Statistica Sinica **8**(1998), 713-728

SOME STATISTICAL AND LOGICAL CONSIDERATIONS WHEN RESCORING TESTS

Eric T. Bradlow and Howard Wainer

University of Pennsylvania and Educational Testing Service

Abstract: When tests or portions of tests are scored subjectively by raters, a rescoring will yield a change in the ratings of some examinees. In a test with a fixed passing score a rescoring will result in the change of some pass/fail decisions. The number of changes depends on: the reliability of the rating system, the number of raters, the variability in examinee abilities, the proportion of examinees that initially pass, and the policy used to incorporate the rescore into the pass/fail decision. In this study, we provide a model that facilitates the evaluation of various rescoring strategies. We consider and compare the efficiency of three rescoring strategies: (1) rescore everyone, (2) rescore failures only, and (3) rescore within some range of the passing cutoff. These rescoring strategies are evaluated by direct simulation. Additionally we consider the optimal allocation of rescores where the probability someone asks to be rescored is inversely proportional to the distance from their initial score to the cutscore. This allocation is evaluated via numerical integration. A further generalization of the basic model is also considered in which a test is comprised of a mixture of objectively and subjectively scored items.

Key words and phrases: Constructed responses, normal linear model, rater reliability, rescoring.

1. Introduction

"*The glorious endeavour that we know today as science has grown out of the murk of sorcery, religious ritual, and cooking. But while witches, priests, and chefs were developing taller and taller hats, scientists worked out a method for determining the validity of their results; they learned to ask, 'Are they reproducible*?' "

Scherr (1983)

An important goal of educational assessments is to produce scores for an individual that are reproducible. It is well known that if someone takes a test twice their two scores will not be identical. The variation observed may be the result of many influences; the person may have learned something between the two administrations, the person may have had a 'bad day' during one administration, subtle differences in the specific questions asked on the two different test forms

may have yielded a somewhat different level of success, etc.. Traditionally the variation in test score expected between two test administrations has been summarized by the statistic 'test reliability.' A more informed look at test reliability can be had through the decomposition of test score variability into its components. The development of 'generalizability theory' (Cronbach, Gleser, Nanda and Rajaratnam (1972)) has done much to popularize this approach. However the practice of summarizing all components of score variability into a single reliability figure remains by far the most common one. The reliability of most professionally prepared tests is approximately .92, which corresponds to rather small stochastic variability of test scores. Moreover, when variability is observed (due to an examinee retaking the test) the relative contribution of each of its causes is obscured leaving open the variability of the person's performance as the principal culprit.

The highly competitive and costly implications of providing university admission, course credits, and to some extent proper course placement has recently expanded the need for forms of assessment which go beyond those that, for the past half century, have been tested principally with multiple-choice items. Such demands may change the nature of many long standing examinations such as the Scholastic Assessment Test (SAT), Graduate Record Examination (GRE), Advanced Placement (AP) tests, and Graduate Management Admissions Test (GMAT) to include items that are responded to in a reasonably free format. Since an examinee must construct a response rather than choose one from among a set of specified choices such items are often called 'constructed response items'. The pantheon of constructed response items includes essays, detailed mathematics problems, portfolio submissions, and, in tests of music and dance, even actual performances. While these modified tests are considered to contain a richer assortment of information about a student's ability, there are difficulties associated with their implementation.

One class of difficulties that such item types engender stem from the fact that they typically require scoring by human judges. Not only is such scoring far more expensive than automated scoring by machine, but also the judge is an additional component of score variability. *Ceteris paribus*, the larger the proportion of the test that is judge-scored the lower the overall reliability. Until recently, it would be very rare for any examinee to request that their test be rescored. The reason for this is that if a test is completely scored through some objective procedure (e.g. multiple choice test items mechanically scored) there is only a minuscule likelihood that rescoring would yield any change. This likelihood increases enormously when the test score results from human judgment. In this latter case rescoring will almost certainly yield a change, indeed a change that has

nothing to do with anything different about the examinee. The issue of rescoring achieves special importance in situations, like those found in licensing tests, in which there is a specified passing score.

If examinees are allowed to have their tests rescored what will be the consequences? One obvious effect is that scores will change upon rescoring. Some who previously failed will pass; for some of these the change will be a correction of an earlier error, for others their passing will be an artifact of the testing system and ought not to have happened. How many errors will be corrected? How many new errors will be introduced? What sort of rescoring strategy provides the most help? How many score changes will we see? The answers to these questions depend on: the quality of the raters, the number of raters, the variability of the examinees, the cutting score, and the rules under which rescoring takes place.

Test rescoring is a practice that typically will differentially affect examinees of varying economic status and ability levels. That is, in the case where examinations are rescored only upon examinee request, such requests are more likely to emanate from those individuals who are able to afford it and who performed below their own expectations. While it is obvious that any assessment of the costs of rescoring must include increased usage of raters' time, one must also include the difficult to assess costs associated with the loss of confidence in the test's quality due to changing test scores, and that with an ill-chosen rescoring policy, the changed scores could be less accurate, in the aggregate, than the original scores. Thus it is difficult to know if the increased revenue that is obtained from rescoring outweighs all of the associated costs for that rescoring. Such outcomes depend on the variability of the system and the rescoring rules. The desire to provide a guide for the policy decisions regarding rescoring of examinations initiated this research.

