DOCUMENT RESUME

ED 401 305 TM 025 855

AUTHOR Thomas, Leslie; Kalohn, John C.
TITLE Weighting Tasks from Job Analysis Data To Derive Test Specifications for Licensure Examinations: Some Methodological and Statistical Considerations.
PUB DATE Apr 96
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Job Analysis; *Licensing Examinations (Professions); *Mathematical Models; *Research Methodology; Scores; *Test Content; Test Results
IDENTIFIERS Mahalanobis Distance Function; *Test Specifications; *Weighting (Statistical)

ABSTRACT

Test specifications dictate the kind of content that should be included on each form of an examination, and the relative weight that each content domain should contribute to the determination of examinees' test scores by specifying the proportion of items to be included in each content area. This paper addresses a step in the development of specifications: combining job analysis rating data to determine an index of importance for each task. The Kane procedure (Kane et al, 1989), a multiplicative model, was compared with a new procedure based on a modified version of the Mahalanobis Distance (MD) measure. A conceptual model was also proposed to evaluate the results when comparing such procedures. A small data set of 5 task statements and 10 respondents was used as an example of data to be analyzed by both procedures. The Kane weights produced orderings of task statements that did not meet the expected ordering based on the model, except when criticality was weighted by a factor of 10. The MD weights approached a limit as the criticality weight was increased. The conceptual model proposed gives a logical rank ordering of tasks, but does not provide a means to translate this ordinal information into reasonable task weights. An ideal conceptual model would include a rationale for evaluating a weighting scheme. Preliminary analyses of the actual job data with each approach demonstrated that both methods failed to produce weights that were compatible with the model presented. Research on a more comprehensive conceptual model is being planned. (Contains 5 figures, 12 tables, and 11 references.) (SLD)

Weighting Tasks from Job Analysis Data to Derive Test Specifications for Licensure Examinations: Some Methodological and Statistical Considerations

Leslie Thomas and John C. Kalohn

American College Testing

April 1996

Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York.

One of the most important concerns in the development of any examination is that the test questions collectively measure what they were intended to measure. This is required for the examination to produce valid scores. To ensure that each form of an examination measures the same content, test specifications are developed and used for guiding the examination development process. Test specifications not only dictate the *kind* of content that should be included on the examination, but also the relative weight that each content domain should contribute to the determination of examinees' test scores by specifying the proportion of items to be included in each content area.

The Development of Test Specifications

Test specifications can be developed using a number of approaches. One process is depicted in Figure 1. The first step in this process is to have subject matter experts (SME) generate task statements to be included on a survey. The survey is then administered to a large, representative sample of practitioners who rate each task statement on each of several rating scales, such as frequency, criticality, or difficulty of learning. The survey data are then analyzed and an index of importance is computed for each task statement by combining the information obtained from the rating scales. SMEs are then asked to generate a list of knowledge, skills, and abilities (KSA) necessary to perform these tasks and then explicitly link each task with the KSAs required to perform that task.[1] Test specifications are then derived from these KSA-task linkages and are reviewed by the SMEs.

