or both. When this is done, one can make very concrete predictions about a person's performance on items to which that person has not previously responded. Also, the results of a test composed of any combination of the items whose locations are represented by the letters "A" through "O" could be translated into equivalent scores for tests T1, T2, and T3 because all the items can be calibrated against each other. This is all possible because, like melting points, the location (difficulty) of items on the latent trait are invariant with respect to their ordering and inter-item distances. Likewise, relative positions of person abilities are invariant with respect to other persons' abilities and item locations. By contrast, conventional item statistics reflect only the relationship between a particular group of examinees (or a similar group) and the particular items on the test. For example, an item's difficulty index (unlike the item's intrinsic difficulty) simply indicates the proportion of examinees that correctly answered it, or would be expected to correctly answer it in a comparable group of examinees. The conventional item discrimination index is similarly limited in meaning and usefulness. Parameter invariance is the characteristic of latent-trait models which make them uniquely useful.

Not surprisingly, many Rasch applications are designed to capitalize upon parameter invariance to generate equivalent tests composed of different items or equate the results of one test with another having overlapping items. This is clearly illustrated by Anderson, Baker, Laguna, and Laguna's (1980) use of the Rasch model to obtain comparable test scores based on overlapping but not identical sets of test items in Neurology clerkship examinations. Anderson et al.'s work is exceptional in that it involved an application of the Rasch model to classroom level data sets containing only 7 to 10 students per exam. More commonly, Rasch techniques are applied when the number of persons who have responded to the items is 200 or more. The uncertainty (measurement error) associated with an item's difficulty tends to be much bigger than that associated with a melting point. In practical work it is not unusual for a small percent of the items in a given test to not fit the Rasch model. These are identified and discarded so that they do not adversely affect the estimation of the intrinsic difficulties of the remaining items.

Anderson et al. cite several Rasch applications in health professions education including: a pharmacy externship (Smith and Kifer, 1980), analysis of the Medical College Admission Test sub-part scores (Cromier,1977), and analyses of tests of the National Board of Medical Examiners (Hughes, 1979; Kreines and Mead, 1979). Schumaker (1979) applied the Rasch model to the problems of equating medical examinations. Harasym (1981) used Rasch techniques in comparing Nedelsky's (1954) and a modified form of Angoff's (1971) procedures for setting passing standards for objective tests.

Our Rudimentary Theory of Performance Rating

We propose that the rating obtained by a subject is a function of the subject's achievement and the rater's leniency and sensitivity. Neither achievement nor leniency is directly observable; but, each underlies and partially accounts for observable behavior. Subject achievement accounts for subject performance only in part. Factors such as illness, inappropriate working conditions, action or inaction of others (e.g., a hostile co-worker or examiner) can either improve or reduce the quality of the observed performance regardless of the subject's true level of achievement. Similarly, the rater's leniency and sensitivity account in part for the ratings given but the ratings also reflect the performance that was observed and rated.

Both rater leniency and subject achievement are measured upon a scale of the same latent trait, factor or attribute. Generically, this underlying trait is called an ability and could be any skill, competency, or disposition, whether innate or acquired. Leniency and achievement may each be represented by points on this ability scale. These points are called the rater reference point (RRP) and the subject achievement point (SAP) respectively.

The rater reference point (RRP) is used by the rater as an implicit standard for judging the perceived performance of the subject. The location of the rater reference point (RRP) embodies the rater's prior knowledge, understanding, and beliefs regarding (a) fundamental, ideal standards relevant to the trait at issue; (b) the subject (person) whose performance or product is to be rated; (c) the task or activity to be performed by the subject; (d) the constraints imposed by the setting upon either or both the rater and subject; (e) where problem solving (broadly construed) is involved in the subject's task, the intrinsic difficulty of the problem; and, (f) related factors. The rater reference point may be viewed as arising from an adjustment the rater makes to some implicit, fundamental standard. The fundamental standard is appropriate only to an ideal set of rating circumstances, i.e., conditions under which nothing but the standard and the performance need be considered in determining the rating. The rater reference point (RRP) results from the rater's effort to take all the discrepancies between an ideal setting and the actual one into account prior to assessing the subject's performance. The rater reference point (RRP) embodies all factors which systematically influence the rating assigned except the subject's performance and effects related to the rater's resolving power and sensitivity.

Implicitly, the rater perceives the subject's performance as a deviation on the relevant ability scale from the rater's RRP. The size and direction (above or below the RRP) directly equals the distance from RRP to the subject achievement point (SAP) on the ability scale, as judged by this rater. The rating assigned is a function of the difference between RRP and SAP.

The rater's resolving power, i.e., the precision of the rater's judgements as embodied in the assigned ratings, is greatest when the difference between RRP and SAP is minimum. Resolving power diminishes in an accelerated manner as the difference between RRP and SAP increases. Generally, small differences in value for SAP's near the RRP result in substantially different assigned ratings. As distance from the RRP increases, larger and larger differences between two SAP's must be present for there to be an appreciable difference in the corresponding assigned ratings. These relationships are analogous but not equivalent to those of visual resolving power. Objects close to the observer need not be separated from each other by very much to be seen as distinctly not at the same distance. But as distance from the observer increases, the distance between objects must increase if they are to be recognized as being at different distances from the observer. Because resolving power diminishes in an accelerated manner as distance from RRP to SAP increases, the rater characteristic curve (RCC), which specifies the rating assigned as a function of the difference between RRP and SAP, is one of a family of smooth, continuous, "s-shaped" curves. (A member of this family of curves is commonly called an ogive, e.g., the normal ogive.)

Some rater's have greater sensitivity than do other raters. Variation in sensitivity between raters is defined by differences in the rate of acceleration in change of resolving power. However, rater sensitivity is somewhat more easily grasped intuitively in terms of the difference in subject achievement associated with a given pair of ratings, for example 10% (of possible points) and 90%. A highly sensitive rater would give these ratings when there was a relatively small difference in two subject's achievement. A less sensitive rater would give these ratings when there was a relatively much larger difference in the achievement of the two subjects. The limit of hypersensitivity is characterized by a rater that gives only minimum or maximum ratings. Any SAP less than the hypersensitive rater's RRP receives a rating of 0%; any SAP equal to or above this rater's RRP receives a rating of 100%. Graphically, the hypersensitive rater's characteristic curve (RCC) is no longer a continuous, smooth curve. It has become two horizontal lines, one at 0% extending down the ability scale from the RRP; the other at 100% extending from the RRP up the ability scale. By contrast, the limit of hypo-sensitivity is characterized by a rater who assigns all SAP's the same value as if they were no different from this rater's RRP. Graphically, the hypo-sensitive rater's characteristic curve has become a horizontal line extending indefinitely in each direction from the RRP parallel to the ability scale at the rating level associated with this rater's RRP.

The measure of rater sensitivity is the slope of the RCC at the point on the RCC directly above the RRP on the ability scale. The hypersensitive limit is defined by the value of the slope having become indefinitely large. The hypo-sensitive limit is defined by a RCC slope of zero. Neither limit occurs in practice, though they may be approached.

The theory of performance rating proposed above may be understood by analogy to latent-trait test theory. Instead of locating test items and examinees (persons), the proposed theory locates raters (persons) and subjects (persons or products) on an underlying trait. Item difficulty is replaced by rater leniency; probability of answering correctly is replaced by rating points assigned; and item discrimination by rater sensitivity. Reconsidering Figure 2, A and B are rater characteristic curves (RCC). Rater A has a leniency of K (i.e., rater A's RRP is located at K). Rater B has a leniency of L. A subject with an achievement point (SAP) located at L would receive a rating of 50% from rater B; and, a rating of near 100% from rater A.

As proposed, our theory is only rudimentary. Many things potentially characterizable as separate factors have been subsumed into the construct of rater leniency. For example, "cases", "problems", and "settings" (i.e., things with which the subject must contend) might be represented as a separate construct. Then we might be able to separate the components of rater leniency regarding the rater's estimation of task demands from the rater's leniency in assigning ratings when task demands do not influence the location of the rater's RRP. An analog to the "guessing parameter" sometimes used in latent-trait test theory might be the presumption of a "minimum existing competence". This would function to limit the minimum rating a rater would assign regardless of how poor the observed performance was. Elaborations such as these hardly seemed justified to us until some empirical tests of the more rudimentary version had been completed.

Simplifying Assumptions

To facilitate our initial empirical investigations we imposed the following simplifying assumptions upon the rudimentary theory presented above:

1. All raters have equal sensitivity. Under this condition the slope of the rater characteristic curve is no longer a measure of rater sensitivity; not even mean rater sensitivity. Any convenient unit (graduation) of measurement may be chosen for the ability scale. Even though a different size unit produces a different value for the slope, this does not imply a change in sensitivity because the relative distances among raters and subjects remain constant. When equal sensitivity is assumed, sensitivity becomes perfectly confounded with leniency and ability.

2. The rater characteristic curve evaluates the difference between a rater reference point (RRP) and subject achievement point (SAP) as the percent (%) of possible rating points.

3. The rater reference point (RRP) for any rater is located under that rater's characteristic curve (RCC) on the

ability scale at that point which evaluates to a rating of 50%. This appears to represent a potentially large and strongly counter-intuitive departure from the construct of the RRP as presented in the proposed theory. Intuitively it might seem that in typical rating circumstances a rater's reference point would be near some traditionally significant value, e.g., 75%. This arises in part from considering the RRP as if it were equivalent to the obstensible, conscious standards in common use. A careful examination of the definition of the RRP given above suggests that its relationship to such conscious standards may be very remote and complex. At any rate, we judged that the gains in mathematical and conceptual tractability had from imposing this assumption justified its use, at least during our initial empirical investigations.

