WHEN IS A T AND E RATING VALID?

James C. Johnson
William L. Guffey
Robert A. Perry

Division of Research
Department of Personnel
State of Tennessee

Presented at the Annual Meeting of the International Personnel Management Association Assessment Council, Boston, July 6-10, 1980

This project was supported in part by a grant provided by the Atlanta Regional Office of the United States Office of Personnel Management under the Intergovernmental Act of 1979

The rating of training and experience as a means of examining applicants for employment has a long history in merit systems. But it is an assessment method which is unique to public organizations. While private sector organizations often establish minimal training and experience criteria for employment, I doubt that a private sector organization can be found which scores applicant training and experience in the manner of a test as occurs in the public sector. Descriptions of T and E rating techniques will not be found in the standard texts or graduate training programs for personnel psychologists, and the limited research evidence which exists on the subject is found almost exclusively in unpublished governmental reports.

I suspect that the reasons we use the methods, and the problems we encounter in doing so, are well known to civil service examiners. T and E rating systems are used in virtually all merit systems. We must use them, largely because we have few options. In Tennessee, for example, about 2,000 different examinations are used to select employees for a corresponding number of competitive job classes. Of these 2,000 examinations, only about 180, or slightly under ten percent, require use of paper and pencil tests or other objective assessment methods. On the other hand, a majority of applicants are assessed by means of paper and pencil tests. A similar pattern of usage seems to exist in other jurisdictions according to Perry's (1980) survey last year of major state, county, and municipal merit systems. We have thus adopted an efficient strategy to meet the demand for massive examination systems: when we have relatively large numbers of applicants to be examined and positions to be filled, we tend to use paper and pencil tests which are costly to develop but inexpensive to administer. When we have relatively fewer applicants to be examined and positions to be filled we tend to use relatively simple T and E rating schemes which are inexpensive to develop though usually more costly to administer and score than paper and pencil tests.

Most jurisdictions could not possibly develop sound written tests or similar objective examining strategies to replace T and E ratings because of a lack of resources to do so. Moreover, it is sometimes though not always true that T and E ratings produce less adverse impact on legally protected groups than do other examining methods, resulting in a lesser perceived need for validity evidence in accord with EEO legislation and hence a lesser cost for validation activities.

We anticipate, however, that this solution to our administrative, legal, and limited resource problems will in turn create another problem for us because of a changing societal environment. Applicant complaints and legal challenges in some areas are now being mounted as frequently against T and E examinations as they are against paper and pencil tests. In a Tennessee sex discrimination case, the plaintiff even argued quite persuasively that a written test would evaluate candidates far more directly then the T and E examination used for the position in question. That one was settled out of court. Also increasing, in our view, is the demand from employing agencies that examinations be useful, that they be valid, that the examining process identify available applicants who are most likely to be effective in the performance of their jobs. Governmental organizations are increasingly faced with the demand that they enhance efficiency and productivity. While employee selection is but one part of the problem of productivity, it is an important part as evidence of test utility continues to demonstrate. To the extent that we use civil service examining methods which are not valid, which are not themselves cost-effective, and which do not contribute to the productivity of the governmental agencies we serve, they are likely to be eliminated altogether in the long run in some jurisdictions, or replaced by ineffectual but less costly methods in others. These may be appropriate consequences unless we can devise better examining procedures to replace those which may be inadequate.

In the 1970s we appropriately focused most of our attention on paper and pencil tests, conducting an enormous amount of research on both the validities of specific tests and methods used to develop them. Though that effort remains incomplete, it seems to me that we are over the major technical hurdles and can demonstrate the value of valid examining procedures to employing organizations. T and E ratings present a special problem. We lack an adequate psychometric theory to guide us, in addition to research evidence on alternative methods an empirical validities. We must continue to use them, however, even for selection of professional and managerial employees who are critical to the functioning of employing organization, because we lack the resources to implement larger numbers of paper and pencil tests, work sample tests, assessment centers, and so on.

The lack of evidence to support either the validities of T and E rating schemes, or to support assumptions underlying use of T and E ratings, has been noted by numerous authors in recent years. In one of the most comprehensive reviews of T and E ratings methods of which we are aware, Beardsley (1976) in Pennsylvania describes the state of affairs she found in 1976 as follows:

> "A formal literature review was conducted to find information (especially empirical studies) that was specifically about E and E's. Very little information was found. For this reason, this report does not have a 'review of the literature'."