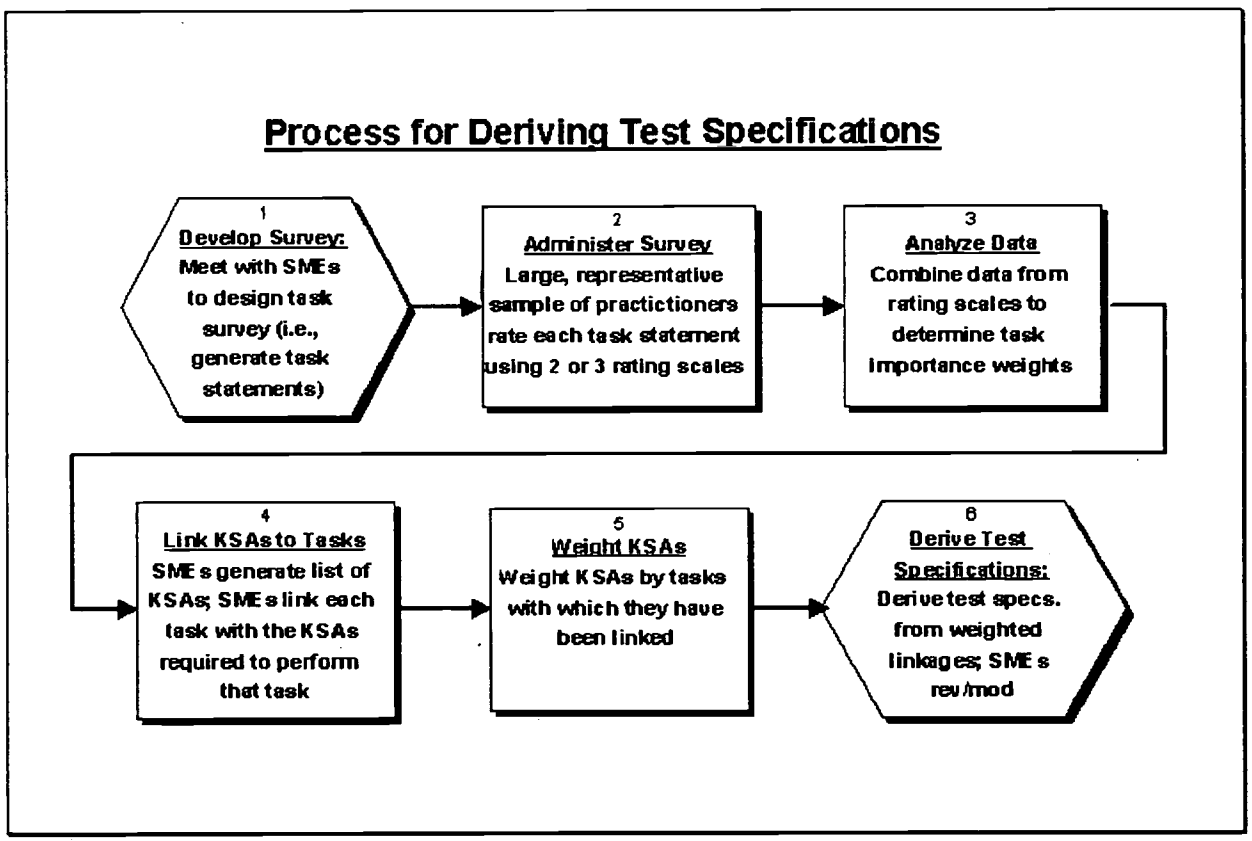


Figure 1.

[1] For convenience, this paper will use the term KSA, even though the job analysis and test specifications for licensure examinations will not focus on abilities in the formal sense of the word. Abilities often refer to concepts such as verbal, quantitative, logical reasoning, and problem-solving. Such traits are beyond the scope of licensure examinations, even though they may be suitable for personnel selection tests.

Purpose of the Present Paper

This paper addresses the third step in the process as depicted in Figure 1: combining job analysis rating data to determine an index of importance for each task. Often task statements are rated using two rating scales such as the *frequency* with which the task is performed and the *criticality* of that task to some standard (e.g., public safety, maximum job performance). The data from these scales are then combined - either statistically or holistically - to create an index of importance for each task (Raymond, in press). A number of statistical models have been proposed for combining the rating scale data, including multiplicative models (Kane, Kingsbury, Colton, & Estes, 1989), Rasch models (Lunz, Stahl, & James, 1989), and simple linear models (Sanchez & Levine, 1989). These models vary in terms of their mathematical complexity, the number and kinds of variables they include and the degree of emphasis given to each variable.

Although studies have indicated that different models may yield similar results (Sanchez & Frazer, 1992), we compared the Kane procedure (Kane et al., 1989) with a new procedure based on a modified version of the Mahalanobis distance measure and found that the results of these procedures *did not* agree. Unfortunately, if the results of two or more combining procedures do not agree, there is no accepted means by which to determine which model is preferred. One purpose of this paper is to propose a conceptual model that can be used to evaluate the results when comparing such procedures. An explicit conceptual model should exist to logically evaluate the results of the statistical models. As Ebel (1977) noted "one should never apologize for having to exercise judgment to validate a test. Data never substitute for good judgment" (p. 59). In addition, measurement attributes that should be included in a statistical model are discussed. Finally, the results of the Kane procedure and the modified Mahalanobis procedure are compared using the conceptual model as the basis for evaluation.

Conceptual Model for Combining Rating Scale Data

A problem that plagues psychological measurement is that we often lack an objective reality by which to judge the accuracy of the measurements. Although various statistical models have been proposed to determine the "true" importance weights for tasks, it is often necessary to step away from the data and decide whether the numerical results are congruent with the purpose of the job analysis and, ultimately, that of the test for which it is being conducted.

Logically, the test specifications for a selection examination should differ from those of a licensure examination due to differences in the purposes of these tests. The purpose of a selection test is to predict maximum job performance while the purpose of a licensure examination is to ensure minimum competence (Kane, 1982). The conceptual model used to evaluate the task importance weights derived from the job analysis data will consequently differ as well. Figure 2 depicts three logical orderings of mean ratings for tasks rated on a 3-point frequency scale and a 3-point criticality scale. In these examples, these job analysis task ratings would be used to develop test specifications to fulfill three different purposes: to reflect actual practice, to predict maximum job performance (selection), or to ensure minimum competence (licensure).[2]

Table 1: Actual practice.

Freq.	Crit.
3	N/A
2	N/A
1	N/A

Table 2: Maximum performance.

Freq.	Crit.
3	3
3	2
3	1
2	3
2	2
2	1

Table 3: Minimum competence.

Crit.	Freq.
3	3
3	2
3	1
2	3
2	2
2	1

[2] For the sake of simplicity, 2- and 3-point scales will be used in the examples in this paper. However in practice the use of a 2- or 3-point scale might artificially restrict the variability of the ratings (see discussion below).

The ordering of mean task ratings for test specifications that are to reflect actual practice (Table 1) need only to include the frequency scale ratings. That is, if the purpose of the job analysis is to determine what is done on the job, then knowing what tasks are performed and how frequently they are performed is sufficient information. However, if the purpose of the job analysis is to devise test specifications for a selection test (i.e., predict maximum job performance), then the depiction of actual practice needs to be constrained to include only those tasks that are critical to effective job performance (EEOC, CSC, DOL, DOJ, 1978). In this case, criticality ratings may act as a secondary ordering variable to order tasks within different levels of frequency (Table 2).[3]

In the third situation, the job analysis is being conducted to create test specifications for a licensure examination. Licensure examinations are intended to protect the public by identifying those examinees who lack critical knowledge or skills that are believed to be minimally necessary for safe and effective practice in a profession (Kane, 1982; Smith & Hambleton, 1990). Therefore, the purpose of the examination dictates that criticality be weighted more than frequency in order to protect the public welfare (Rakel, 1979). As opposed to the selection situation depicted in Table 2, the primary ordering variable in a licensure context should be the criticality ratings and the frequency ratings should be used to order tasks within the different levels of criticality (Table 3).