Our Simplified Performance Rating Model

More formally, we propose that the ability scale upon which rater reference points (RRP) and subject achievement points (SAP) are located is an equal interval scale of arbitrary graduation (unit) and arbitrary origin (zero point). For the purposes of this research, we operationally define the rater characteristic curve (RCC) as the product of an arbitrary positive, constant scaling factor (SF) and the cumulative unit-normal deviate ogive. The scaling factor is abitrarily set equal to 100. The difference between a rater reference point (RRP) and subject achievement point (SAP) divided by the scaling factor (SF) gives an ability scale deviation value (z):

Formula 1

$$z=(SAP - RRP)/SF$$

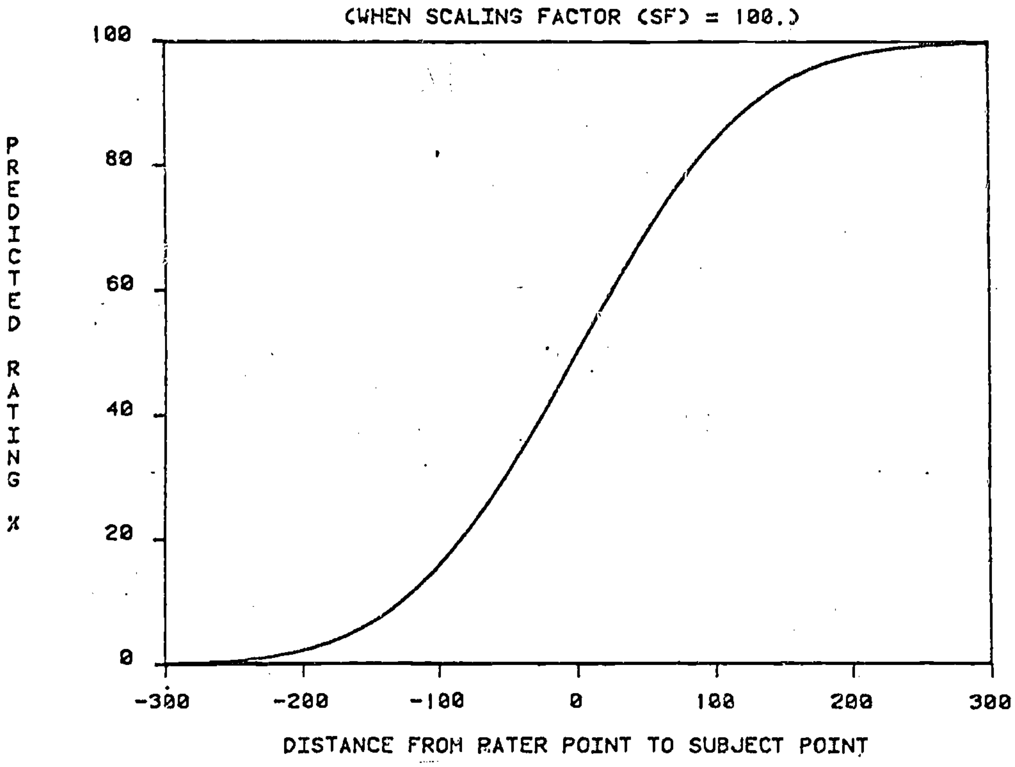
The proportion of possible rating points assigned for a given value of z is equal to the total proportion of area under the unit-normal curve below z, that is p(z). Multiplying the proportion p(z) by 100 gives the expected subject rating (ESR) in percent units:

Formula 2

$$ESR=p(z) * 100$$

The relationship between the expected subject rating (ESR) and the discrepancy between RRP and SAP is depicted graphically in Figure 4.

Figure 4. Expected Rating as a Function of Distance Between RRP and SAP



There may be variation in the rater's perception, knowledge, judgement, and so forth. Therefore, the observed subject rating (OSR) may contain error:

Formula 3

$$OSR=ESR+error$$

In Hambleton, Swaminathan, Cook, Eignor, and Gifford's (1978) terms, our model is somewhere between Lord's (1952; 1953) two parameter normal ogive model and Rasch's (1966) one parameter logistic model. Conceptually it is somewhat closer to Rasch's model, although it uses the normal ogive as does Lord's. It was not until our model was developed essentially to the level presented above that we somewhat belatedly recognized some of its conceptual and formal relationships to Rasch's and Lord's objective test measurement models.

Objectives

The objectives of the research reported here were to determine the extent to which a normal-ogive model of a proposed latent-trait theory of performance rating: (a) fit data of a type common to health professions education, i.e., dirty and incomplete ratings of clinical performance; (b) clarified and quantified the separate contribution of (1) all rater characteristics as embodied in the single theoretical construct of leniency and (2) the construct of the subject's underlying (i.e., latent) true achievement to the observed dirty and incomplete ratings; and, (c) appears to provide a basis for generating more reliable and valid measures of performance than the mean of the observed ratings on a subject when the rating data is not only dirty and incomplete but the subsets of raters are unrepresentative of the whole relevant rater pool.

Method

Data Source. Data analyzed were samples of convenience available from a project whose objective was to develop a machine based system for processing clinical performance data. As part of that project, a prototype machine readable (optically scanned) form was used experimentally (Cason and Cason, 1979). Data collected on this experimental form were analyzed here.

Subjects and Cohorts. The subjects upon whom rating data were available were third year medical students enrolled in a medicine clerkship, i.e., a clinically oriented course in internal medicine. Data were available from the third and fourth cohorts (i.e., groups of students concurrently taking the course) in academic year 1978-79 and the second cohort in 1979-80. The third cohort took the course during the winter months; the fourth during the spring; and, the second during the fall. Table 2 gives the number of students in each cohort.

Clerkship and Setting. The medicine clerkship was 12 weeks long with six weeks spent at each of two training sites: University Hospital and Little Rock Veterans Administration Hospital. In the wards, instruction was entirely tutorial and small group based. Faculty attending physicians and residents each had a small number of (usually at least two but less than six) medical students randomly assigned to them for instruction. Residents tended to have more contact with students than did the faculty.

Rating Instrument. The machine processable form contained a 33 item clinical performance rating inventory. The items were divided into seven non-overlapping categories. Raters could assign a rating value of from 1 to

5 to each item or indicate that it was either not observed or was not applicable. Rating values were defined in explicitly norm-referenced terms rather than being behaviorally anchored. For example, a rating of "4" was defined as "A little better than the typical student in the typical class (i.e., would be in the top 25% but below the top 10%)". Appendix A contains a facsimile of the form. For scores on the full inventory (i.e., mean of valid ratings to all items), previous research (Cason and Cason, 1979) indicated a mean interrater correlation of r=0.50; ranging from a high of r=0.71 between residents and faculty at the same training site to a low of r=0.23 for ratings given by residents at one site and faculty at another.

Raters and Rating Procedures. The raters were the faculty attending physicians and residents who trained the medical students. Most students were rated by two attending physicians and one resident at University Hospital and by one attending and one resident at the VA Hospital (mode=5 ratings/student). Raters received a 20 minute oral explanation of the proper use of the rating form (from G. Cason) and a written memorandum restating the details. No other rater training was used. At the conclusion of the six weeks students spent at a training site, raters completed a form on each student with whom they had contact. Raters entered only rating data. The various identification data grids were completed by a departmental clerk. After the forms were optically scanned and an electronic (computer disk file) copy made, they were placed in the respective students' permanent files. The number of raters for each cohort is given in Table 2. The number of raters overlapping cohorts (i.e., rating students in more than one cohort) is given in Table 3.

Dependent Measure. The dependent measure of clincial performance was operationally defined as the mean valid rating across all items in the inventory, rated by one rater, expressed in percent form. A valid rating was any rating of 1 through 5. Blanks, multiple marks, not applicable and not rated were non-valid ratings. Although the inventory contained items of both the affective, interpersonal skills type and the cognitive, technical, problem solving type which prior research (e.g., Davis, Hull, Davidge, and Dielman, 1979) indicated belong to statistically independent (orthogonal) factors, the global trait represented by the mean across all items, i.e., overall achievement in clinical performance, was chosen. This was done because (a) with missing data at the item level, unbiased estimates of the separate factor scores could not be obtained with any certainity; (b) extracting factor scores (by factor analysis) is a scaling procedure which results in "cleaner" scores, thus results of further analyses based upon these factor scores might be contaminated by and attributed to the effects of the factor analysis; (c) the only available unbiased measure of both student performance and rater judgement was the mean of the valid ratings across all items on the inventory.

Estimation of RRP's and SAP's. Program MERLIN (Cason, 1980) was used in conjunction with subroutine STEPIT (Chandler, 1965) to obtain least-squares estimates of the rater reference points (RRP) and subject achievement points (SAP). Briefly, MERLIN operates as follows. An observed data table with one row per subject and one column per rater is input. All observed subject ratings (OSR) are contained in this data table. A set of "best guesses" for the RRP's and SAP's are input. In actual practice, we started with very bad guesses:. all RRP's and all SAP's equal to 500. The program uses these starting guesses for the SAP's and RRP's and the function depicted in Figure 4 to calculate an expected subject rating (ESR) for every cell in an expected data table. Then, the discrepancy between each value in the observed data table and its corresponding value in the expected data table is found and squared. When all the squared values are summed, the result is the error sum-of-squares (ESSQ) for the fit between the predicted ratings generated from the current set of "guesses" for the SAP's and RRP's and those ratings actually observed. STEPIT is used to successively alter (i.e., step) the guesses for the parameters and evaluate the impact on the resulting fit. When changes to the parameter values no longer produce appreciable improvement in the fit (reduction in the error-sum-of-squares) between the observed and predicted, MERLIN outputs a series of reports. These reports include the least-squares estimates of the RRP's and SAP's, the complete table of predicted ratings, measures of final fit (r and ESSQ), results of an F-test between the proposed model and the null hypothesis, and so forth. This process requires that one parameter be fixed (i.e., held at a constant value throughout the estimation process) to anchor the scale. A senior faculty member who rated at least 6 students in each of the cohorts was used for this. This rater's RRP was held fixed at 500.