The existing published research evidence on the criterion validities of T and E ratings is, in general, not even mixed: studies we were able to locate concluded almost unanimously that statistically significant validity coefficients are rare, and most frequently they are zero. There is similarly little evidence that the basic models underlying most T and E rating systems are appropriate, and evidence which exists suggests that these assumptions are, in fact, often incorrect.

There are several assumptions or theories presumed to justify T and E rating schemes. Porter, Levine, and Flory (1976) suggest that T and E evaluation systems are based on two basic assumptions. One is that past performance is the best predicator of future performance. It is logical to assume that information about relevant training and work experience is related to performance of a job, and therefore predictive of job performance. Their second assumption is that as an individual gains more experience in an occupation, he or she demonstrates greater commitment to it, and is thus "more likely to wish to pursue, perform well in, and gain advancement in the occupation." (Porter, Levine, and Flory, 1976; p l).

Beardsley (1976) notes, in addition to general assumption that past performance is the best predictor of future performance, several specific assumptions attributed to Yost (1967):

1. Training and experience directly pertinent to the job is more predictive of success than training and experience which is less pertinent. Some training and experience is so wholly unrelated as to have no predictive value whatever.

2. Training and experience which is progressive is more valuable than the same amount on the same job or in jobs of decreasing responsibility. Here the question to be answered is whether a candidate has ten years of experience or one year of experience ten times.

3. Recent training and experience is more valuable than training and experience of the same type which is not so recent.

4. More responsible training and experience is more predictive of success than less responsible training and experience, assuming that the previous job is not more responsible than the position for which application is made.

5. The probability of success increases with the mere aggregate length of training and experience.

6. There is a maximum of experience beyond which no increase in competence is either required or demonstrated.

7. In theory, a corresponding hypothesis holds for training - there is a maximum of training beyond which no increase in job performance is likely to result and there may be limit beyond which increased training actually indicates reduced probability of success.

The authors of a technical section of an Exam Preparation Manual for the U. S. Civil Service Commission dated June, 1977 have identified yet another kind of assumption underlying use of T and E ratings.

1. Amount and quality of education and experience are indirect indicants of knowledges, skills, abilities, and other characteristics (KSAO's). Education and experience are correlated with KSAO's, since they are among the causes of KSAO's.

2. KSAO's are correlated with job performance.

But these authors point up disconcerting implications of these two assumptions with respect to the expected predictive validities of a T and E rating scheme.

$$r_{t\&e.jp} = r_{t\&e.ksao} \times r_{ksao.jp}$$

$$r_{t\&e.ksao} \leq .40$$

$$r_{ksao.jp} \leq .50$$

$$r_{t\&e.jp} \leq (40)(.50) \leq .20$$

The correlation of a T and E rating and job performance is equal to the product of the correlation of the rating and KSAOs, and the correlation of the KSAOs and criteria of job performance. KSAOs are not likely to correlate higher than about .50 with measures of job performance. A T and E rating is not likely to correlate higher than .40 with the KSAOs (and we think that is a generous estimate). They conclude, then, that the maximum validity of a T and E rating scheme is (.50 x .40), or .20. Their conclusion is clearly consistent with the research evidence suggesting that validity coefficients for most T and E ratings are zero.

Faced with the need to overhaul the T and E rating procedure in Tennessee two years ago, we initiated a project to develop what we hoped would be "state of the art" procedures for use by our examiners. That objective faded rapidly when it became apparent that no one has yet demonstrated that relatively simple T and E ratings are valid, nor could we find models to clearly establish a content-validity rational for T and E ratings. Our objectives there become, first, to develop more detailed models to guide us, and then to study psychometric properties of rating schemes derived from the models. The major focus of this paper is a description of these models and the limited data we have accumulated to date about them, as well as our tentative conclusions.

The directions we have taken is based on a specific view of what validity of an employment test means, and in particular how content validity must be defined in an employment setting. Tenopyr (1974) has pointed out that the fundamental rational of an employment test must rest on predictive validity - the extent to which scores on a test predict performance on the job at a later point in time. Our concepts of validity with respect to T and E measures derive from this view, and define content validity as the extent to which an examination samples behaviors which are predictive of job performance. To be content valid, then, a T and E rating scheme must get at those elements of past behavior which will be predictive of future behavior. Face validity is insufficient. Content validation strategies must be supported by evidence that the methods used to develop the T and E rating scheme produce an examination which possesses predictive criterion validity, even though it is typically impossible to evaluate the criterion validity of each individual examination. In our view, criterion validity evidence for each specific examination is unnecessary. It is the method of developing an examination, rather than the examination itself, which must be subjected to empirical "validation" research (Johnson, 1978).