Variability of Ratings

Tables 1, 2, and 3 indicated a logical ordering of scale *means* based upon the purpose of the job analysis. Another important consideration is the variability of individual ratings around these means. Although there is a tendency to conclude that variability in a measure is due to random error, it is important to note that there may be significant within-title variability that causes differences in job analysis ratings (Harvey, 1991). For instance, two individuals with the *same* job title working for two companies (or even within the same company) may perform different tasks or perform the same tasks with a different degree of frequency. Although one might hope that there would less variability with criticality ratings (it seems there should be more rater agreement as to whether ineffective performance of a task could cause harm), there is no way to know whether variability in frequency ratings is due to *error* or *true within-title variability*.

In addition, there are other factors that may influence the variability of ratings such as the number of scale points used (one would expect more variability in a 6-point scale than a 3-point scale solely due to differences in the number of scale points) and the degree of specificity with which the tasks are written for the job analysis survey (the more ambiguously a task is written the more chance that respondents might interpret and, therefore, respond to the task differently). Similarly, the types of terms or phrases used as anchors for these scales can also influence ratings in that scales anchored using *relative* terms rather than *absolute* terms (e.g., "seldom, occasionally, sometimes, often, frequently" as opposed to "yearly, monthly, weekly, daily, hourly") may demonstrate either a positive bias (i.e., everything rated highly) or more variability because there is less agreement among ratings due to the ambiguity of the terminology (e.g., the anchor "often" is less concrete than "daily").[4]

Although it is difficult, if not impossible, to discern the cause of rating variability (i.e., whether it is due to random error, properties of the measurement instrument or true within-title variability), from a

---

[3] The argument could also be made that frequency and criticality should be weighted equally or that criticality be weighted more than frequency for a selection test. This is because there is no agreed upon definition as to what constitutes maximum job performance. This criterion may vary as a function of the job (e.g., criticality may be more important in jobs where harm is a probable result of ineffective performance [e.g., nursing] whereas in other types of jobs [e.g., clerical] the frequency of task performance may be a more important determinant of the successful job performance)

[4] In addition, statistical methods assume that the numbers (e.g., 4 = often) mean the same across raters and across activities (e.g., two *different* activities rated as performed "often" by the same rater, *or* two respondents rating the *same* activity as performed "often" are interpreted to mean that the activities posses *identical levels or magnitudes of the characteristic being rated*). However, raters may use different "internal" metrics when rating an activity using relative scales. Using absolute scales (e.g., performed "daily") rather than relative scales (e.g., performed "often") makes this assumption more tenable.

substantive standpoint, variability of ratings is an important consideration. This is because examinations are designed to be administered to the "typical" examinee within a particular field. Less variability in a task's ratings relative to other task ratings may be an indication that most people agree on the ratings.[5] For example, within-title variability may cause the frequency ratings of a set of tasks to vary considerably (e.g., the task of administering CPR may vary as a function of the age group with which a nurse is most often working) whereas other tasks (e.g., monitor and record vital signs) may cut across settings and therefore result in more observed agreement as to its frequency of occurrence. When testing the typical examinee it makes sense to give more weight to those tasks with which there is less variability (and perhaps more agreement) which may be more indicative of general practice than those with more variability (tasks that perhaps vary as a function of setting, client population, or region). Because the goal is to test for minimum competence across various settings and regions it is logical to use variability to order within the mean ratings of both scales (Table 4, 5, & 6). In these examples, tasks with less variability are ranked higher than those with more variability.

Table 4: Actual practice.

Freq.	Var	Crit.
3	low	N/A
3	med	N/A
3	high	N/A
2	low	N/A
2	med	N/A
2	high	N/A

Table 5: Maximum performance.

Freq.	Var	Crit.	Var
3	low	3	low
3	med	3	med
3	high	3	high
3	low	2	low
3	med	2	med
3	high	2	high

Table 6: Minimum competence.

Crit.	Var	Freq.	Var
3	low	3	low
3	med	3	med
3	high	3	high
3	low	2	low
3	med	2	med
3	high	2	high

Covariability of Ratings

Some researchers advocate allowing the covariability of the rating scales to directly impact a task's estimated importance (Kane, 1989). However, there is no conceptual basis for considering tasks with highly covarying scales as either more or less important than tasks for which the rating scales are independent. Furthermore, the degree to which scales are observed to covary differs as a function of the job and of the tasks within that job (Harvey, 1991). Moreover, whether these differences reflect truly covarying rating scales or simply the measurement properties inherent in those scales is difficult to ascertain. The same measurement artifacts that influence the variability of ratings also strongly affect the covariability of the scales (e.g., number of scale points, the types of anchors used, and the way in which the tasks are written).

The contingency tables below emphasize that the difficulty of logically interpreting covariance within the conceptual framework of job analysis. These tables demonstrate how the covariance between the rating scales is *independent* of the mean ratings. Whether mean levels are similar or dissimilar across scales implies nothing about their covariance. While each of these tables has an identical mean of 1.5 on both scales, the covariance between scales ranges from perfect positive through no covariance to perfect negative covariance. Thus, identical mean ratings on both criticality and frequency scales does not imply a positive covariance. A negative or nonexistent covariance can arise when means are equal across scales. Similarly, a mean criticality rating of 3 can have a strong positive covariance with a mean frequency rating of 1 and vice versa.