MERLIN was run on a Digital Equipment Corporation System 10 (DEC-10). Parameter estimates were determined on each cohort's data separately. Central processing unit (CPU) time required to find least-squares estimates was as follows: Cohort 1978-79:3 with 75 free parameters to be estimated required 82 minutes of CPU time; Cohort 1978-79:4 with 47 free parameters required 29 CPU minutes; Cohort 1979-80:2 with 63 free parameters required 36 CPU minutes.

Results

Fit was determined for four models on each cohort's data separately. Thus, each cohort represented an independent replication.

Model A was the model proposed above with one free (RRP) parameter per rater (except for one which was fixed at 500 to anchor the scale) and one free (SAP) parameter per student. Model A permitted, but did not require that, both rater leniency and subject achievement contributed to the fit between the predicted and observed ratings. If there were no appreciable differences in raters' leniency, the least-squares values of the RRP's found by MERLIN would all be near the same value (i.e., 500). Similarly, if there were no appreciable differences in students' achievement, the least-squares values for all the SAP's found by MERLIN would be near the same value. Table 1 provides descriptive statistics (means and standard deviations) for the estimated values of Model A's RRP's, SAP's, as well as observed ratings for each cohort. Means for each of these variables were quite similar across all three cohorts. Model A was the most general model considered. Models B and C were derived by imposing restrictions upon Model A.

Table 1

Means and Standard Deviations (SD) for RRP's, SAP's, and Ratings Based upon the Full Data Set

	RRP		SAP		Observed Ratings	
Cohort	Mean	SD	Mean	SD	Mean	SD
78-79:3	485.50	38.17	558.03	37.72	73.49	11.75
78-79:4	476.60	37.64	549.48	23.99	74.19	8.16
79-80:2	486.49	21.17	545.52	27.13	72.55	7.77

Model B imposed the restriction that all raters are equally lenient, i.e., all RRP's equal 500, while allowing SAP's to vary. This restriction forces the predicted ratings for the raters of a single student to be the unweighted mean of the observed ratings of these raters on this subject. This is the model corresponding to the common practice of using the mean of the observed ratings as the best measure of the student's true performance. Note however that it is accurate only within the context of equal rater leniency. When contrasted with Model D (null hypothesis), Model B provided a mechanism for determining how well variation in student performance could account for observed ratings. Also, statistical contrast of Model B (achievement) with Model A (both achievement and leniency), provides a way to determine if rater leniency contributed to observed ratings beyond that accounted for by student achievement. A statistical difference between A and B indicates a "leniency main effect".

Model C imposed the restriction that all students had equal achievement, i.e., all SAP's equal 500, while permitting all the RRP's to vary. When contrasted with Model 0 (null hypothesis), Model C provided a way to determine the extent to which variation in the observed ratings may be accounted for by variation in rater leniency. Also, when contrasted with Model A (i.e., both achievment and leniency), Model C (leniency) provides a way to determine if student achievement makes a significant contribution to observed ratings beyond that which could be ascribed to variations in rater leniency. A statistical difference between Models A and C indicates an "achievement main effect".

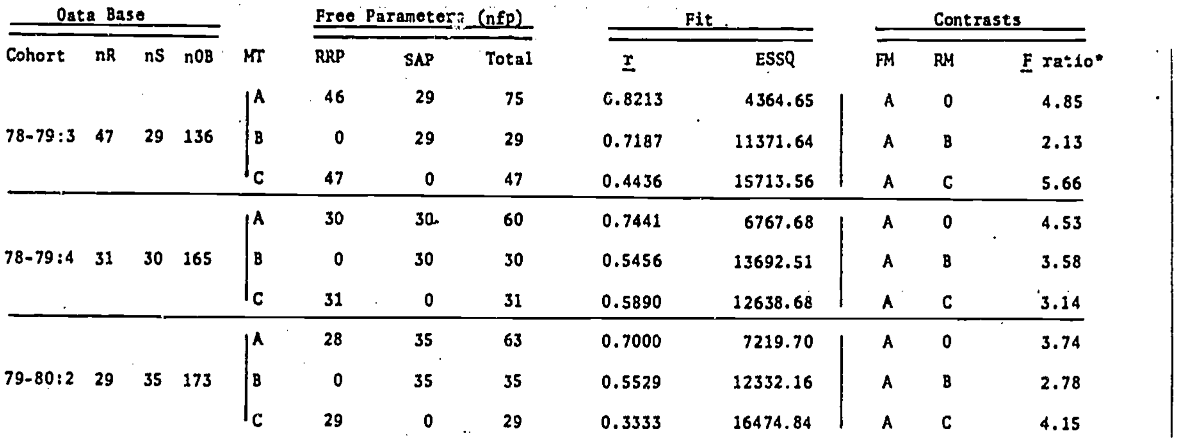
Model 0 embodies the null hypothesis, i.e., a model which accounts for the observed data as chance (random) variation from the overall mean rating (across all raters and students). Models B and C were not "straw-men" intended to make the proposed model (A) look good. All three hypothetical models must be used in contrast with each other and with the null hypothesis to determine the relationships of interest.

Table 2 presents the results of formal, statistical contrasts between the proposed model (A), as the full model (FM) and each of the others (e.g., B, C, and 0) as the restricted (RM) model (Ward and Jennings, 1973; see also Sternberg, 1967). All the F-tests resulting from the contrasts reported in Table 2 produced statistically significant F's ($p<0.01$). Table 2 also provides measures of the fit between each model and the three data bases. The fit is indicated both by the correlation (r) between the observed and predicted ratings and by the associated error-sum-of-squares (ESSQ). In all three cohorts, the proposed model (A) fit better ($r=0.82$, 0.74, 0.70) than chance ($p<0.01$), better than Model B ($r=0.72$, 0.55, 0.55; $p<0.01$), and better than Model C ($r=0.44$, 0.59, 0.33; $p<0.01$). The contrasts between models A, B, and C indicated that both rater leniency and student achievement made statistically significant ($p<0.01$), independent contributions to the observed ratings in all three cohorts. In conventional analysis-of-variance terminology, the results supported the conclusion of a signficant ($p<0.01$) rater leniency main effect and a significant ($p<0.01$) student achievement main effect in each of the three cohorts.

Table 2

Contrast of Fit of Models A, B, and C to Data from Each of Three Junior Year Medicine Clerkship Cohorts

Table 1. Contrast of Fit of Models A, B, and C to Data from Each of Three Junior Year Medicine Clerkship Cohorts



Oata Base					Free Parameters (nfp)			Fit		Contrasts		
Cohort	nR	nS	nOB	MT	RRP	SAP	Total	$\underline{r}$	ESSQ	FM	RM	$\underline{F}$ ratio*
				A	46	29	75	G.8213	4364.65	A	0	4.85
78-79:3	47	29	136	B	0	29	29	0.7187	11371.64	A	B	2.13
				C	47	0	47	0.4436	15713.56	A	C	5.66
				A	30	30	60	0.7441	6767.68	A	0	4.53
78-79:4	31	30	165	B	0	30	30	0.5456	13692.51	A	B	3.58
				C	31	0	31	0.5890	12638.68	A	C	3.14
				A	28	35	63	0.7000	7219.70	A	0	3.74
79-80:2	29	35	173	B	0	35	35	0.5529	12332.16	A	B	2.78
				C	29	0	29	0.3333	16474.84	A	C	4.15

*For all reported $\underline{F}$'s, $p \leq 0.01$. MT= model type; n=number; R=raters; S=students; OB=observations(ratings); r=Pearson correlation; ESSQ=error sum of squares; FM=full model; RM=restricted model; df_1=nfpFM-nfpRM; df_2=nOB-nfpFM.

Because the study was replicated on three independent data bases and the same results were obtained on each, the joint probability across all three cohorts for each of the results cited above was p<0.000001. The probability values given in the prior paragraph refer to each data base considered separately. When all were considered together the smaller value just given should be substituted for the earlier ones. Partitioning the variance by contrasting models A, B, and C, we found that in these data about 20% of the variability in clinical performance ratings could be attributed to variations in rater leniency. An additional 35% could be attributed to variation in student achievement. Taken together these results strongly indicated that while a knowledge of either leniency or achievement provided a significantly better than chance basis for predicting ratings, each was a statistically independent factor, and the best accuracy in prediction was achieved on the basis of a knowledge of both. These results directly support the proposed model and thereby indirectly the proposed theory: performance ratings were a function of both rater leniency and subject achievement.

As some raters rated students in more than one cohort, it was possible to calculate a "test-retest" reliability coefficient for the rater reference points (RRP) of these raters. Table 3 provides the reliability coefficients (r) determined on pairs of RRP's for each rater. The number (n) of raters who rated students in two cohorts is indicated in parentheses under the corresponding r value. The probability (p) of the observed correlation arising by chance is also given. All these reliabilities are positive but below r=0.30. Although no single one of these r's departed from a value of r=0.00 to a statistically significant degree (i.e., individual probabilities were p>0.15), at least two of these r's were statistically independent. From their joint, independent occurrence it was found that the set of r values differed significantly (p<0.04) from an r=0.00. This very important result provides directly validating evidence for the theoretical construct of leniency and indirect validation for the construct of achievement. For these raters, we found that while their RRP's were labile or difficult to measure with precision, their RRP's corresponded to some feature of their rating behavior that persisted over at least a six month period of time.