Consider, first, the problems inherent in what Beardsley (1976) and others have termed the "traditional" rating scheme. The traditional methods are probably more common than any of the others, although there are variations in detail among the various traditional schemes in use by different organizations. In a traditional rating, potentially relevant indicators of training and work experience are identified and assigned a relevance level.

"RATING OF EDUCATION AND EXPERIENCE"

Position _____

College

1	2	3	4	5	6	7	8		Totals
20	24	30	36	40	40	40	40		
Degrees	Associate Degree		Bachelors Degree	1 Year Masters	2 Year Masters	DDS or JD	MD or PhD		
	1		2	3	4	5	6		

High School	1	2	3	4	Diploma
	10	10	10	10	1

	19	19	19	19		
A	5-10-15-20	6-12-18-24	7-14-21-28	7-14-21-28		
B	3- 6 -9 -12	4- 8- 12- 16	5-10-15-20	5-10-15-20		
C	2- 4- 6- 8	3- 6- 9- 12	3- 6- 9- 12	3- 6- 9- 12		

	19	19	19	19	19	19	19	19	19	19	19	
A	8-16-24-32	8-16-24-32	8-16-24-32	9-18-27-36	9-18-27-36	9-18-27-36	9-18-27-36	10-20-30-40	10-20-30-40	10-20-30-40	10-20-30-40	
B	6-12-18-24	6-12-18-24	6-12-18-24	7-14-21-28	7-14-21-28	7-14-21-28	7-14-21-28	8-16-24-32	8-16-24-32	8-16-24-32	8-16-24-32	
C	4- 8-12-16	4- 8-12-16	4- 8-12-16	5-10-15-20	5-10-15-20	5-10-15-20	5-10-15-20	6-12-18-24	6-12-18-24	6-12-18-24	6-12-18-24	

Examiner _____ Date _____ Final Score:

Figure 1. Tennessee's traditional rating guide

Tennessee's basic rating guide is about as simple as any which we have encountered (Figure 1). Three levels of relevance, designated "A," "B," or "C" are reflected in the scoring of past work experience. The assumption is made that recent experience is associated with higher competence, and this assumption is reflected in the rating guide by assigning more points to recent than to earlier experience. Regardless of the class for which the examination is being conducted, the number of points assigned for each year of work experience, or year of education, is fixed; the only difference in the structure of these examinations for different jobs, then, is the specific work experience which is creditable, and the assigned relevance levels of creditable work experience.

The fundamental assumption underlying a traditional rating guide is that exposure to the opportunity to acquire job-related competencies is associated with the acquisition of these competencies. Each score element of the guide, which in this case is each three months of work experience, and each period of time of education, contributes to the total score of the examination. As in the case of a paper and pencil test, then, the validity of the examination scores depend on the validities of the individual scored elements or what we can term indicants (Figure 2). TRADITIONAL MODEL

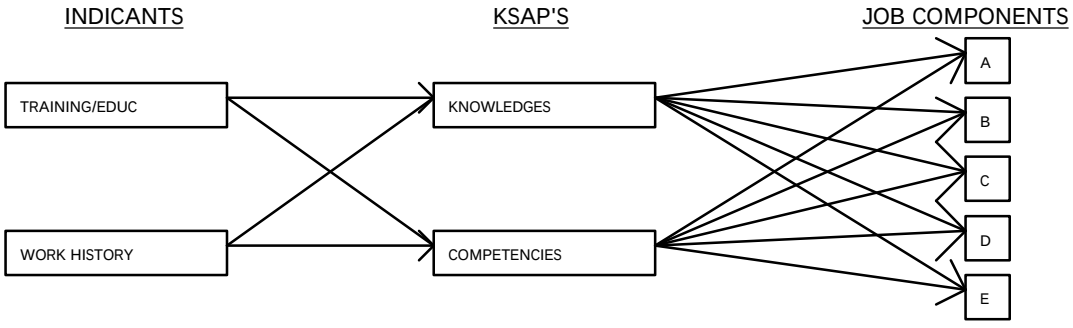

Figure 2. Linkages of indicants to KSAPs and job components in the traditional model.