[5] Properties of the measurement instrument [e.g., no. of scale points, types of scale anchors] would influence all task ratings. However, ambiguous wording of a task would only influence the ratings for that task. A review by the SMEs of each task and the mean and variance of its ratings should help to identify those tasks that might have received poor ratings due to its wording.

Table 7: Perfect positive covariance.

		Criticality 1	2	Total		Mean
Frequency	1	4	0	4	Frequency	1.5
	2	0	4	4	Criticality	1.5
	Total	4	4			

Table 8: No covariance.

		Criticality 1	2	Total		Mean
Frequency	1	2	2	4	Frequency	1.5
	2	2	2	4	Criticality	1.5
	Total	4	4			

Table 9: Perfect negative covariance.

		Criticality 1	2	Total		Mean
Frequency	1	0	4	4	Frequency	1.5
	2	4	0	4	Criticality	1.5
	Total	4	4			

Furthermore, it can be argued that a *lack* of covariability might be desirable because it could indicate that there is more agreement among ratings on a task. As an extreme example, suppose all raters gave the highest rating for criticality and the lowest rating for frequency, as might be the case with a task such as administering CPR, there would be perfect agreement among ratings, meaning that there would be no variability and, as a result, no covariability.

Some Statistical Considerations:

While there is no conceptual reason for considering tasks with covarying scales as more or less important than tasks for which the rating scales fail to covary, there are statistical benefits to incorporating the covariance between rating scales into the procedures for estimating task importance. From a statistical perspective, if two scales covary then each offers information about the other. This information can be used to *strengthen* the estimates of the mean ratings on either scale. Therefore, while covariability is not included in the conceptual model, it can and should be considered by the statistical procedures used to combine rating scales and estimate task importance.

In addition to providing an ordering of tasks that is congruent with the conceptual model, a statistical combining procedure will ideally have the following attributes. First, raters with response patterns that are more internally consistent should be given more weight in the determination of task importance than those with internally inconsistent ratings. In addition, raters whose responses *across tasks* (rather than for specific tasks) consistently agree with those of other raters (i.e., are externally consistent) should be weighted more than those whose ratings do not agree (e.g., person that does not differentiate between tasks and rates everything as highly critical or frequently performed).

Numerical Example: A Comparison of Two Statistical Models for Combining Rating Data

To demonstrate how the conceptual ordering of tasks based on their means and variances can be used to roughly evaluate the results of statistical models, we compared the Kane procedure (Kane et al. 1989) with a new combining procedure based on a modification to the Mahalanobis distance measure. An overview

of both procedures is given followed by a comparison of their results using the conceptual model as the basis for this discussion.

Overview of the Kane Procedure

Kane, Kingsbury, Colton, and Estes (1989) proposed a method for the combination of frequency and criticality data collected from job analysis surveys to produce overall importance weightings for task statements. The method proposed was a multiplicative model that applied a weighting to the criticality scale. This weight was designed to compensate for differences between the two scales. A primary argument for the support of this weight was "the contributions to be made by criticality and frequency are a matter of judgment rather than an empirical question, it seems clear that the relative contributions of these two variables should not be determined by the properties of the data collection procedures" (Kane et al., 1989). The final determination of the raw importance weight for a task statement was:

$$I_i = C_i^{a} F_i \tag{1}$$

where: I is the importance weight for task statement i, C is the mean criticality for task statement i, F is the mean frequency for task statement i, and a is the weighting coefficient. When it is desired that the criticality and frequency scales receive equal weighting the a coefficient is defined as:

$$a = \sqrt{\sigma^2(\ln \overline{F}_i) / \sigma^2(\ln \overline{C}_i)} \tag{2}$$

where: $\ln \overline{F}_i$ is the natural log of the mean of the frequency scale for task statement i and $\ln \overline{C}_i$ is the natural log of the mean of the criticality scale for task statement i. When it is desirable to have more weight given to the criticality scale so that criticality is weighted k times frequency, the a coefficient is:

$$a = \frac{\left((k-1.0)(\sigma_{(F,C)}) + \sqrt{\left((k-1.0)^2(\sigma_{(F,C)})^2 + 4k\sigma_F^2\sigma_C^2\right)}\right)}{(2)(\sigma_C^2)} \tag{3}$$

where: k is the weighting factor, $\sigma_{(F,C)}$ is the covariance of natural logs of the means of the frequency and criticality scales, σ_F^2 is the variance of the natural log of the means of the frequency scale, and σ_C^2 is the variance of the natural log of the means of the criticality scale. It can be seen from these formulae that the Kane weighting procedure utilizes the means of the rating scales for each task statement and the variability and covariability of the means of the task statements.

Overview of the Mahalanobis Distance

The Mahalanobis distance (MD) is a multivariate statistic used to determine distances between objects or task statements in a multivariate space. We have modified the Mahalanobis formula to reflect distances from the origin (0,0). This distance is defined as:

$$MD_j = \sqrt{(R_j - \underline{0})\Sigma_i^{-1}(R_j - \underline{0})'} \tag{4}$$

where: $R_i = \begin{pmatrix} \overline{F}_i \\ \overline{C}_i \end{pmatrix}$ is the mean vector, $\left(R_i - \underline{0}\right)$ is the vector difference between $(\overline{F}_i, \overline{C}_i)$ point and the origin, and Σ is the sample variance-covariance matrix for each task statement *i*.