Table 3

Correlations between RRP's for same Instructors
Across Independent Cohorts of Students
All Available Data Used to Estimate RRP's

		78-79:4	79-80:2
	r	0.2796	0.2883
78-79:3	n	(15)	(13)
	p	0.1560	0.1700
	r		0.2483
78-79:4	n		(9)
	p		0.2600

The mean observed rating on each student was moderately well correlated (r<0.95) with the rating that the proposed model predicted a rater of mean leniency would assign. Assuming (on the strength of the evidence thus far reported) that the proposed model was valid, this result indicates that the leniency of the various sets of raters who rated these students were moderately representative of the whole pool of 75 different raters. This would be expected as assignments of students to raters was random. But, random assignment could produce highly different subsets of raters. Apart from the model under investigation here, there was no other technique for determining the representativeness of the rater subsets. The results only suggest that the rater subsets were moderately representative.

To further test the proposed model, a cross-validation of model predictions against an independent criterion was conducted. A restricted data set was created from the full data set. The full data set contained all the observed ratings on the three cohorts used in the analyses reported above. The restricted data set was formed by setting aside (i.e., "saving") one randomly chosen rating per student (with the constraint that the remaining restricted data set contained no rater who rated less than two students nor a student rated by less than two raters). Parameters (RRP's and SAP's) were then estimated on each cohort's restricted data set separately. Descriptive statistics (means and standard deviations) for the observed ratings, and RRP's and SAP's estimated for Model A from the restricted data set are given in Table 4. When compared with the values obtained on the full data set (Table 1) , the reduction of one observation per student had no significant impact on the means.

Table 4

Means and Standard Deviations (SD) for RRP's, SAP's and Ratings Based upon the Restricted Data Set

	RRP		SAP		Observed Ratings	
Cohort	Mean	SD	Mean	SD	Mean	SD
78-79:3	490.57	45.65	566.25	45.70	74.44	12.70
78-79:4	466.78	74.11	547.52	24.48	74.73	8.54
79-80:2	487.84	19.91	549.90	37.94	72.29	8.91

The saved ratings were then correlated with the corresponding elements in two different sets of predicted ratings: (a) those given by the proposed model (when its parameters had been estimated from the restricted data set); and, (b) those given by the model underlying the most common rating practice, i.e., Model B, which is equivalent to the mean of the ratings each student received in the restricted data set. In each case the saved ratings were independent of the predictions with which they were correlated.

This procedure could put the proposed model at a substantial disadvantage when contrasted with the alternate model (B). This arises from the reduction in data available to estimate parameters. By consulting Table 2, it can be deduced that in the full data set the ratio of observations to free parameters (to be estimated for Model A) was 1.8, 2.7, and 2.7 respectively in the three cohorts. In the restricted data set, these ratios declined to 1.4, 2.2, and 2.2. In cohort 1978-79:3 the ratio fell from an already marginal 1.8 observations per parameter in the full data set to a very doubtful 1.4 in the restricted data set. A low ratio could place the proposed model at a disadvantage because it had more parameters to be estimated. Less data per paramter would reduce the accuracy of the parameter estimates and thus the accuracy of the model's predictions. The alternate model having only about half as many parameters to estimate had an advantage in obtaining more accurate estimates of its parameters (i.e., one mean per student).

Table 5 reports the results of correlating an independent rating of each student with the prediction of the proposed model (A) and the prediction implicit in the common practice of taking the unweighted mean of the observed ratings (Model B) as the best available measure of performance. In two of the three cohorts the results appear to favor the proposed model, but in cohort 1978-79:3, Model B seems to be superior to the proposed model. This means that in two of the three cohorts predictions based upon a knowledge of both rater leniency and student performance (i.e., Model A) appeared superior to a knowledge of student performance alone (Model B). In cohort 1978-79:3, the prediction of Model A was not only less accurate (i.e., less well correlated with the criterion), the observed correlation (r=0.26) for Model A was not significantly different from r=0.0. Considering that Model B was a restricted case of Model A, Model A should do no worse than Model B.

Table 5

Correlations of Prediction of Models A and B with an Independent Rating on Each Subject

Cohort	A	B
78-79:3	0.2555	0.5020
78-79:4	0.6699	0.5531
79-80:2	0.4027	0.2022
Mean 1	0.5136	0.4465
Mean 2	0.6128	0.4055

For cohort 1978-79:3, the data indicate that very poor estimates for Model A's parameters were obtained from the restricted data set. The result of Model A fitting worse than Model B was directly attributable to the lack of sufficient data in the restricted data set for simultaneously estimating SAP's and RRP's. This "negative finding" was serendipitously suggestive of a useful rule of thumb. Anytime the correlation between the proposed model's predictions (when based upon the parameter estimation procedures in MERLIN) and independent criterion ratings fails to at least equal the correlation between the criterion and each subject's mean observed rating (i.e., the prediction of Model B), then there are insufficient data available to make useful estimates of the parameters of the Model A. In the case at issue, this interpretation was corroborated by an analysis of correlations between the values estimated for Model A's parameters (RRP's and SAP's) based on the full data set with estimates for the same parameters based on the restricted data set. The results of these analyses are reported in Table 6.

Table 6

Correlations between Parameters Estimated from the Full Data Set with Those Estimated from the Restricted Data Set

Cohort	RRP	SAP	Both	RSQ
78-79:3	0.8300	0.7991	0.8178	0.6688
78-79:4	0.9173	0.9691	0.9508	0.9040
79-80:2	0.8676	0.9625	0.9329	0.8703

These correlations would be high if the parameter estimates were stable. The correlations for RRP's and SAP's separately and combined indicated that there was good stability for the parameter estimates in cohorts 1978-79:4 and 1979-80:2. Taking the square of the correlation (RSG) between the two conditions (i.e., full and restricted data sets) as a measure of common variance, the stability of cohort 1978-79:3's parameter estimates was clearly poor (RSQ=0.67). Deleting one observation per student produced substantially different parameter estimates. Better estimates could not be had from less data; therefore, the estimates from the restricted data set must have been substantially worse than from the full data set. It is important to emphasize the extreme conditions under which the parameter estimation procedure failed. Complete data on the cohort would have contained: 47 raters x 29 students =

1363 observed ratings. In the reduced data set there were 107 observations. In other words, 7.85% of the possible data were present and 92.15% of the data were missing from the observed data table input to MERLIN. In the other two cohorts, the respective data tables were 17.74% and 17.04% complete.

With the clear evidence that it was the parameter estimation process rather than the proposed model that failed and that the failure was due to lack of sufficient data to make useful estimates of the proposed model's parameters, we reconsidered the results reported in Table 5.

Means 1 and 2 were computed using the weighted r to z mean correlation procedure recommended by McNemar (1966, p. 139). The mean correlation between Model A and an independent criterion (i.e., the saved ratings) across all three cohorts (mean 1) was higher than that obtained by Model B, but not significantly higher ($p>0.15$). However, ample evidence had been found which required the exclusion of the 1978-79:3 data from this comparison. Therefore, Mean 2 was calculated only upon the results for cohorts 1978-79:4 and 1979-80:2. This resulted in $r=0.62$ for the proposed model, while the mean correlation between the criterion and Model B predictions was $r=0.41$. Each of these correlations was significantly greater than $r=0.0$ ($p<0.004$). Further, the proposed model predicted the criterion significantly better ($z=2.62$; $p<0.004$) than did the alternative model. This result directly validates the theoretical constructs of both rater leniency and subject achievement.

Model A's predictions correlated higher with the independent criterion ratings, $r=0.61$, because Model A's predictions were more nearly valid. The raw ratings contained two components: subject achievement and rater leniency. As measures of true subject performance, the raw ratings were contaminated with rater leniency and were therefore less valid and reliable measures of true subject performance. The reliability of $r=0.50$ for raw ratings reported in earlier work (Cason and Cason, 1979) was an overestimate because it did not take the leniency effect into account. The best available estimate for the reliability of raw ratings as measures of performance alone was the mean correlation between Model B and the criterion ratings in the last two cohorts (mean 2): $r=0.41$. Our model attained higher correlations with the criterion because it explicitly used both rater leniency and achievement data to make its predictions. The model depicted the data more validly than could the mean of raw ratings in incomplete data sets. Therefore, the best available measure of student performance or student achievement was the rating that our model predicted a rater of average leniency would assign a given subject (or, its equivalent on the latent scale, this subject's SAP).

Applying our model, the reliability of a single rating as a measure of true performance was r=0.61. Leniency effects had been removed; therefore, Spearman-Brown's formula was appropriate to conservatively estimate the reliability of a rating based upon several independent raters. Specifically, our model's predicted mean rating for each subject based on 5 ratings had an estimated reliability of r=0.89. By the same logic, the reliability of the mean of 5 raw ratings as a measure of true performance was calculated taking r=0.41 as the reliability of a single raw rating. Applying Spearman-Brown's formula, this gave r=0.78 for the reliability of the mean of 5 observed ratings as a measure of student true performance. Because validity cannot exceed reliability these results clearly indicated our model could produce substantially more nearly valid measures of student true performance from an incomplete data table than could the mean of observed ratings on each student.

Conclusions and Implications

All the a priori objectives of the research were attained. With respect to clinical performance rating data sets of a type which are common to health professions education (i.e., dirty and incomplete), the proposed model was empirically demonstrated to have: (a) closely fit the data ($p<0.000001$), (b) clarified and quantified the separate contributions of rater leniency and subject achievement (e.g., 20% and 35% of variance accounted for respectively in these data; empirical cross-validation of both constructs, and so forth); and, (c) provided a usable mechanism for generating more reliable and valid ratings-based measures of clinical performance as indicated by the reliability of r=0.89 (based on 5 independent ratings) attained from application of the proposed model as compared to r=0.78 attained for the most commonly used current alternative, i.e., the mean of the 5 observed ratings.

The results clearly demonstrated the superiority of the proposed model when data sets were incomplete and subjects were rated by unrepresentative subsets of raters. In addition, an empirical method for judging the adequacy of the data for the application of the model was demonstrated. When the proposed model failed to provide fit with the data at least as good as the mean of each subject's observed ratings, the data set was insufficient to provide adequate estimates of the proposed model's parameters. Nevertheless, the proposed model provided improved measures of performance when the data set was as little as 17% complete.