In this study we examine the effects that each of the relevant variables has on both error rates and on the number of pass/fail classifications that change with rescoring. We describe an initial system to be used throughout for exemplary purposes in Section 2. In Section 3, we describe a model for subjectively rated item scores. Section 4 describes the relevant cross-classified table that specifies the entire set of error rates due to initial scoring and rescoring. Simulation results for a variety of specified rating systems (Section 5) are given in Section 6. A basic result for tests which are a mixture of multiple-choice and subjectively rated items is given in Section 7. Section 8 contains some concluding remarks.

2. An Example

To facilitate the description of the model, error rates, and notation we shall begin with an example that represents a typical testing paradigm. Although the

example is specific, it is done so without loss of generality; however when an explicit generalization is helpful we include it.

Consider a test administered to a population of I examinees comprised of one subjectively rated item (e.g. an essay). Each examinee's essay is *initially* assigned to a set of raters, say t_1 of them. This assignment and subsequent rating by t_1 raters is called the initial scoring period. After the initial scoring, each of the I examinees is assigned to one of two categories (IP) = Initial Pass and (IF)= Initial Fail. We consider the case where the assignment to IP and IF is based on the observed mean score of the first t_1 ratings, denoted $\bar{y}_{it_1}$, where examinee i is passed initially if $\bar{y}_{it_1}$ is greater than or equal to a cutoff c and failed initially if $\bar{y}_{it_1} < c$. Typically c is chosen so that a desired proportion of the population, p, is passed initially. After the initial scoring there are those examinees who are passed whose true ability (defined as an average over infinitely many tasks as rated by infinitely many raters; Lord and Novick (1968)) does not warrant passing (TF, True Failures) and similarly those that initially fail who should pass (TP, True Passers). These cases are initial scoring errors.

In this example, we consider the simple strategy of rescoring all examinees after t_1 scorings, called the rescoring period, with $T - t_1$ rescorings. More sophisticated (and efficient) strategies are described and evaluated in Section 6. This results in a total of T ratings for every examinee i; each score is denoted y_{ijt}, the score from examinee i from rater j in the tth scoring period, $t = 1, \ldots, T$. After all T ratings are observed, a final assignment is made to passing (RP = Rescoring Pass) or failing (RF = Rescoring Failure) based on $\bar{y}_{iT} \geq c$ and $\bar{y}_{iT} < c$ respectively. Errors at this stage (i.e. RP if TF or RF if TP) are called rescoring errors. Some rescoring errors are remaining from the initial scoring and some are caused by the rescoring process. A description of results for alternative scoring rules such as taking the maximum, and minimum over all observed scores appears in Wainer and Bradlow (1996).

For example, suppose we are faced with having to admit 50 students out of an applicant pool of 100, on the basis of an essay test that is scored initially by $t_1 = 2$ expert raters. A common finding in such situations (Ruggles (1911), Bock (1991), Linn (1994)) is that the variability due to judges is at least as large as that associated with students. It is natural to ask how many of the admissions decisions that we make on the basis of these ratings are likely to be in error. And, as a follow-on question, how many of those errors can be corrected if we rescore each test with $T - t_1 = 2$ additional raters and use the mean of all four ratings for our decision? This situation corresponds with simulation 13 in Table 2 and is more fully described in Section 6. The answer is that we would expect about 72 of the decisions based on the initial scoring to be correct and 77 to be

correct after rescoring. Thus the 200 extra rescorings results in a reduction of 5 of the incorrect decisions. If we know the cost of rescoring and the costs of errors we can decide on an efficacious scoring strategy.

Since the error rates mentioned are all functions of the unobserved examinee true score, we must specify a model relating the observed scores y_{ijt} to the true scores for each examinee denoted τ_i. This is done next.

3. The Model

We propose a simple additive linear model for observed scores y_{ijt} given by

$$\text{observed score} = \text{true score} + \text{error}$$
$$y_{ijt} = \tau_i + e_{ijt}.$$

True score (τ_i) is given an additive form with random examinee intercept (ability), $\tau_i = \mu + \alpha_i$. Error is decomposed into two parts, $e_{ijt} = \beta_j + \epsilon_{ijt}$; with β_j characterizing rater j severity, and ϵ_{ijt} random variance. This yields the random effects model for examinee scores

$$y_{ijt} = \mu + \alpha_i + \beta_j + \epsilon_{ijt}. \tag{1}$$

The assumptions in (1) of person by rater and rater by scoring period independence are likely to hold if rater-blind scoring and efficient allocation of exams to raters is utilized.

Additional specification of the model is assumed by asserting independent Gaussian prior distributions for the random effects with

$$\alpha_i \sim N(0, \sigma_\alpha^2), \beta_j \sim N(0, \sigma_\beta^2), \text{ and } \epsilon_{ijt} \sim N(0, \sigma_\epsilon^2). \tag{2}$$

We treat μ the average level of response as a fixed effect. The sensitivity of results to the Gaussian distribution assumption are given in detail in Wainer and Bradlow (1996).