The use of the MD appears to be appropriate because rating scale values from two or more scales are multivariate in nature. A natural consideration for the analysis of such data are methods which combine the multiple scales (vectors) into some scalar value is the multivariate analog of the univariate z-score or standardized distance from zero. In the case of multiple scales, the origin or zero vector represents the lowest combination of ratings possible. Thus, scale vector means can be measured from this lowest point, and this "distance" represents the strength of that task statement from the lowest possible scale values. The distance, however, is "standardized" in the multivariate sense, by dividing it by the inverse of the covariance matrix of the scales for each task statement. The mean vector is then given more weight if the variance of its scales is smaller and the covariance is near zero. Modifications of this are possible (e.g., setting the covariances to zero automatically or using the absolute value of the covariance).

To give weight to one scale more than the other the mean vector, R, can be premultiplied by a weighting factor. This weighting factor, α_j, where for example $\alpha_1 = 1$ and $\alpha_2 = 2$, indicating that the criticality scale is to receive twice the weight of the frequency scale. Then,

$$MD_i = \sqrt{(\alpha_1 \overline{F}_i \alpha_2 \overline{C}_i)\Sigma_i^{-1}(\alpha_1 \overline{F}_i \alpha_2 \overline{C}_i)'} \quad , \tag{5}$$

for any nonzero value for α_j.

Methods

A small data set of five task statements and ten respondents served as an example of data to be analyzed by the Kane procedure and the MD. These data are presented in Table 10 with summary statistics for each scale and task statement. The were analyzed using four different weights assigned to the criticality scale. These weights, were 1, 2, 5, and 10. The first two weights were selected as common weights that are typically applied in job analyses, while the latter two were selected to demonstrate the impact of higher criticality weights on the task statement weights produced by each method. Each method produced raw weights which were then normed to provide task statement weights that sum to a value of 1.0.

Table 10 - Sample data and summary statistics.

	Task Statements									
	1		2		3		4		5	
Respondent	Freq*	Crit**	Freq	Crit	Freq	Crit	Freq	Crit	Freq	Crit
1	5	4	2	1	3	4	5	1	4	3
2	5	5	2	2	3	2	4	2	1	4
3	4	4	3	1	4	3	4	1	5	2
4	4	5	1	2	4	2	5	2	3	1
5	4	5	1	1	2	2	3	1	1	4
6	5	4	2	1	3	3	4	1	1	3
7	5	4	1	2	3	4	3	2	2	5
8	3	5	1	3	4	2	4	3	5	1
9	4	4	1	2	4	3	5	2	4	3
10	5	5	2	1	3	3	4	1	3	3
Mean	4.400	4.500	1.600	1.600	3.300	2.800	4.100	1.600	2.900	2.900
SD	0.699	0.527	0.699	0.699	0.675	0.789	0.738	0.699	1.595	1.287

* Freq = Frequency
** Crit = Criticality

Results

Table 11A-D present the results of the Kane procedure and MD calculations utilizing various weightings on the criticality scale. Table 11A presents results when the scales were weighted equally, that is a ratio of 1:1, criticality to frequency. Tables 11B, 11C, and 11D present results when the scales were weighted 2:1, 5:1, and 10:1, respectively. In Table 11A, the results for each method demonstrated the same ranking of task statement (TS) 1 and TS 3, but differed on the ranking of the remaining task statements. The normed weights for each method were considerably different. These differences appear to be due to how the data were incorporated into each method. The results presented in Tables 11A - 11D demonstrated how the weights for both methods were impacted by changing the weight on the criticality scale.

Table 11A - Criticality weighted 1:1.

TS	Raw Kane Weight	Normed Kane Weight	Rank	Raw MD	Normed MD	Rank
1	17.327	0.411	1	12.625	0.363	1
2	2.455	0.058	5	5.060	0.146	4
3	8.433	0.200	2	6.299	0.181	2
4	6.292	0.149	4	5.846	0.168	3
5	7.652	0.182	3	4.908	0.141	5

Table 11B - Criticality weighted 2:1.

TS	Raw Kane Weight	Normed Kane Weight	Rank	Raw MD	Normed MD	Rank
1	45.319	0.512	1	20.871	0.400	1
2	3.316	0.037	5	7.697	0.147	4
3	16.287	0.184	2	8.981	0.172	2
4	8.497	0.096	4	6.913	0.132	5
5	15.113	0.171	3	7.758	0.149	3

Table 11C - Criticality weighted 5:1.

TS	Raw Kane Weight	Normed Kane Weight	Rank	Raw MD	Normed MD	Rank
1	550.488	0.729	1	47.186	0.425	1
2	7.236	0.010	5	16.023	0.144	4
3	89.996	0.119	2	18.865	0.170	2
4	18.542	0.025	4	12.327	0.111	5
5	88.519	0.117	3	16.591	0.149	3

Table 11D - Criticality weighted 10:1.

TS	Raw Kane Weight	Normed Kane Weight	Rank	Raw MD	Normed MD	Rank
1	25348.491	0.905	1	91.755	0.431	1
2	23.946	0.001	5	30.128	0.142	4
3	1238.146	0.044	3	36.361	0.171	2
4	61.360	0.002	4	23.164	0.109	5
5	1331.654	0.048	2	31.458	0.148	3

For the Kane procedure the normed weights were greatly affected by the change in the weight of the criticality scale. As the criticality weight was increased, the normed Kane weight for TS 1 increases from 0.411 when the scales were equally weighted case to 0.905 when the criticality scale was weighted 10:1 over the frequency scale. Changes were also noted in the MD weights, but these changes were far less remarkable. As the criticality scale weighting was increased, the normed weights for the individual task statements changed in smaller increments than were observed for the Kane weights. The normed MD weights appeared to be changing toward a limit for each task statement as the weight for the criticality scale was increased. To further explore this notion, an additional weighting scheme was explored for the MD. This was to weight the criticality scale with a value of 1.0 and the frequency scale with a value of 0.0, thereby removing the frequency scale from the MD. Table 12 presents the results for this weighting scheme.