The conditions of the tests contrasting the proposed model (A) with the mean of the observed ratings were biased against the proposed model. Assignment of students had been random so variation of average leniency in rater subsets would tend to be small. This tended to reduce the rater main effect in these data. In settings where non-random assignment occurs, larger discrepancies in mean rater leniency could easily occur. In such settings, the power of the proposed model in producing more valid measures would be even more pronounced. Assuming the proposed model was valid, Table 7 provides a "worst case" example of the potential impact of rater leniency upon ratings received by students. This example was based on the extreme (lenient and stringent) raters and extreme (low and high achieving) students in cohort 1979-80:2. The top row depicts the ratings that the most stringent rater would assign; the bottom row the most lenient. The left column gives the corresponding rater reference points for the two raters. The middle column gives the expected rating for the low achieving student; the rightmost column, the expected ratings for the high achieving student. Both the raters see the high achieving student much the same; there is only a 10% difference in ratings. But, the low achieving student is predicted to receive drastically different ratings. There is a 30% difference in ratings. Predictions rather than observed discrepancies were used in the illustration because it was the model that was validated in this research. Whether discrepancies as large as this occurred in this data was a chance matter. The model's predictions were a better general indicator of the possible magnitude than coincidental data because the model captured a set of relationships in whole data sets.

Table 7

Maximum Effect of Rater Leniency on Predicted Student Ratings in Cohort 79-80:2

	RRP	Low Student (SAP 497.9)	High Student (SAP 653.7)
Stringent rater	(534.9)	35.59%	88.27%
Lenient rater	(452.1)	67.65%	97.81%

In spite of the consistency, strength, and coherence of the results supporting the proposed model found in these data, these data were limited. Only one setting, an internal medicine clerkship was represented. Only one rating inventory was used. Still 75 different raters were involved and 94 different students were rated. It would not be prudent to conclude that the proposed model will fit every conceivable performance rating setting. Neither would it be reasonable to ignore the strength of the results from these limited data. There are too many commonalities between these data and many others not to expect that this model may prove very useful in a wide variety of settings and contexts.

Extrapolating optimistically from these early, promising results, a number of useful possibilities occur to us. Our model might meet Meskauskas and Norcini's (1980) requirements for a methodology for "handicapping" judges in both standards setting and performance assessment procedures better than do Stanley's (1961) methods. Our results suggest that in some settings rater leniency may not be sufficiently stable to use Stanley's methods. However, because our model can be applied to incomplete data sets, it provides a means of "adjusting" judges' ratings on the bases of their current behavior rather than on their past ratings.

An intriguing possibility is the application of our model to the problem of assessing the test items in a large item bank. Some test item banks now have thousands of items in them. But, these items are not equally relevant to the objectives of specific training programs which may use these test item banks. Our model would permit a more uniform standard to be applied in judging the difficulty or relevance of items in the item bank while reducing the extent to which redundant judgements were required. For example, our model might permit judges to consider only slightly overlapping subsets of items while applying Angoff's (1971) or a similar standards setting method. The judges' judgements could be calibrated through the common items that they judged. This would permit a small number of judges (e.g., the faculty in a department) to evaluate a larger item bank without either taking years or imposing an unrealistic burden on the individuals.

Our model provides a technique whereby it would be possible to "track" the rating performance of individual raters and provide them with feedback on how their ratings compared with other raters in settings where not all raters rate all subjects. This might even be useful in settings where raters had been trained to a very high level of skill so that only few raters would rate each subject. So long as there were adequate overlaps in the ratings, the model would provide a way of monitoring raters that was non-intrusive and inexpensive since it requires only their routine rating data.

There are at least two general ways in which our model may prove to be of research interest. First, the model itself, in so far as it is a simplification of a somewhat more elaborate theory, deserves investigation. Perhaps incorporation of differential rater sensitivity, an explicit representation of problem or situation difficulty, or other elaborations of the proposed model would lead to further improvements in ratings-based measures of complex human performance. However, such elaborations would involve adding parameters and this would require more nearly complete data sets if useful estimates of the parameters were to be achievable. In spite of the success of the simplified model in fitting and explaining the relationships in these data, the model is a gross simplification of even the rudimentary performance rating theory that we have proposed.

Second, the proposed model may be useful as an analytic method in research involving complex human performance as either a criterion or predictor variable. With notable exeptions such as Sheehan, Husted, Candee, Cook, and Bargen's (1980) report, prior investigations of the relationships between complex performance variables (such as clinical performance) and variables measured by more reliable methods (such as objectively scored aptitude and achievement tests) have found only very modest relationships or none at all. This may have arisen in part because of the relatively low reliability and/or validity of the available ratings-based measures of complex performance. The proposed model may have a substantial contribution to make to these investigations by providing a way to get more nearly valid and highly reliable measures of complex performance than have been available in the past. This prospect is especially exciting for those areas of performance where there are already large but dirty and incomplete data sets available and/or those areas which, for practical reasons, may be unable to concurrently produce both clean and complete data sets regardless of the resources available.

While it is desirable that the judgements of individual judges be made as reliable and valid as is possible, there will almost certainly always be more assessment programs that generate incomplete, dirty data sets than complete, clean ones. The model we have presented here shows real promise for improving the quality of the assessment information that may be extracted under these less than ideal and unfortunately common circumstances.

References

Anderson, D.O., Baker, H.H., Laguna, J.E., and Laguna, J.F. Applying the Rasch model to improve health science clerkship evaluations. Presented at the Annual Meeting of the Rocky Mountain Educatioral Research Association, Las Cruces, N. M., 1980.

Angoff, W.H. Scales, norms and equivalent scores. In R.L. Thorndike (Ed.) Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Baker, F.B. Advances in item analysis. Review of Educational Research, 1977, 47, 151-178.

Cason, G.J. MERLIN: A FORTRAN IV program for finding least-squares estimates of rater reference points, subject achievement points, and goodness-of-fit for Cason and Cason's model of performance rating. Copyright 1980 by Gerald J. Cason. (Available from author.)

Cason, G.J., and Cason, C.L. Rating students' clinical performance: Interim report number 2. Presented at the Annual Meeting of the Mid-South Educational Research Association, Little Rock, Arkansas, 1979.

Chandler, J.P. STEPIT: A FORTRAN II subroutine for finding local minima of real functions. Copyright by J.P. Chandler. (Available from Quantum Chemistry Program Exchange, Indiana University: Bloomington, Indiana.)

Cromier, G. A study of the applicability of a truly objective model in medical education. In Proceedings of the Sixteenth Annual Conference on Research in Medical Education. Washington, D.C.: American Association of Medical Colleges, 1977, 123-128.

Davidge, A.M., Davis, W.K., and Hull, A.L. A system for the evaluation of medical students' clinical competence. Journal of Medical Education, 1980, 55, 65-67.

Davis, W.K., Hull, A.L., Davidge, A.M., and Dielman, T.E. Variables influencing ratings of medical student's clinical performance. Presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.

Dielman, T.E., Hull, A.L., and Davis, W.K. Psychometric properties of clinical performance rating. Evaluation and the Health Professions, 1980, 3(1), 103-117.

Ebel, R.L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R., and Gifford, J.A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Harasym, P. A comparison of the Nedelsky and modified Angoff standard-setting procedure on evaluation outcome. In Proceedings of the Nineteenth Annual Conference on Research in Medical Education. Washington, D.C.: American Association of Medical Colleges, 1980, 3-8.

Hughes, F.P. The Rasch model applied to the equating of several examination forms. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.

Kreines, D.C., and Mead, R.J. Equating tests with the Rasch model. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.

Landy, F., and Barnes, J. Scaling behavioral anchors. Applied Psychological Measurement, 1978, 3(2), 193-200.

Lord, F.M. A theory of test scores. Psychometric Monographs, 1952, No.7.

Lord, F.M. An application of confidence intervals and maximum likelihood to the estimation of an examinee's ability. Psychometrika, 1953, 18, 57-75.

McNemar, G. Psychological statistics (3rd Ed.). New York: Wiley, 1966.

Mead, R.J., Wright, B.D., and Bell, S.R. BICAL-Version 3. Computer program to perform Rasch item analysis. Chicago: University of Chicago, 1979.

Meskauskas, J.A., and Norcini, J.J. Standard-setting in written and interactive (oral) specialty certification examinations: Issues, models, methods, challenges. Evaluation and the Health Professions, 1980, 3(3), 321-360.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Nunnally, J.C. Psychometric theory. New York: McGraw-Hill, 1967.

O'Donohue, W.J., and Wergin, J.F. Evaluation of medical students during a clinical clerkship in internal medicine. Journal of Medical Education, 1978, 53, 55-58.

Pierleoni, R.G., Clark, G.M., and Dudding, B.A. A comparison of faculty, resident, and nurse practitioner ratings of ambulatory pediatric students. Presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.

Printen, K.J., Chappell, W., and Whitney, D.R. Clinical performance evaluation of junior medical students. Journal of Medical Education, 1973, 48, 343-348.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Remmers, H.H., Shock, N.W., and Kelly, E.L. An empirical study of the vaidity of the Spearman-Brown formula as applied to the Purdue Rating Scale. Journal of Educational Psychology, 1927, 18, 187-195.

Schumaker, C.F., et al. Applying the Rasch model to equate examinations in the field of medicine. Presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.

Sheehan, J.T., Husted, S.D.R., Candee, D., Cook, C.D., and Bargen, M. Moral judgement as a predictor of clinical performance. Evaluation and the Health Professions, 1980, 3(4), 393-404.

Smith, H.A., and Kifer, E. Student evaluation in an externship utilizing the Rasch model for test calibration. American Journal of Pharmaceutical Education, 1980, 44, 6-11.

Smith, P., and Kendall, L. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

Snedecor, G.W. Statistical methods. (4th Ed.). Ames, Iowa: Iowa State College Press, 1946.