To the extent that any indicant fails to contributed to the validity of the examination by distinguishing between more competent ad less competent applicants, the indicant is contributing only error variance to the total score. This premise is illustrated in Figure 3, in which the relationship between indicants and performance of a job or competence is illustrated for three different indicants, one possessing no validity, a second possessing moderate validity, and a third possessing a great deal of validity. In practice, of course, we are rarely able to conduct an empirical study to determine whether these relationships between individual indicants and competence exist. The rational test of the validity of an indicant is whether there is reason to assume, on the basis of expert judgment and prior experience with employees, that these relationships exist. If it is not rational to assume that two years of college training will result in less competence than three years of college training, then this difference in training should <u>not</u> produce a difference in examination score.

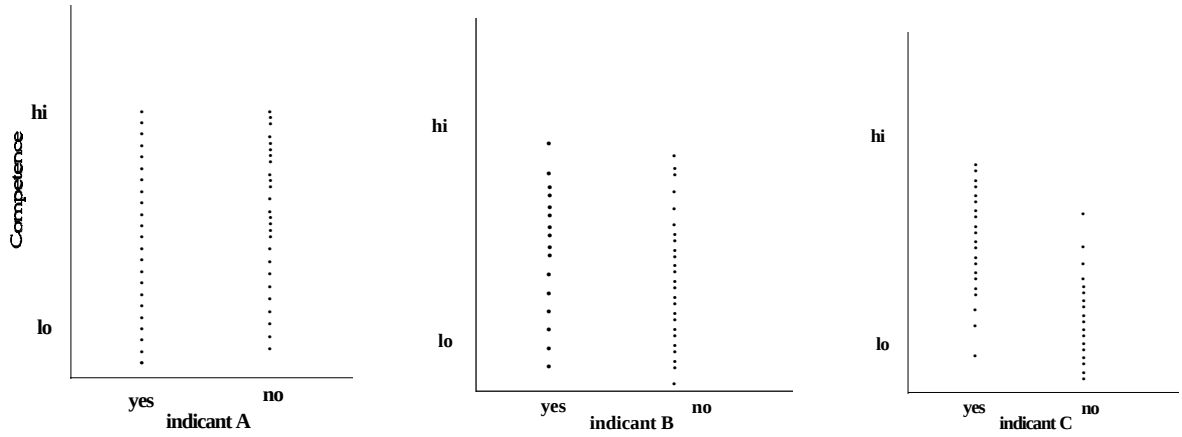

Figure 3. Indicant A is not valid; indicant B is moderately valid; indicant C is highly valid.

The same logic is appropriately applied to indicants which are more specific than is typically the case in a traditional rating. If we compare individuals who have had a course in accounting with individuals who have had no course in accounting with respect to their competence is applying accounting principles, the competence scores of the two groups might reasonably be expected to result in distributions as illustrated in Figure 4. It is reasonable to presume that there will be a substantial difference between the means for the two groups, but note also that there is considerable overlap between the two distributions. To the extent that this overlap exists, of course, the indicant is not a valid predictor of competence in accounting principles. In this illustration, however, the indicant is valid.

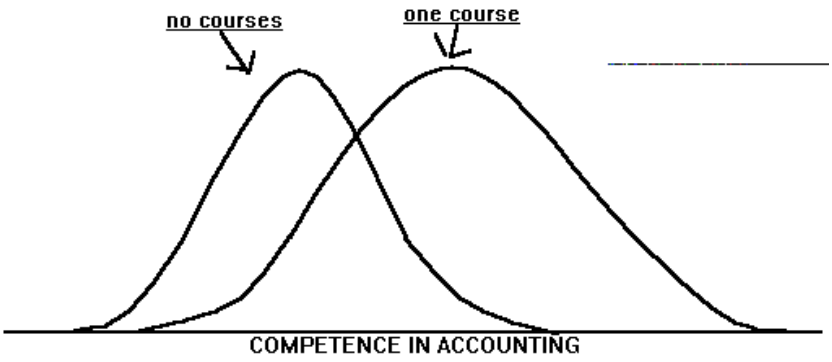

Figure 4. Comparison of persons with no training in accounting and those with one course in accounting with respect to competence in accounting principles.

If we now compare persons who have had twelve hours of accounting with individuals who have had eighteen hours of accounting, our result is more likely to

be as illustrated in Figure 5. In this case, there is substantially greater overlap between these two groups, and in fact, some individuals with eighteen hours of accounting are less competent than the average individual with twelve hours of accounting, and conversely some of those with twelve hours of accounting are more competent than the average person with eighteen hours of accounting. In this case, assigning more points in individuals with a higher number of quarter hours in accounting may add little or no valid variance to the T and E scores.