Table 12 - Mahalanobis distances with the frequency scale receiving no weight.

TS	Raw MD	Normed MD	Rank
1	8.955	0.434	1
2	2.836	0.137	4
3	3.562	0.173	2
4	2.297	0.111	5
5	2.983	0.145	3

In comparing the results from Tables 11A-11D and Table 12 it can be seen that as the criticality weight was increased, values for the normed MD moved toward the normed weights presented in Table 12. The Kane procedure is formulated to only respond to changes in weighting in the criticality scale. To simulate the frequency scale having zero weight, the criticality scale was given an extremely high value. The impact on the normed Kane weights was to give all of the weight to the first task statement. This result was consistent with the trend of the results presented in Tables 11A - 11D.

The Kane procedure for assigning task statement weights clearly states that one of its goals is to equalize the contribution of each scale or to provide for additional weighting if desired. Both procedures accomplish this goal but with different methods. The MD uses the raw data differently from the Kane procedure. Each procedure considers variability and covariability at two different levels: MD at the respondent level and Kane at the survey level. The MD uses information pertaining to the variability and covariability of the individual respondents ratings of each task statement individually, while the Kane procedure considers the variability and covariability of the natural logs of the task statement means. It is

illustrative to view plots of the raw data to begin to get a sense for how the data was incorporated by the MD. Figures 1 - 5 (located at the end of the paper) provide bivariate plots of the respondents data and mean ratings for each of the five task statements. In comparing these plots to the results presented in Table 11A, it can be seen that task statements with lower variability in the ratings assigned to a particular scale are given more weight. The primary example of this is represented by TS 2 and TS 5. The MD gives more weight to TS 2 than TS 5. The average ratings of TS 5 are higher than TS 2 but the variability and covariability of the ratings associated with TS 5 are larger than those observed in TS 2. This result is problematical. Although the MD compensated for the variability in the ratings of TS 2, this method provided too much weighting for TS 2.

Discussion

The Kane weights produced orderings of task statements that did not meet the expected ordering based on the model, except when criticality was weighted by a factor of 10. The ordering of task statements in the first three weighting schemes, TS 3 had a higher weight than TS 5. Based on our model we would expect to TS 5 to be weighted more than TS 2 because TS 5 had a larger criticality mean.

However, as the weight for criticality is continually increased the *ordering* does not change for the Kane procedure but the *weights* do change. The MD weights, as stated earlier, approach a limit as the criticality weight is increased. However, increasing the weight of criticality with the Kane procedure causes all of the Kane weight to be given to TS 1 (i.e., a weight of 1.0) whereas all the other task statements receive zero weights. This is problematic. Although the conceptual model we have proposed gives a logical rank ordering of tasks, it does not provide a means to translate this ordinal information into reasonable task weights. Although the statistical model may be congruent with this conceptual model in terms of ordering the task statements, the task *weights* must also be reviewed to ensure they are reasonable. Ideally, a conceptual model would also include a rationale for evaluating a weighting scheme.

We also reviewed Kane weights and MD applied to actual job analysis data. Preliminary analyses demonstrated that each method failed to produce weights that were compatible with the model presented in this paper. Future research is being planned to develop a more comprehensive conceptual model (e.g., include method to evaluate weighting scheme) and a statistical model that would produce weights that are compatible with this conceptual model. Obviously, more work needs to be done; however, we hope that approaching this problem *first logically* and *then statistically* is a step in the right direction.

REFERENCES

Ebel, R. L. (1977). Comments on some problems of employment testing. *Personnel Psychology, 30,* 55-68.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43(166),* 38290-38315.

Harvey, R. J. (1991). Job analysis. In M. Dunnette and L. Hough (Eds.), *Handbook of industrial and organizational psychology (2ed.).* Palo Alto: Consulting Psychologists Press.

Kane, M. T. (1982). The validity of licensure examinations. *American Psychologist, 37(8),* 911-918.

Kane, M. T., Kingsbury, C., Colton, D., & Estes, C. A. (1989). Combining data on criticality and frequency in developing plans for licensure and certification examinations. *Journal of Educational Measurement, 26,* 17-27.

Lunz, M. E., Stahl, J. A. & James, K. (1989). Content validity revisited: transforming job analysis data into test specifications. *Evaluation and the Health Professions, 12,* 192-206.

Rakel, R. E. (1979). Defining competence in specialty practice: The needs for relevance. In *Definitions of competence in specialties of medicine, conference proceedings.* Chicago: American Board of Medical Specialties.

Raymond, M. R. (in press). Establishing weights for test plans for licensure and certification examinations. *Applied Measurement in Education.*

Sanchez, J. I. & Frazer, S. L. (1992). On the choice of scales for task analysis. *Journal of Applied Psychology. 77,* 545-553.

Sanchez, J. I. & Levine, E. L. (1989). Determining important tasks within jobs: a policy capturing approach. *Journal of Applied Psychology, 74,* 736-342.