Stanley, J.C. Analysis of unreplicated three-way classifications with applications to rater bias and trait independence. Psychometrika, 1961, 26(2), 203-209.

Sternberg, S. Stochastic learning theory. In R.D. Luce, R.R. Bush and E. Galanter (Eds.) Handbook of Mathematical Psychology, Volume II. New York: Wiley, 1967.

Stillman, P.L. Arizona Clinical Interview Medical Rating Scale. Medical Teacher, 1980, 2(5), 248-251.

Stillman, P.L., Brown, D.R., Redfield, D.L., and Sabers, D.L. Construct validation of the Arizona Clinical Interview Rating Scale. Educational and Psychological Measurement, 1977, 37, 1031-1038.

Symonds, P.M. Diagnosing personality and conduct. New York: Century, 1931.

Ward, J., and Jennings, E. Introduction to linear models. Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Wright, B.D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968.

Wright, B.D., and Stone, M.H. Best test design. Chicago: MESA Press, 1979.

Acknowledgements

We gratefully acknowledge the encouragement, assistance, and co-operation of: George Ackerman, Harry Ackerman, Jerry Blackburn, Roger Bone, Tom Bruce, John Delk, Ross Dykman, Ron Hale, Lisa Hale, Peter Kohler, Tom Lewis, Tom Monson, Jim Phillips, Bob Shannon, Lois Tipton, and Ture Schoultz.

This research was supported in part by the taxpayers of the United States through grant No. 90AL0005-01 from the Department of Health and Human Services.

Director: MD

University of Arkansas for Medical Sciences

STUDENT'S LAST NAME

DATE

STUDENT NUMBER

12 MCH/VAH — Instructor

XXXXJr Yr

XXXXX

USE SCALE AT RIGHT
USE SOFT PENCIL
MAKE NO MORE THAN
ONE MARK PER ITEM
ERASE COMPLETELY
TO CORRECT.

5 = Substantially better (i.e., would be at top 10% of typical class)
4 = A little better (i.e., would be in top 25% but below top 10% of typical class)
3 = No better or worse (i.e., would be in middle 50% of typical class)
2 = A little worse (i.e., would be in bottom 25% but above bottom 10% of typical class)
1 = Substantially worse (i.e., would be in bottom 10% of typical class)

Than Typic... in Typical...

XXXXXXXXXXXXXXXXXXX X = Not Applicable Y = Not Rated Because Not Observed

GENERAL COGNITIVE SKILLS:				
Knowing facts, rules, etc	1 5 4 3 2 1	X Y	Recording Physical Exam	21 5
Understanding facts, rules	2 5 4 3 2 1	X Y	Requesting Studies/Tests	22 5
Applying facts, rules, etc	3 5 4 3 2 1	X Y	Requesting Consults	23 5
Problem Solving: analysis, synthesis evaluation	4 5 4 3 2 1	X Y	Interpreting: History Results	24 5
COMMUNICATION (with):				
Peers (Jr Med Students)	5 5 4 3 2 1	X Y	Physical Exam	25 5
Patients	6 5 4 3 2 1	X Y	Studies/Tests	26 5
Faculty	7 5 4 3 2 1	X Y	Consult Results	27 5
Residents	8 5 4 3 2 1	X Y	Synthesizing Problem/ Formulating Diagnosis	28 5
Clinical Team: RNs, Techs, etc	9 5 4 3 2 1	X Y	*THERAPEUTIC DESIGN/PROCEDURES:* Selecting/formulating treatment	29 5
ATTITUDE (toward):				
Peers (Jr Med Students)	10 5 4 3 2 1	X Y	Manual Skills & Executing procedures........	30 5
Patients	11 5 4 3 2 1	X Y	Follow-up, evaluation, revision of treatment regimen........	31 5
Faculty	12 5 4 3 2 1	X Y	*PERFORMANCE UNDER STRESS*	32 5
Residents	13 5 4 3 2 1	X Y	*POTENTIAL FOR ADVANCED TRAINING*	33 5
Clinical Team: RNs, Techs, etc	14 5 4 3 2 1	X Y	*PROVISIONAL OVERALL GRADE* A=5 / B=4 / C=3 / D=2 / F=1	34 5
Assigned duties	15 5 4 3 2 1	X Y		35 5
Implicit responsibilities	16 5 4 3 2 1	X Y		36 5
Being corrected	17 5 4 3 2 1	X Y		37 5
BASIC PATIENT "WORK-UP"				
Conducting History	18 5 4 3 2 1	X Y		38 5 4 3 2 1 X Y
Conducting Physical Exam	19 5 4 3 2 1	X Y		39 5 4 3 2 1 X Y
Recording History	20 5 4 3 2 1	X Y		40 5 4 3 2 1 X Y

UAMS STUDENT PERFORMANCE RATING FORM

RATER SIGN AND COMMENT ON REVERSE SIDE

RATER'S COMMENTS

PLEASE ENTER ANY COMMENTS YOU FEEL ARE RELEVANT TO THE EVALUATION OF STUDENT WHOSE RATINGS YOU ENTERED ON OBVERSE OF THIS FORM. COMMENTS ON THIS STUDENT'S SPECIFIC STRENGTHS AND/OR WEAKNESSES AND DOCUMENTATION FOR RATINGS ASSIGNED ARE MOST USEFUL.
RATINGS FOR EACH ITEM ON OBVERSE MUST BE ENTERED. COMMENTS ARE ONLY A SUPPLEMENT NOT A SUBSTITUTE FOR RATINGS.

PROVISIONAL OVERALL GRADE: In marking item 34 on the obverse, use the definitions for 5, 4, 3, 2, and 1 given below. For all other items, use the definitions provided above the items on the obverse of this form.

5 = A = OUTSTANDING overall performance for Med. School Jr.
4 = B = ABOVE AVERAGE " " for Med. School Jr.
3 = C = AVERAGE " " for Med. School Jr.
2 = D = BELOW AVERAGE " " for Med. School Jr.
1 = F = UNSATISFACTORY " " for Med. School Jr.

IF NO COMMENTS, CHECK HERE □ X ____________________ RATER'S SIGNATURE

____________________ DATE ____________________ RATER'S NAME TYPED OR PRINTED

Appendix A

to free parameters (to be estimated for Model A) was 1.8, 2.7, and 2.7 respectively in the three cohorts. In the restricted data set, these ratios declined to 1.4, 2.2, and 2.2. In cohort 1978-79:3 the ratio fell from an already marginal 1.8 observations per parameter in the full data set to a very doubtful 1.4 in the restricted data set. A low ratio could place the proposed model at a disadvantage because it had more parameters to be estimated. Less data per paramter would reduce the accuracy of the parameter estimates and thus the accuracy of the model's predictions. The alternate model having only about half as many parameters to estimate had an advantage in obtaining more accurate estimates of its parameters (i.e., one mean per student).

Table 5 reports the results of correlating an independent rating of each student with the prediction of the proposed model (A) and the prediction implicit in the common practice of taking the unweighted mean of the observed ratings (Model B) as the best available measure of performance. In two of the three cohorts the results appear to favor the proposed model, but in cohort 1978-79:3, Model B seems to be superior to the proposed model. This means that in two of the three cohorts predictions based upon a knowledge of both rater leniency and student performance (i.e., Model A) appeared superior to a knowledge of student performance alone (Model B). In cohort 1978-79:3, the prediction of Model A was not only less accurate (i.e., less well correlated with the criterion), the observed correlation (r=0.26) for Model A was not significantly different from r=0.0. Considering that Model B was a restricted case of Model A, Model A should do no worse than Model B.

Table 5

Correlations of Prediction of Models A and B with an Independent Rating on Each Subject

Cohort	A	B
78-79:3	0.2555	0.5020
78-79:4	0.6699	0.5531
79-80:2	0.4027	0.2022
Mean 1	0.5136	0.4465
Mean 2	0.6128	0.4055

For cohort 1978-79:3, the data indicate that very poor estimates for Model A's parameters were obtained from the restricted data set. The result of Model A fitting worse than Model B was directly attributable to the lack of sufficient data in the restricted data set for simultaneously estimating SAP's and RRP's. This "negative finding" was serendipitously suggestive of a useful rule of thumb. Anytime the correlation between the proposed model's predictions (when based upon the parameter estimation procedures in MERLIN) and independent criterion ratings fails to at least equal the correlation between the criterion and each subject's mean observed rating (i.e., the prediction of Model B), then there are insufficient data available to make useful estimates of the parameters of the Model A. In the case at issue, this interpretation was corroborated by an analysis of correlations between the values estimated for Model A's parameters (RRP's and SAP's) based on the full data set with estimates for the same parameters based on the restricted data set. The results of these analyses are reported in Table 6.

Table 6

Correlations between Parameters Estimated from the Full Data Set with Those Estimated from the Restricted Data Set

Cohort	RRP	SAP	Both	RSQ
78-79:3	0.8300	0.7991	0.8178	0.6688
78-79:4	0.9173	0.9691	0.9508	0.9040
79-80:2	0.8676	0.9625	0.9329	0.8703

These correlations would be high if the parameter estimates were stable. The correlations for RRP's and SAP's separately and combined indicated that there was good stability for the parameter estimates in cohorts 1978-79:4 and 1979-80:2. Taking the square of the correlation (RSQ) between the two conditions (i.e., full and restricted data sets) as a measure of common variance, the stability of cohort 1978-79:3's parameter estimates was clearly poor (RSQ=0.67). Deleting one observation per student produced substantially different parameter estimates. Better estimates could not be had from less data; therefore, the estimates from the restricted data set must have been substantially worse than from the full data set. It is important to emphasize the extreme conditions under which the parameter estimation procedure failed. Complete data on the cohort would have contained: 47 raters x 29 students =

1363 observed ratings. In the reduced data set there were 107 observations. In other words, 7.85% of the possible data were present and 92.15% of the data were missing from the observed data table input to MERLIN. In the other two cohorts, the respective data tables were 17.74% and 17.04% complete.