Under the model given by (1) and (2), the entire scoring system is specified by the parameter vector $\eta = (c - \mu, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\epsilon^2)$. We utilize a reparameterization of η that is more easily understood (and more commonly used) among education researchers given by:

$$\begin{aligned}
\text{Var}(y_{ijt}) &= \sigma_y^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\epsilon^2 \\
\text{Cor}(y_{ijt}, y_{ij't}) &= \sigma_\alpha^2/\sigma_y^2 \\
\text{Cor}(y_{ijt}, y_{i'jt}) &= \sigma_\beta^2/\sigma_y^2 \\
\text{Prob(IP)} &= \text{Prob}(\bar{y}_{it_1} \geq c)
\end{aligned}$$

the marginal variance of the observed scores, the inter-rater reliability, the percent of the total variation due to judges, and the desired initial passing rate respectively.

One can simulate the entire system for any set of values desired and let prospective users pick those values that represent their specific situation. In Section 6 of this study we provide simulation results for a moderate and extreme value (0.5 and 0.9) of the passing rates.

4. Description of Error Rates

Since our problem can be specified in terms of an initial pass decision followed by a final pass decision given some true ability status, all of the events of interest can be summarized in the 2 X 2 X 2 event table, shown here as Table 1.

Table 1. Event table for pass/fail system with initial and resoring periods.

True Pass

	Rescore Pass	Rescore Fail
Initial Pass	IV	III
Initial Fail	I	II

True Fail

	Rescore Pass	Rescore Fail
Initial Pass	II	I
Initial Fail	III	IV

Type I events are initial scoring errors (ISEs) which are fixed by rescoring. These are a gain due to rescoring. Type II events are ISEs which are not fixed by rescoring. Although no score changes occur here there is an associated cost for rescoring as well as an opportunity loss to rescoring a faulty initial decision and not correcting it. Certainly, Type II events would be more common for examinees whose true scores are near the cut score and/or the condition where the number of rescores $T - t_1$ is small compared to the number of initial scorings t_1. Type III events are those cases where errors are created due to rescoring. A cost is clearly associated with these cases and typically one goal of a rescoring system would be to minimize Type III events. Type IV events are cases in which erroneous decisions are never made. We note that under these definitions, ISEs are simply the union of Type I and Type II events and that rescoring errors are the union of Type II and Type III events. Therefore, if the percentage of Type III events is less than that of Type I events there will be fewer errors after the rescoring period than after the initial scoring period.

All of the probabilities corresponding to the events in Table 1 can be computed sequentially from the set of probabilities containing the initial conditions: P(TP) and P(TF), ISEs: P(IP|TF) and P(IF|TP), and rescoring probabilities: P(RF|TF, IP), P(RP|TP, IF), P(RP|IF, TF) and P(RF|IP, TP). The initial conditions would typically be set by design and are taken as given. We compute the remaining probabilities via direct simulation for various values of η in Section 6.

5. Various Rescoring Policies

The example given in Section 2 and its description which follows assumes that all examinees are scored by an initial set of t_1 ratings followed by a rescoring set of $T - t_1$ ratings. This policy, although equitable in the sense of number of scores per individual, is probably neither practical or efficient. We present two alternatives to the "rescore everybody" policy; "rescore only those who fail", and "rescore only those whose initial score is near the cutscore". We also provide some results on the optimal allocation of judges where the probability someone asks to be rescored is inversely proportional to their distance from the cutscore.

If there were no monetary cost for the examinee associated with rescoring, then rescoring all failures would correspond to rescoring all who would ask to be rescored. As the monetary cost increases we would expect that those who ask to be rescored would be connected to their socio-economic status as well as their distance from the cutscore. Under this strategy, we get some simplification of error rates in that people who initially pass are not rescored and therefore must pass after rescoring (i.e. Prob(RP|IP, TP) = Prob(RP|IP, TF) = 1 and Prob(RF|IP, TP) = Prob(RF|IP, TF) = 0). Rescoring only those examinees within some range of the cutscore could correspond to a quality control strategy implemented by a testing organization regardless of whether or not the examinee asks (or pays) to be rescored. Rescoring with probability inversely proportional to the distance from the cutscore is a realistic mechanism for modeling who would ask for rescoring.

6. Simulation

6.1. Simulation design

The evaluation of the error rates described in Sections 4 and 5 were accomplished through direct simulation. The simulation factors that were manipulated correspond to parameter η given in Section 3. To standardize the simulation values the marginal variance of the observed scores, $\text{var}(y_{ij})$, was set equal to one. The number of initial raters was set at $t_1 = 2$ and the number of rescoring raters at $T - t_1 = 2$ as in Section 2. Changing the number of initial and resoring raters would lead to a rescaling of the rater variance; however, since the marginal variance is set to one, the effect of the change in number of raters is measured by varying the rater variance between 0 and 1.