Smith, I. L. & Hambleton, R. K. (1990). Content validity studies of licensing examinations. *Educational Measurement: Issues and Practice, 9(4),* 7-10.

Figure 1 - Plot of Item 1 Respondent Ratings

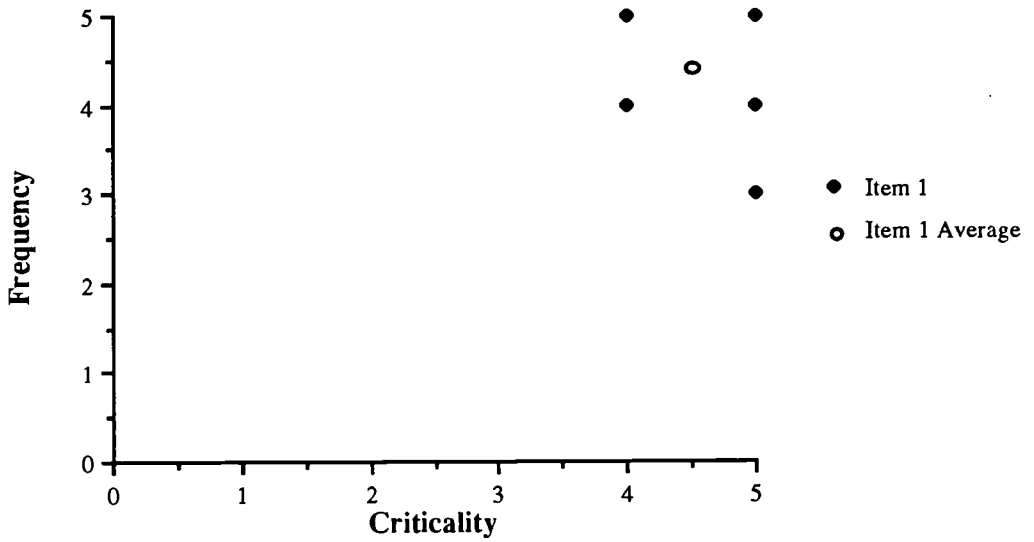


Figure 2 - Plot of Item 2 Respondent Ratings

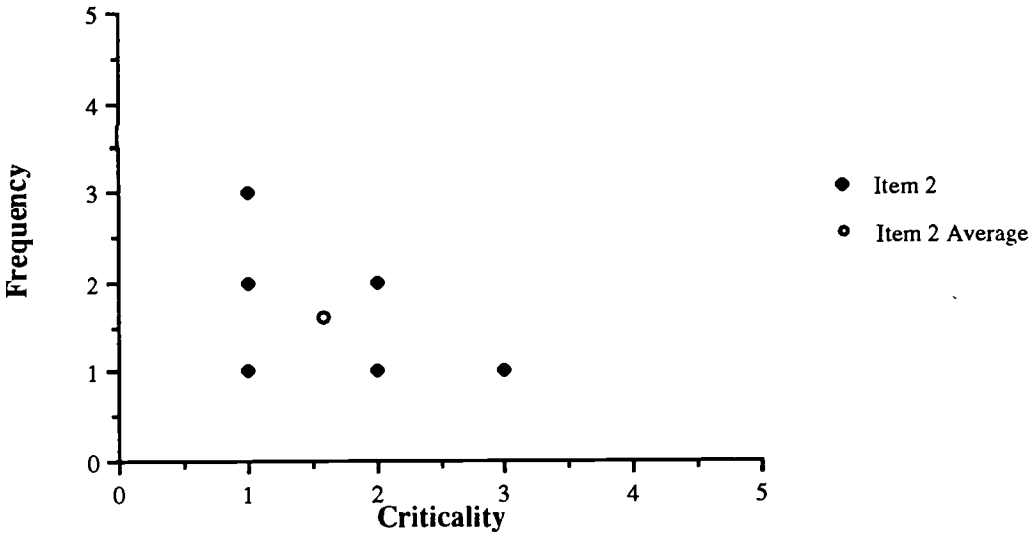


Figure 3 - Plot of Item 3 Respondent Ratings

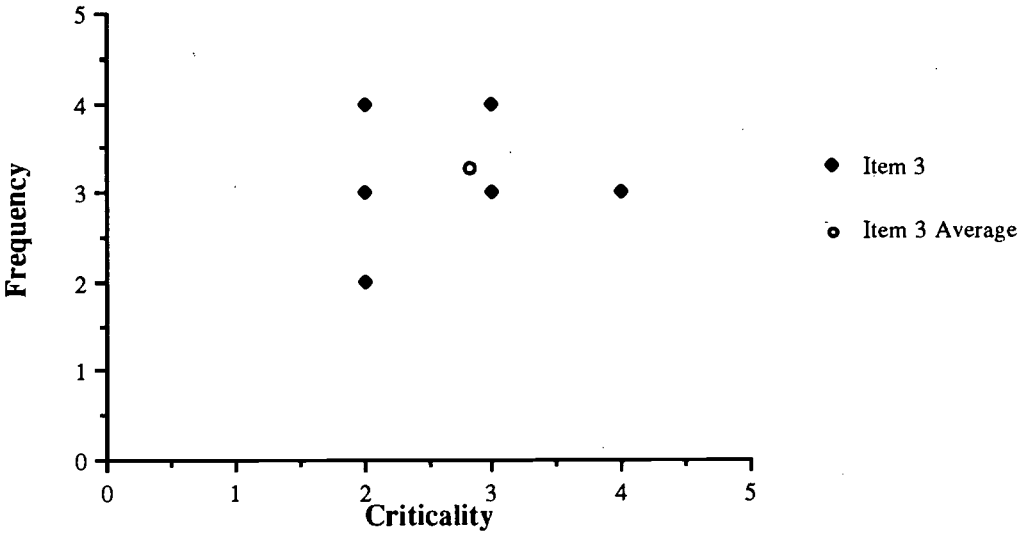


Figure 4 - Plot of Item 4 Respondent Ratings

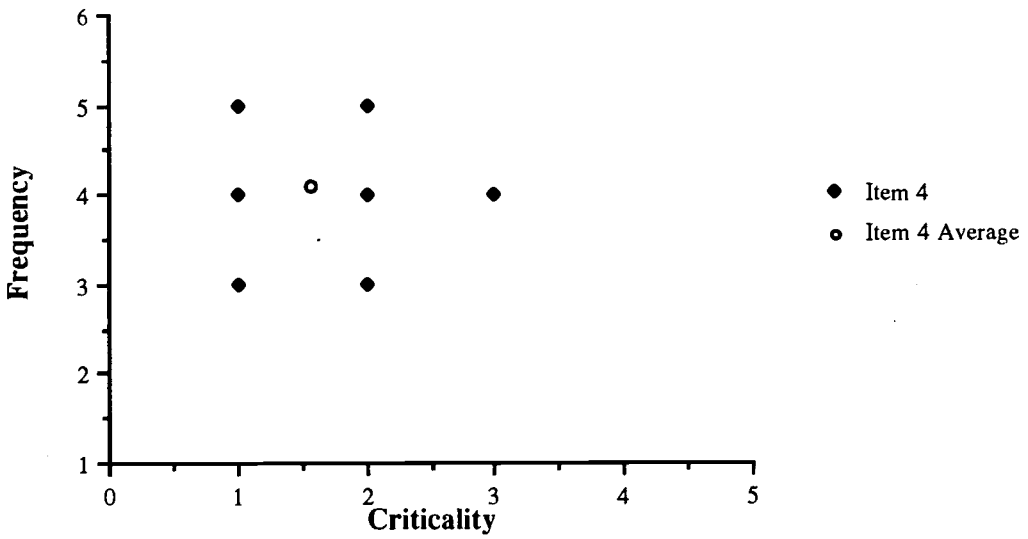
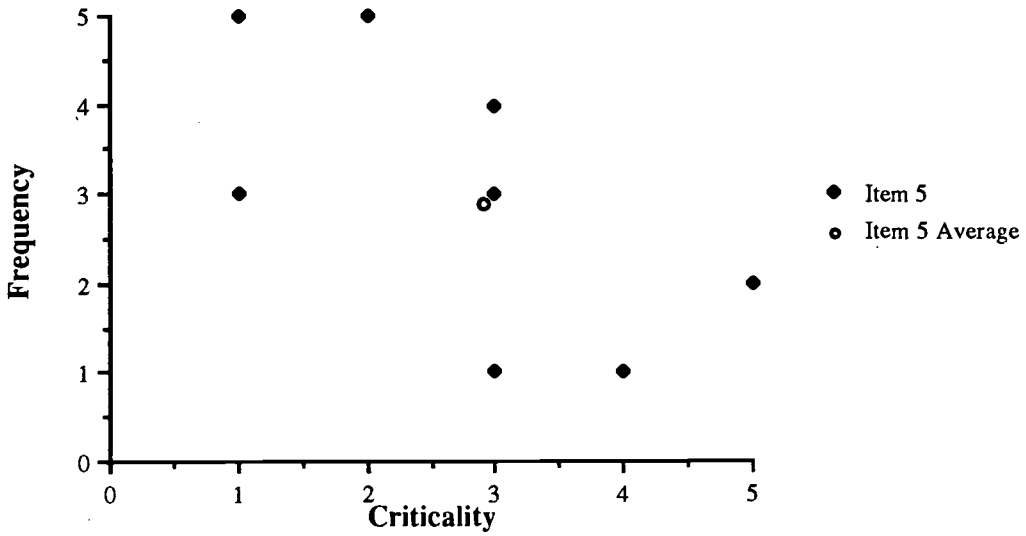


Figure 5 - Plot of Item 5 Respondent Ratings



TM025855 AERA April 8-12, 1996

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Weighting Tasks from Job Analysis Data to Derive Test Specifications for Licensure Examinations: Some Methodological & Statistical Considerations	
Author(s): Leslie Thomas and John C. Kalohn	
Corporate Source:	Publication Date: April 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☒ ← **Sample sticker to be affixed to document**

Check here

Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document → ☐

or here

Permitting reproduction in other than paper copy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Leslie Thomas	Position: Research Psychologist
Printed Name: Leslie Thomas	Organization: American College Testing
Address: American College Testing P.O. Box 168 Iowa City, IA 52243	Telephone Number: (319) 337-1137
	Date: 5/15/96

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

February 27, 1996

Dear AERA Presenter,

Congratulations on being a presenter at AERA[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a written copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (23)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

AERA 1996/ERIC Acquisitions
The Catholic University of America
O'Boyle Hall, Room 210
Washington, DC 20064

This year ERIC/AE is making a **Searchable Conference Program** available on the AERA web page (http://tikkun.ed.asu.edu/aera/). Check it out!

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

---

[1]If you are an AERA chair or discussant, please save this form for future use.