With the clear evidence that it was the parameter estimation process rather than the proposed model that failed and that the failure was due to lack of sufficient data to make useful estimates of the proposed model's parameters, we reconsidered the results reported in Table 5.

Means 1 and 2 were computed using the weighted r to z mean correlation procedure recommended by McNemar (1966, p. 139). The mean correlation between Model A and an independent criterion (i.e., the saved ratings) across all three cohorts (mean 1) was higher than that obtained by Model B, but not significantly higher ($p>0.15$). However, ample evidence had been found which required the exclusion of the 1978-79:3 data from this comparison. Therefore, Mean 2 was calculated only upon the results for cohorts 1978-79:4 and 1979-80:2. This resulted in $r=0.62$ for the proposed model, while the mean correlation between the criterion and Model B predictions was $r=0.41$. Each of these correlations was significantly greater than $r=0.0$ ($p<0.004$). Further, the proposed model predicted the criterion significantly better ($z=2.62$; $p<0.004$) than did the alternative model. This result directly validates the theoretical constructs of both rater leniency and subject achievement.

Model A's predictions correlated higher with the independent criterion ratings, $r=0.61$, because Model A's predictions were more nearly valid. The raw ratings contained two components: subject achievement and rater leniency. As measures of true subject performance, the raw ratings were contaminated with rater leniency and were therefore less valid and reliable measures of true subject performance. The reliability of $r=0.50$ for raw ratings reported in earlier work (Cason and Cason, 1979) was an overestimate because it did not take the leniency effect into account. The best available estimate for the reliability of raw ratings as measures of performance alone was the mean correlation between Model B and the criterion ratings in the last two cohorts (mean 2): $r=0.41$. Our model attained higher correlations with the criterion because it explicitly used both rater leniency and achievement data to make its predictions. The model depicted the data more validly than could the mean of raw ratings in incomplete data sets. Therefore, the best available measure of student performance or student achievement was the rating that our model predicted a rater of average leniency would assign a given subject (or, its equivalent on the latent scale, this subject's SAP).

Applying our model, the reliability of a single rating as a measure of true performance was r=0.61. Leniency effects had been removed; therefore, Spearman-Brown's formula was appropriate to conservatively estimate the reliability of a rating based upon several independent raters. Specifically, our model's predicted mean rating for each subject based on 5 ratings had an estimated reliability of r=0.89. By the same logic, the reliability of the mean of 5 raw ratings as a measure of true performance was calculated taking r=0.41 as the reliability of a single raw rating. Applying Spearman-Brown's formula, this gave r=0.78 for the reliability of the mean of 5 observed ratings as a measure of student true performance. Because validity cannot exceed reliability these results clearly indicated our model could produce substantially more nearly valid measures of student true performance from an incomplete data table than could the mean of observed ratings on each student.

Conclusions and Implications

All the a priori objectives of the research were attained. With respect to clinical performance rating data sets of a type which are common to health professions education (i.e., dirty and incomplete), the proposed model was empirically demonstrated to have: (a) closely fit the data ($p<0.000001$), (b) clarified and quantified the separate contributions of rater leniency and subject achievement (e.g., 20% and 35% of variance accounted for respectively in these data; empirical cross-validation of both constructs, and so forth); and, (c) provided a usable mechanism for generating more reliable and valid ratings-based measures of clinical performance as indicated by the reliability of r=0.89 (based on 5 independent ratings) attained from application of the proposed model as compared to r=0.78 attained for the most commonly used current alternative, i.e., the mean of the 5 observed ratings.

The results clearly demonstrated the superiority of the proposed model when data sets were incomplete and subjects were rated by unrepresentative subsets of raters. In addition, an empirical method for judging the adequacy of the data for the application of the model was demonstrated. When the proposed model failed to provide fit with the data at least as good as the mean of each subject's observed ratings, the data set was insufficient to provide adequate estimates of the proposed model's parameters. Nevertheless, the proposed model provided improved measures of performance when the data set was as little as 17% complete.

The conditions of the tests contrasting the proposed model (A) with the mean of the observed ratings were biased against the proposed model. Assignment of students had been random so variation of average leniency in rater subsets would tend to be small. This tended to reduce the rater main effect in these data. In settings where non-random assignment occurs, larger discrepancies in mean rater leniency could easily occur. In such settings, the power of the proposed model in producing more valid measures would be even more pronounced. Assuming the proposed model was valid, Table 7 provides a "worst case" example of the potential impact of rater leniency upon ratings received by students. This example was based on the extreme (lenient and stringent) raters and extreme (low and high achieving) students in cohort 1979-80:2. The top row depicts the ratings that the most stringent rater would assign; the bottom row the most lenient. The left column gives the corresponding rater reference points for the two raters. The middle column gives the expected rating for the low achieving student; the rightmost column, the expected ratings for the high achieving student. Both the raters see the high achieving student much the same; there is only a 10% difference in ratings. But, the low achieving student is predicted to receive drastically different ratings. There is a 30% difference in ratings. Predictions rather than observed discrepancies were used in the illustration because it was the model that was validated in this research. Whether discrepancies as large as this occurred in this data was a chance matter. The model's predictions were a better general indicator of the possible magnitude than coincidental data because the model captured a set of relationships in whole data sets.

Table 7

Maximum Effect of Rater Leniency on Predicted Student Ratings in Cohort 79-80:2

	RRP	Low Student (SAP 497.9)	High Student (SAP 653.7)
Stringent rater	(534.9)	35.59%	88.27%
Lenient rater	(452.1)	67.65%	97.81%

In spite of the consistency, strength, and coherence of the results supporting the proposed model found in these data, these data were limited. Only one setting, an internal medicine clerkship was represented. Only one rating inventory was used. Still 75 different raters were involved and 94 different students were rated. It would not be prudent to conclude that the proposed model will fit every conceivable performance rating setting. Neither would it be reasonable to ignore the strength of the results from these limited data. There are too many commonalities between these data and many others not to expect that this model may prove very useful in a wide variety of settings and contexts.

Extrapolating optimistically from these early, promising results, a number of useful possibilities occur to us. Our model might meet Meskauskas and Norcini's (1980) requirements for a methodology for "handicapping" judges in both standards setting and performance assessment procedures better than do Stanley's (1961) methods. Our results suggest that in some settings rater leniency may not be sufficiently stable to use Stanley's methods. However, because our model can be applied to incomplete data sets, it provides a means of "adjusting" judges' ratings on the bases of their current behavior rather than on their past ratings.

An intriguing possibility is the application of our model to the problem of assessing the test items in a large item bank. Some test item banks now have thousands of items in them. But, these items are not equally relevant to the objectives of specific training programs which may use these test item banks. Our model would permit a more uniform standard to be applied in judging the difficulty or relevance of items in the item bank while reducing the extent to which redundant judgements were required. For example, our model might permit judges to consider only slightly overlapping subsets of items while applying Angoff's (1971) or a similar standards setting method. The judges' judgements could be calibrated through the common items that they judged. This would permit a small number of judges (e.g., the faculty in a department) to evaluate a larger item bank without either taking years or imposing an unrealistic burden on the individuals.

Our model provides a technique whereby it would be possible to "track" the rating performance of individual raters and provide them with feedback on how their ratings compared with other raters in settings where not all raters rate all subjects. This might even be useful in settings where raters had been trained to a very high level of skill so that only few raters would rate each subject. So long as there were adequate overlaps in the ratings, the model would provide a way of monitoring raters that was non-intrusive and inexpensive since it requires only their routine rating data.

There are at least two general ways in which our model may prove to be of research interest. First, the model itself, in so far as it is a simplification of a somewhat more elaborate theory, deserves investigation. Perhaps incorporation of differential rater sensitivity, an explicit representation of problem or situation difficulty, or other elaborations of the proposed model would lead to further improvements in ratings-based measures of complex human performance. However, such elaborations would involve adding parameters and this would require more nearly complete data sets if useful estimates of the parameters were to be achievable. In spite of the success of the simplified model in fitting and explaining the relationships in these data, the model is a gross simplification of even the rudimentary performance rating theory that we have proposed.

Second, the proposed model may be useful as an analytic method in research involving complex human performance as either a criterion or predictor variable. With notable exeptions such as Sheehan, Husted, Candee, Cook, and Bargen's (1980) report, prior investigations of the relationships between complex performance variables (such as clinical performance) and variables measured by more reliable methods (such as objectively scored aptitude and achievement tests) have found only very modest relationships or none at all. This may have arisen in part because of the relatively low reliability and/or validity of the available ratings-based measures of complex performance. The proposed model may have a substantial contribution to make to these investigations by providing a way to get more nearly valid and highly reliable measures of complex performance than have been available in the past. This prospect is especially exciting for those areas of performance where there are already large but dirty and incomplete data sets available and/or those areas which, for practical reasons, may be unable to concurrently produce both clean and complete data sets regardless of the resources available.

While it is desirable that the judgements of individual judges be made as reliable and valid as is possible, there will almost certainly always be more assessment programs that generate incomplete, dirty data sets than complete, clean ones. The model we have presented here shows real promise for improving the quality of the assessment information that may be extracted under these less than ideal and unfortunately common circumstances.

References

Anderson, D.O., Baker, H.H., Laguna, J.E., and Laguna, J.F. Applying the Rasch model to improve health science clerkship evaluations. Presented at the Annual Meeting of the Rocky Mountain Educatioral Research Association, Las Cruces, N. M., 1980.

Angoff, W.H. Scales, norms and equivalent scores. In R.L. Thorndike (Ed.) Ecucational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Baker, F.B. Advances in item analysis. Review of Educational Research, 1977, 47, 151-178.