A new independent judge was drawn for each rating. This only partially corresponds with practice, in that when an indivdual exam is rescored a new judge is sampled from a pool of judges without replacement. This correponds to our simulation precisely. We depart from practice when we aggregate across examinees, for in practice the same judges are used repeatedly for different examinees. We did not attempt to model this since the within judge effects are quite

complex including, as they do, such factors as the individual foibles of each judge, as well as differential fatigue and learning effects. The size of such judge effects are uncertain, and we felt the role of rescoring, uncontaminated by other issues, would be more clearly visible with independntly chosen judges. We therefore left the complex task of modeling idiosyncratic judge behavior to other accounts.

The simulation experiment was a $3 \times 2 \times 3 \times 3$ full factorial design with levels

(i) rescore everyone, rescore failures, rescore within some range of the cutscore
(ii) Prob(IP) = 0.5, 0.9
(iii) $\sigma_\epsilon^2 = 0.1, 0.5, 0.9$
(iv) $\sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\beta^2) = 1/6, 1/2, 5/6$.

yielding a total of 54 simulation conditions. For each of the three rescoring strategies (item (i)), we simulated 250,000 examinees at each of the 18 simulation conditions (a sample size chosen to make the maximum standard error equal to 0.001). The IP conditions were chosen at 0.5 to yield the largest binomial variability, and 0.9 to represent a practical and liberal passing standard (e.g. a certification test). The condition with Prob(IP) = 0.1 is symmetric with Prob(IP) = 0.9 and hence not included. The values of the random variance, σ_ϵ^2, were chosen to represent a small amount of random variability ($\sigma_\epsilon^2 = 0.1$) corresponding to a carefully structured test with many objectively scored items (e.g., the SAT), a test with a weakly defined structure ($\sigma_\epsilon^2 = 0.9$) and rather few items (e.g., the Vermont Portfolio Assessment), and a test with a balanced structure ($\sigma_\epsilon^2 = 0.5$). The ratio of true score variance to the sum of true score variance and variance due to raters ($\sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\beta^2)$) was assigned three values ranging from 1/6 – true score variance is very small compared to rater variance, 1/2 – true score variance is the same as rater variance, and 5/6 – true score variance is very large compared to rater variance. Because, in this simulation, we define the sum of the three variances to be one, the exact values of the true score variance and the rater variance will vary as a function of this ratio but also as a function of the amount of random variance.

For the rescoring strategy, "rescore with some range of the cutscore", we chose to rescore the nearest 10% of the examinees on either side of the cutscore. As the simulation conditions were set at Prob(IP) = 0.5, 0.9 this corresponds to rescoring all examinees in the range $[\hat{F}^{-1}(0.4), \hat{F}^{-1}(0.6)]$ for the Prob(IP) = 0.5 condition, and all examinees in the range $[\hat{F}^{-1}(0.8), \hat{F}^{-1}(1.0)]$ for the Prob(IP) = 0.9 condition where $\hat{F}$ is the empirical CDF of the initial score distribution. It is important to note that for the conditions with Prob(IP) = 0.9 rescoring within 10% of the cutscore rescores all those who fail (the lowest decile of the CDF) and the 10% who barely passed. This is in contrast to the Prob(IP) = 0.5 conditions in which rescoring within 10% of the cutscore rescores only a small subset of the failures (1/5 of them in this case).

6.2. Simulation results

The simulation output is presented in terms of six useful summaries computed from each of the 54 $2 \times 2 \times 2$ event tables (as in Table 1). The summaries included (along with abbreviated notation) are the *percentage* of:

1. Initial scoring errors (ISE),
2. Errors remaining after rescoring (EReR),
3. Error improvement due to rescoring (EIR),
4. Errors corrected by rescoring (ECoR),
5. Errors created by rescoring (ECrR) and
6. Scores changed due to rescoring (SCR).

By definition, we know that EIR = ISE - EReR, ECoR = EIR + ECrR, and SCR = ECoR + ECrR.

Since we are most interested in comparing these summaries across the three rescoring strategies ("Rescoring everyone" is denoted by EVERY, "Rescoring Failures" by FAIL, and "Rescoring those within 10% of the cutscore" by BOUNDS) and by initial passing probability, we group the results into 9 triplets of results for Prob(IP) = 0.5 (Table 2) and 9 triplets for Prob(IP) = 0.9 (Table 3). The inferences presented are based on the significant effects from an ANOVA run on the summary results.

A number of confirmatory and interesting findings are exhibited in Table 2 for the Prob(IP) = 0.5 conditions. As expected we observe: (a) for every condition all of the changes due to rescoring in EVERY are approximately double that in FAIL (symmetry), (b) ISE, SCR, and ECrR are an increasing function of random variance σ_ϵ^2, (c) the percentage of score changes which are corrections (ECoR/SCR) is a decreasing function of σ_ϵ^2, and (d) the number of errors remaining after rescoring (EReR) is in its expected order EVERY<BOUNDS<FAIL. The more interesting findings were: (a) the high ISE and SCR rates across all conditions, and (b) the increased efficiency (EIR per rescore) of the BOUNDS rescoring strategy indicated by $(1/0.2) \cdot \text{BOUNDS} >> (1/0.5) \cdot \text{FAIL} \approx (1/1) \cdot \text{EVERY}$ across all simulation conditions; however this efficiency gain decreases as a function of random variance σ_ϵ^2.