Cason, G.J. MERLIN: A FORTRAN IV program for finding least-squares estimates of rater reference points, subject achievement points, and goodness-of-fit for Cason and Cason's model of performance rating. Copyright 1980 by Gerald J. Cason. (Available from author.)

Cason, G.J., and Cason, C.L. Rating students' clinical performance: Interim report number 2. Presented at the Annual Meeting of the Mid-South Educational Research Association, Little Rock, Arkansas, 1979.

Chandler, J.P. STEPIT: A FORTRAN II subroutine for finding local minima of real functions. Copyright by J.P. Chandler. (Available from Quantum Chemistry Program Exchange, Indiana University: Bloomington, Indiana.)

Cromier, G. A study of the applicability of a truly objective model in medical education. In Proceedings of the Sixteenth Annual Conference on Research in Medical Education. Washington, D.C.: American Association of Medical Colleges, 1977, 123-128.

Davidge, A.M., Davis, W.K., and Hull, A.L. A system for the evaluation of medical students' clinical competence. Journal of Medical Education, 1980, 55, 65-67.

Davis, W.K., Hull, A.L., Davidge, A.M., and Dielman, T.E. Variables influencing ratings of medical student's clinical performance. Presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.

Dielman, T.E., Hull, A.L., and Davis, W.K. Psychometric properties of clinical performance rating. Evaluation and the Health Professions, 1980, 3(1), 103-117.

Ebel, R.L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R., and Gifford, J.A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Harasym, P. A comparison of the Nedelsky and modified Angoff standard-setting procedure on evaluation outcome. In Proceedings of the Nineteenth Annual Conference on Research in Medical Education. Washington, D.C.: American Association of Medical Colleges, 1980, 3-8.

Hughes, F.P. The Rasch model applied to the equating of several examination forms. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.

Kreines, D.C., and Mead, R.J. Equating tests with the Rasch model. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.

Landy, F., and Barnes, J. Scaling behavioral anchors. Applied Psychological Measurement, 1978, 3(2), 193-200.

Lord, F.M. A theory of test scores. Psychometric Monographs, 1952, No.7.

Lord, F.M. An application of confidence intervals and maximum likelihood to the estimation of an examinee's ability. Psychometrika, 1953, 18, 57-75.

McNemar, G. Psychological statistics (3rd Ed.). New York: Wiley, 1966.

Mead, R.J., Wright, B.D., and Bell, S.R. BICAL-Version 3. Computer program to perform Rasch item analysis. Chicago: University of Chicago, 1979.

Meskauskas, J.A., and Norcini, J.J. Standard-setting in written and interactive (oral) specialty certification examinations: Issues, models, methods, challenges. Evaluation and the Health Professions, 1980, 3(3), 321-360.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Nunnally, J.C. Psychometric theory. New York: McGraw-Hill, 1967.

O'Donohue, W.J., and Wergin, J.F. Evaluation of medical students during a clinical clerkship in internal medicine. Journal of Medical Education, 1978, 53, 55-58.

Pierleoni, R.G., Clark, G.M., and Dudding, B.A. A comparison of faculty, resident, and nurse practitioner ratings of ambulatory pediatric students. Presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.

Printen, K.J., Chappell, W., and Whitney, D.R. Clinical performance evaluation of junior medical students. Journal of Medical Education, 1973, 48, 343-348.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Remmers, H.H., Shock, N.W., and Kelly, E.L. An empirical study of the vaidity of the Spearman-Brown formula as applied to the Purdue Rating Scale. Journal of Educational Psychology, 1927, 18, 187-195.

Schumaker, C.F., et al. Applying the Rasch model to equate examinations in the field of medicine. Presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.

Sheehan, J.T., Husted, S.D.R., Candee, D., Cook, C.D., and Bargen, M. Moral judgement as a predictor of clinical performance. Evaluation and the Health Professions, 1980, 3(4), 393-404.

Smith, H.A., and Kifer, E. Student evaluation in an externship utilizing the Rasch model for test calibration. American Journal of Pharmaceutical Education, 1980, 44, 6-11.

Smith, P., and Kendall, L. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

Snedecor, G.W. Statistical methods. (4th Ed.). Ames, Iowa: Iowa State College Press, 1946.

Stanley, J.C. Analysis of unreplicated three-way classifications with applications to rater bias and trait independence. Psychometrika, 1961, 26(2), 203-219.

Sternberg, S. Stochastic learning theory. In R.D. Luce, R.R. Bush and E. Galanter (Eds.) Handbook of Mathematical Psychology, Volume II. New York: Wiley, 1967.

Stillman, P.L. Arizona Clinical Interview Medical Rating Scale. Medical Teacher, 1980, 2(5), 248-251.

Stillman, P.L., Brown, D.R., Redfield, D.L., and Sabers, D.L. Construct validation of the Arizona Clinical Interview Rating Scale. Educational and Psychological Measurement, 1977, 37, 1031-1038.

Symonds, P.M. Diagnosing personality and conduct. New York: Century, 1931.

Ward, J., and Jennings, E. Introduction to linear models. Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Wright, B.D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968.

Wright, B.D., and Stone, M.H. Best test design. Chicago: MESA Press, 1979.

Acknowledgements

We gratefully acknowledge the encouragement, assistance, and co-operation of: George Ackerman, Harry Ackerman, Jerry Blackburn, Roger Bone, Tom Bruce, John Delk, Ross Dykman, Ron Hale, Lisa Hale, Peter Kohler, Tom Lewis, Tom Monson, Jim Phillips, Bob Shannon, Lois Tipton, and Ture Schoultz.

This research was supported in part by the taxpayers of the United States through grant No. 90AL0005-01 from the Department of Health and Human Services.

Director: MD

University of Arkansas for Medical Sciences

USE SCALE AT RIGHT
USE SOFT PENCIL
MAKE NO MORE THAN
ONE MARK PER ITEM
ERASE COMPLETELY
TO CORRECT.

5 = Substantially better (i.e., would be in top 10% of typical class)
4 = A little better (i.e., would be in top 25% but below top 10% of typical class)
3 = No better or worse (i.e., would be in middle 50% of typical class)
2 = A little worse (i.e., would be in bottom 25% but above bottom 10% of typical class)
1 = Substantially worse (i.e., would be in bottom 10% of typical class)

} Than Typic
in Typical

XXXXXXXXXXXXXXXXXXXX

X = Not Applicable Y = Not Rated Because Not Observed

XXXX Jr Yr 12 MCH/VAH INSTRUCTOR

GENERAL COGNITIVE SKILLS:					
Knowing facts, rules, etc	1 5 4 3 2 1	X Y	Recording Physical Exam	21	5
Understanding facts, rules	2 5 4 3 2 1	X Y	Requesting Studies/Tests	22	5
Applying facts, rules, etc	3 5 4 3 2 1	X Y	Requesting Consults	23	5
Problem Solving: analysis, synthesis evaluation	4 5 4 3 2 1	X Y	Interpreting: History Results	24	5
COMMUNICATION (with):					
Peers (Jr Med Students)	5 5 4 3 2 1	X Y	Physical Exam	25	5
Patients	6 5 4 3 2 1	X Y	Studies/Tests	26	5
Faculty	7 5 4 3 2 1	X Y	Consult Results	27	5
Residents	8 5 4 3 2 1	X Y	Synthesizing Problem/ Formulating Diagnosis	28	5
Clinical Team: RNs,Techs,etc	9 5 4 3 2 1	X Y	*THERAPEUTIC DESIGN/PROCEDURES:* Selecting/formulating treatment	29	5
ATTITUDE (toward):					
Peers (Jr Med Students)	10 5 4 3 2 1	X Y	Manual Skills & Executing procedures	30	5
Patients	11 5 4 3 2 1	X Y	Follow-up, evaluation, revision of treatment regimen.........	31	5
Faculty	12 5 4 3 2 1	X Y	*PERFORMANCE UNDER STRESS*	32	5
Residents	13 5 4 3 2 1	X Y	*POTENTIAL FOR ADVANCED TRAINING*	33	5
Clinical Team: RNs,Techs,etc	14 5 4 3 2 1	X Y	*PROVISIONAL OVERALL GRADE* A=5 / B=4 / C=3 / D=2 / F=1	34	5
Assigned duties	15 5 4 3 2 1	X Y		35	5
Implicit responsibilities	16 5 4 3 2 1	X Y		36	5
Being corrected	17 5 4 3 2 1	X Y		37	5
BASIC PATIENT "WORK-UP"					
Conducting History	18 5 4 3 2 1	X Y		38	5 4 3 2 1 X Y
Conducting Physical Exam	19 5 4 3 2 1	X Y		39	5 4 3 2 1 X Y
Recording History	20 5 4 3 2 1	X Y		40	5 4 3 2 1 X Y

UAMS STUDENT PERFORMANCE RATING FORM

RATER SIGN AND COMMENT ON REVERSE SIDE

RATER'S COMMENTS

PLEASE ENTER ANY COMMENTS YOU FEEL ARE RELEVAN
STUDENT WHOSE RATINGS YOU ENTERED ON OBVERSE O
ON THIS STUDENT'S SPECIFIC STRENGTHS AND/OR
UMENTATION FOR RATINGS ASSIGNED ARE MOST USE
RATINGS FOR EACH ITEM ON OBVERSE MUST BE ENTER
A SUPPLEMENT NOT A SUBSTITUTE FOR RATINGS.

PROVISIONAL OVERALL GRADE: In marking item
the definitions for 5, 4, 3, 2, and 1 given
items, use the definitions provided above the
of this form.

5 = A = OUTSTANDING overall performanc
4 = B = ABOVE AVERAGE " "
3 = C = AVERAGE " "
2 = D = BELOW AVERAGE " "
1 = F = UNSATISFACTORY " "

IF NO COMMENTS, CHECK HERE □ X
RATER'S SIGNAT

DATE RATER'S NAME