The findings for the Prob(IP) = 0.9 conditions (Table 3) were more surprising. Specifically, we found that (a) in many conditions Prob(ISE) = Prob(IF) indicating that no initial scoring is needed to achieve the same ISE rate (!!) - a further description of this result is given in Section 7.2, (b) the number of errors remaining after rescoring (EReR) is ordered by FAIL<BOUNDS<EVERY, for all but simulations 28-30 in which there is near equality, indicating that increasing the number of rescores for those who pass initially actually *increases* the number of errors after rescoring (!!), and (c) almost all score changes due to rescoring

are due to corrections as seen by the large values of ECoR/SCR across all conditions. Result (b) is mostly due to the fact that under the rescore FAIL strategy, considerably fewer errors are created due to rescoring (ECrR) than under the EVERY or BOUNDS strategy.

Table 2. Summary of percentage errors for three rescoring strategies with Prob(IP) = 0.5. Each simulation is based on 250,000 simulees. We abbreviate ISE = Initial scoring errors, EReR = Errors remaining after rescoring, EIR = Error improvement due to rescoring, ECoR = Errors corrected by rescoring, ECrR = Errors created by rescoring, and SCR = Scores changed due to rescoring.

sim	Strat.	σ_ϵ^2	$\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\beta^2}$	ISE	EReR	EIR	ECoR	ECrR	SCR
1	EVERY	0.1	5/6	12.3	8.9	3.4	6.0	2.6	8.6
2	FAIL	0.1	5/6	12.2	10.6	1.6	2.9	1.3	4.2
3	BOUNDS	0.1	5/6	12.2	9.4	2.8	4.7	1.9	6.6
4	EVERY	0.1	1/2	21.1	16.0	5.1	9.7	4.6	14.3
5	FAIL	0.1	1/2	21.0	18.5	2.5	4.8	2.3	7.1
6	BOUNDS	0.1	1/2	21.1	18.0	3.1	5.5	2.4	8.0
7	EVERY	0.1	1/6	32.9	27.8	5.1	12.9	7.9	20.8
8	FAIL	0.1	1/6	33.0	30.4	2.6	6.5	3.9	10.4
9	BOUNDS	0.1	1/6	32.7	30.3	2.4	5.5	3.1	8.7
10	EVERY	0.5	5/6	22.1	17.0	5.1	10.1	4.9	15.0
11	FAIL	0.5	5/6	22.3	19.7	2.6	5.0	2.5	7.5
12	BOUNDS	0.5	5/6	22.2	19.1	3.1	5.6	2.5	8.0
13	EVERY	0.5	1/2	28.2	22.7	5.5	12.0	6.5	18.4
14	FAIL	0.5	1/2	28.1	25.5	2.6	5.9	3.3	9.2
15	BOUNDS	0.5	1/2	28.3	25.5	2.8	5.7	2.9	8.5
16	EVERY	0.5	1/6	37.3	32.9	4.4	13.5	9.1	22.5
17	FAIL	0.5	1/6	37.1	35.0	2.1	6.6	4.5	11.2
18	BOUNDS	0.5	1/6	37.1	35.2	1.9	5.4	3.4	8.9
19	EVERY	0.9	5/6	37.3	32.9	4.4	13.5	9.1	22.6
20	FAIL	0.9	5/6	37.3	35.0	2.3	6.9	4.5	11.4
21	BOUNDS	0.9	5/6	37.3	35.4	1.9	5.4	3.5	8.9
22	EVERY	0.9	1/2	40.0	36.3	3.7	13.6	9.9	23.4
23	FAIL	0.9	1/2	40.0	38.2	1.8	6.8	5.0	11.8
24	BOUNDS	0.9	1/2	40.2	38.6	1.6	5.3	3.7	8.9
25	EVERY	0.9	1/6	44.2	41.9	2.3	13.4	11.1	24.5
26	FAIL	0.9	1/6	44.3	42.9	1.4	6.8	5.4	12.2
27	BOUNDS	0.9	1/6	44.2	43.2	1.0	4.9	3.9	8.9

Table 3. Summary of percentage errors for three rescoring strategies with Prob(IP) = 0.9. Each simulation is based on 250,000 simulees. We abbreviate ISE = Initial scoring errors, EReR = Errors remaining after rescoring, EIR = Error improvement due to rescoring, ECoR = Errors corrected by rescoring, ECrR = Errors created by rescoring, and SCR = Scores changed due to rescoring.

sim	Strat.	σ_ϵ^2	$\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\beta^2}$	ISE	EReR	EIR	ECoR	ECrR	SCR
28	EVERY	0.1	5/6	5.2	3.5	1.7	2.7	1.1	3.7
29	FAIL	0.1	5/6	5.2	3.7	1.5	1.9	0.4	2.3
30	BOUNDS	0.1	5/6	5.2	3.6	1.6	2.6	1.0	3.6
31	EVERY	0.1	1/2	8.1	5.3	2.8	4.3	1.6	5.9
32	FAIL	0.1	1/2	8.1	4.8	3.3	3.7	0.4	4.1
33	BOUNDS	0.1	1/2	8.0	5.1	2.9	4.2	1.3	5.5
34	EVERY	0.1	1/6	9.9	5.2	4.7	6.4	1.7	8.0
35	FAIL	0.1	1/6	9.9	3.7	6.2	6.3	0.1	6.3
36	BOUNDS	0.1	1/6	9.9	4.6	5.3	6.3	1.0	7.3
37	EVERY	0.5	5/6	8.3	5.4	2.9	4.5	1.6	6.2
38	FAIL	0.5	5/6	8.3	4.8	3.5	3.9	0.4	4.3
39	BOUNDS	0.5	5/6	8.4	5.3	3.1	4.4	1.3	5.7
40	EVERY	0.5	1/2	9.6	5.7	3.9	5.6	1.7	7.3
41	FAIL	0.5	1/2	9.5	4.4	5.1	5.3	0.2	5.6
42	BOUNDS	0.5	1/2	9.5	5.2	4.3	5.5	1.2	6.7
43	EVERY	0.5	1/6	10.0	4.6	5.4	7.0	1.6	8.6
44	FAIL	0.5	1/6	10.0	3.0	7.0	7.0	0.0	7.0
45	BOUNDS	0.5	1/6	10.0	3.9	6.1	7.0	0.9	7.8
46	EVERY	0.9	5/6	10.0	4.5	5.5	7.0	1.5	8.5
47	FAIL	0.9	5/6	10.0	3.0	7.0	7.0	0	7.0
48	BOUNDS	0.9	5/6	10.0	3.9	6.1	6.9	0.8	7.8
49	EVERY	0.9	1/2	10.0	4.2	5.8	7.3	1.4	8.7
50	FAIL	0.9	1/2	10.0	2.8	7.2	7.2	0.0	7.2
51	BOUNDS	0.9	1/2	10.0	3.5	6.5	7.3	0.8	8.1
52	EVERY	0.9	1/6	10.0	3.7	6.3	7.6	1.3	8.9
53	FAIL	0.9	1/6	10.0	2.4	7.6	7.6	0	7.6
54	BOUNDS	0.9	1/6	10.0	3.1	6.9	7.6	0.7	8.3

6.3. Optimal allocation of judges

In this section, we consider the following sequential allocation problem. Assume that the initial scoring period has occurred with each of I examinees receiving t_1 initial scorings and a resulting score $\bar{y}_{it_1}$. Further, suppose that a fixed

budget of $K = I \cdot T > I \cdot t_1$ scorings is allowed for the total scoring period. This setup is identical to that of Sections 2 - 6. We address the problem of how to optimally allocate the remaining $I \cdot (T - t_1)$ scorings to the I examinees given their observed initial scores in order to minimize the total expected misclassification error. This analysis is in contrast to Sections 6.1 and 6.2 in which a one-time allocation decision leads to each examinee getting 0 or $T - t_1$ additional rescores. Intuitively, one would expect that an examinee whose score is initially far from the cutscore would require a lower share of resource allocation than one initially near the cutscore. We demonstrate the computation and provide details below.

The total expected misclassification error after rescoring is given by:

$$\text{Prob}(ME) = \sum_{i=1}^{I} \text{Prob}(\bar{Y}_{in_i} \geq c, \mu + \alpha_i < c) + \text{Prob}(\bar{Y}_{in_i} < c, \mu + \alpha_i \geq c), \quad (3)$$

where $\bar{Y}_{in_i}$ is the final average rating of the ith examinee after $n_i \geq t_1$ scorings, and μ, α_i, and c are the grand mean, examinee ability, and cutoff as previously described. This classification rule though simple has the drawback of equally penalizing serious and minor misclassifications equally. A generalization of our results to other error rules is a nice area for future consideration.

Noting that

$$\bar{Y}_{in_i} = \Big(\sum_{j=1}^{t_1} Y_{ij} + \sum_{j=t_1+1}^{n_i} Y_{ij} \Big) / n_i = (t_1 \bar{Y}_{it_1} + n_{i2} \bar{Y}_{in_{i2}}) / (t_1 + n_{i2})$$

we obtain from (3)

$$\begin{aligned}
&\text{Prob}(ME) \\
&= \sum_{i=1}^{I} \Big(\text{Prob}(\bar{Y}_{in_{i2}} \geq k_1, \mu + \alpha_i < c) + \text{Prob}(\bar{Y}_{in_{i2}} < k_1, \mu + \alpha_i \geq c) \Big) \\
&= \sum_{i=1}^{I} \Big(\int_{-\infty}^{c} \text{Prob}(\bar{Y}_{in_{i2}} \geq k_1 | \mu + \alpha_i = t) \phi_{\mu+\alpha_i}(t) dt \\
&\qquad + \int_{c}^{\infty} \text{Prob}(\bar{Y}_{in_{i2}} < k_1 | \mu + \alpha_i = t) \phi_{\mu+\alpha_i}(t) dt \Big) \\
&= \sum_{i=1}^{I} \Big(\int_{-\infty}^{c} (1 - \Phi(\frac{k_1 - t}{\sqrt{k_2}})) \phi_0(\frac{t - \mu}{\sigma_\alpha}) dt + \int_{c}^{\infty} (1 - \Phi(\frac{k_1 - t}{\sqrt{k_2}})) \phi_0(\frac{t - \mu}{\sigma_\alpha}) dt \Big), \quad (4)
\end{aligned}$$

where $k_1 = \frac{(c - \bar{Y}_{it_1}) t_1 + c n_{i2}}{n_{i2}}$, $k_2 = (\frac{\sigma_\beta^2 + \sigma_\epsilon^2}{n_{i2}})^{0.5}$, and Φ, ϕ_z are the standard normal cdf and normal pdf with mean z respectively. We computed the one-dimensional integral in (4) by numerical integration over a grid of 200 equally spaced grid

points. We found that a larger grid and/or more wisely chosen grid points yielded no significant difference in accuracy. The constants k_1 and k_2 are functions of the unknown rescoring counts n_{i2}; yet, for any given allocation $n_{12}, \ldots, n_{I2}$ the expected error rate can be computed. Since there exists a very large class of allocation functions from which to choose we consider a specific class of allocation functions given by $n_{i2} \propto d(c, \bar{Y}_{it_1})^r$, where $d(c, \bar{Y}_{it_1})$ is the absolute distance between the initial score for examinee i and the cutoff, and $r \leq 0$. We search among values of r which yields the smallest value of (4). For example, if $r = 0$ minimizes (4) then this suggests that all examinees receive the same number of rescores, $(K - t_1 \cdot I)/I$ regardless of their initial score. The more negative the value of r, the larger the share of resources that are allocated to observations near the cut score. We provide a simulation experiment below.

6.4. A simulation experiment

Our simulation experiment consisted of generating three parallel sequences of test scores using the same plausible value of $\eta = (\sigma_\epsilon^2 = 0.33, \sigma_\alpha^2/(\sigma_\alpha^2+\sigma_\beta^2) = 1/2)$ and Prob(IP) $= 0.5$ where η is as defined in Section 2. The three parallel simulations all set $t_1 = 1$, the number of initial scorings, and $I = 100$, the number of examinees; however, K the total number of scorings was varied at 200, 300, and 400 respectively (i.e. 100, 200, and 300 rescorings to allocate among the 100 examinees). This process was repeated 50 times to account for the sampling variability of test scores. The mean optimal value of r was $-.15$, $-.09$, and, $-.09$ for $K = 200, 300, 400$ respectively. The corresponding mean reductions in expected number of errors over equal allocation of rescores (i.e. $r = 0$) were 3.6, 4.5, and 5.0 respectively.

The allocation function $d(c, \bar{Y}_{it_1})^r$, the parameter η, and values of K were chosen to demonstrate (a) the potential for reduction in errors through a simple yet more clever allocation of judges scorings, and (b) the order of magnitude of these gains in an exemplary case. Intuition suggests that in cases where the residual error variance σ_ϵ^2 is larger, there is more potential for gains. Furthermore, we also consider that our allocation process is a "one-time" rescoring and that a more sequential procedure with many stages of resource allocation may be desired. However, careful inspection suggests that our procedure would still be applicable in a sequential nature where the inputs would simply change from $\bar{Y}_{it_1}$, t_1 to their updated values.

7. Further Results

7.1. Tests containing both objectively and subjectively scored items

We consider a set of basic results for tests which are comprised of a mixture of objectively (e.g. multiple choice) and subjectively (e.g. an essay) scored items.

These results can be used in conjunction with the simulation results of Section 6 by noting that increasing the proportion of the test with objectively scored items would lead to lower values of random variance σ_ϵ^2.

Let x_i be the score for examinee i on a subtest that is scored without rater error, and ω be the proportion (i.e., weight) of the total score due to x_i.

Assume that

$$x_i = \mu + \alpha_i + \delta_i,$$

where $\mu + \alpha_i$ is the true score as defined in (1) and δ_i is the true score residual. The observed score is

$$\begin{aligned} y_{ij}^* &= \omega x_i + (1-\omega) y_{ij} \\ &= \mu + (\alpha_i + \omega\delta_i) + (1-\omega)(\beta_j + \epsilon_{ij}). \end{aligned}$$

Notice that the new model has the same form as the original model. The inter-rater reliability for the new model is

$$r^* = \frac{\sigma_\alpha^2 + \omega^2\sigma_\delta^2}{\sigma_\alpha^2 + \omega^2\sigma_\delta^2 + (1-\omega)^2(\sigma_\beta^2 + \sigma_\epsilon^2)}.$$

When $\omega > 0$,

$$r^* - r = \frac{(\sigma_\beta^2 + \sigma_\epsilon^2)[\omega(2-\omega)\sigma_\alpha^2 + \omega^2\sigma_\delta^2]}{(\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\epsilon^2)[\sigma_\alpha^2 + \omega^2\sigma_\delta^2 + (1-\omega)^2(\sigma_\beta^2 + \sigma_\epsilon^2)]} > 0.$$

That is, the inter-rater reliability for the new model is larger than that for the original model. Moreover, r^* is a strictly increasing function of ω $(0 \le \omega \le 1)$. As a matter of fact,

$$\frac{\partial r^*}{\partial \omega} = \frac{2(1-\omega)(\sigma_\alpha^2 + \omega\sigma_\delta^2)(\sigma_\beta^2 + \sigma_\epsilon^2)}{[\sigma_\alpha^2 + \omega^2\sigma_\delta^2 + (1-\omega)^2(\sigma_\beta^2 + \sigma_\epsilon^2)]^2} > 0 \qquad \text{for} \quad 0 < \omega < 1.$$

Obviously since the initial error probabilities are decreasing functions of the inter-rater reliability, the larger ω is, the lower the initial error probabilities are.

7.2. Other interesting results

An interesting scenario related to ISEs can occur when the judges scoring the examinees are of mediocre quality. Specifically, suppose there exists an examination with a very high true passing rate, say P(TP) = 0.90, and judges who perform poorly enough so that P(ISE) > P(TF) = 0.10. In this case, a lower ISE rate can be achieved by not scoring the examinees and passing everyone! Clearly

if P(TP)=0.90, then P(ISE) associated with passing everyone is P(ISE)=0.10. This result is obtained when

$$\begin{aligned} P(ISE) > P(TF) \Leftrightarrow \\ P(IP,TF) + P(IF,TP) > P(IP,TF) + P(IF,TF) \Leftrightarrow \\ P(IF,TP) > P(IF,TF), \end{aligned} \tag{5}$$

the probability of failing a true passer is greater than the probability of failing a true failure. We note that in Simulations 43-54 (Table 3) equality of the ISE rate and true failure rate (TF) was achieved.

A related result for a slightly different problem can also be examined using the ISEs. Consider the case in which an examination is given, the test scores are observed, and the highest scoring $x\%$ of the examinees are passed. If the P(ISE) in this case is greater than $2x(1-x)$, the error rate associated with the random assignment of examinees to pass with probability x and fail with probability $1-x$, then random assignment is a cost effective strategy. This did not occur in any of the simulation conditions; however for Simulations 22-27 (Table 2) the ISE rate is certainly approaching 2 (0.5)(0.5) = 0.50.

8. Concluding Remarks

There are a number of potentially useful results of this research. First, for tests in which the number of initial failers and passers are approximately equal, the most efficient strategy is to only rescore examinees near the cutscore. Second, for tests in which almost all examinees are expected to pass, the total misclassification rate after rescoring is minimized by only rescoring those who fail. These two findings have immediate implications for rater allocations. Additional findings in this research also suggest that when the error associated with scoring a subjective item dominates the variability in the true scores that (a) it may not pay to score at all yielding a strategy to pass everyone, and (b) random assignment of the pass/fail condition may lead to lower error rate (and certainly lower cost) than the use of the test scores.

Acknowledgement

The authors thank Jinming Zhang for his work on earlier versions of this research. This work grew out of a discussion of new problems facing licensing exams at the monthly COPA Research Seminar. We are grateful to the participants of that seminar for their helpful discussions. This research was supported by ETS's Research allocation to the Research Statistics Group.

References

Bock, R. D. (1991). The california assessment. A talk given at the Educational Testing Service, Princeton, NJ. on June 17, 1991.

Boodoo, G. (1995). Post Hoc reliability analyses of a range scale portfolio assessment, presented at NCME 1995 - San Francisco.

Cronbach, L. J., Gleser, G., Nanda, H. and Rajaratnam, N. (1972). *The Dependability of Behavioral Measures: Theory of Generalizability for Scores and Profiles*. Wiley, New York.

Feldt, L. S. and Brennan, R. L. (1993). Reliability. In *Educational Measurement*, 3rd edition (Edited by R. L. Linn), 105-146. Phoenix, AZ: Oryx Press.

Kelley, T. L. (1947). *Fundamentals of Statistics*. Cambridge: Harvard University Press.

Linn, R. L. (1994). Performance assessment: policy promises and technical measurement standards. *Educational Researcher* **23(9)**, 4-14.

Lord, F. M. and Novick, M. (1968). *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison Wesley.

Ruggles, A. M. (1911). *Grades and Grading*. : Teacher's College, New York.

Scherr, G. H. (1983). Irreproducible science: editor's introduction. *The Best of the Journal of Irreproducible Results*. Workman New York Publishing.

Wainer, H. and Bradlow, E. T. (1996). On the consequences of some test rescoring policies. ETS Technical Report 96-5.

Department of Marketing and Statistics, Wharton School of Business, University of Pennsylvania, 3620 Locust Walk, Philadelphia, PA 19104-6371, U.S.A.

E-mail: ebradlow@wharton.upenn.edu

Statistics and Psychometric Research Group, Educational Testing Service, Rosedale Road, MS 15-T, Princeton, NJ 000-000, U.S.A.

E-mail: hwainer@ets.org

(Received March 1996; accepted October 1997)