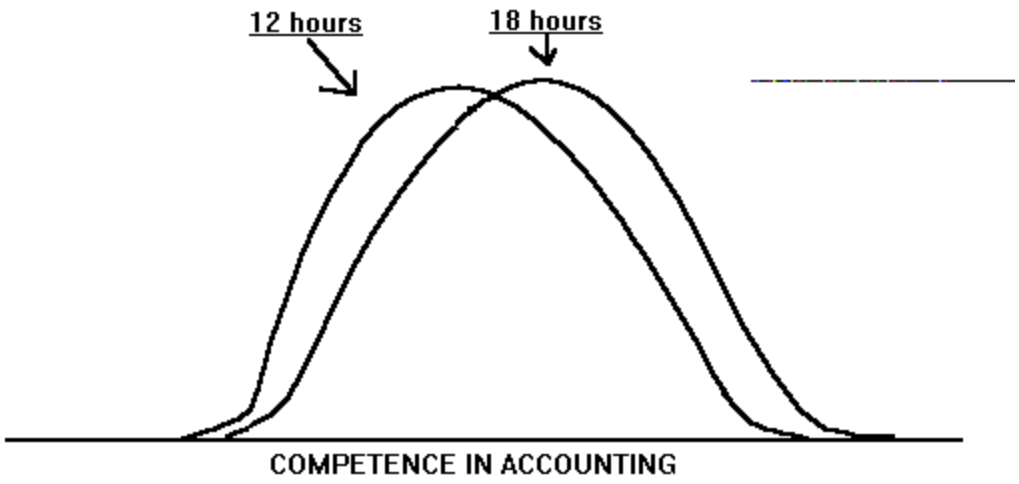

Figure 5. Comparison of persons with 12 and with 18 credit hours in accounting with respect to competence in accounting principles.

There is, of course, a substantial amount of empirical evidence suggesting that the relationships between indicants of academic achievement and job performance are at best modest and often zero. With respect to objective tests of educational achievement, the evidence also supports the contention that there is enormous overlap in score distributions among individuals at differing grade levels.

particularly in high school, college, and graduate programs, again leading to the conclusion that such indicators as number of years of education, or number of credit hours in specific courses or training programs, are likely to possess little or no validity because of their indirect linkage to the competencies of individuals.

The same conclusion can be drawn with respect to the work history segment of a traditional rating guide. The published research evidence suggests that there is generally little or not relationship between the amount of relevant experience an individual has had and job performance. To assign an individual with two years of supervisory experience more points than an individual with one year of supervisory experience is based on an unsupported assumption that the additional year of experience increases the individual's competence. In our studies of the question of the relationships between amount of experience, or amount of training to job performance, we have not yet encountered a significant positive relationship, and we have even encountered instances of negative relationships between length of experience and objective test scores of job performance measures. An example of this phenomenon appears in a study carried out several years ago involving both entry and second-level probation and parole officers (Johnson and Hill, 1976; see Table 1).

The relationships between educational background and alternation rankings provided by supervisors of five job performance domains plus overall performance are essentially zero. Even possession of a college degree, though associated with a test under development to predict performance in these jobs, is essentially uncorrelated with performance. The only educational variable we found to be associated with job performance was overall grade point average, but the magnitudes of these relationships are quite small.

TABLE 1

RELATIONSHIPS BETWEEN ACADEMIC BACKGROUND VARIABLES, TEST SCORES, AND JOB PERFORMANCE RATINGS

Background Variable	Test Score (N=125)	Supervisory Ratings (N = 120)					
		Counseling	Caseload	Communication	Judgment	Interpersonal	Overall
College Degree	27**	13	00	16*	12	-10	02
Criminal Justice Courses	-06	06	16*	15*	17*	12	10
Counseling Courses	-14	-07	-09	-10	-06	-09	-04
Psychology Courses	22**	10	07	13	10	-02	10
Sociology Courses	-08	10	18*	10	05	-10	-11
Graduate Courses	15*	12	02	13	11	04	11
Graduate Degree	12	13	04	06	03	01	08
GAP	27**	15*	12	17*	12	02	18*

Note. Decimals omitted. Frequency distributions for background questions are given in Appendix J.

** p < .01
* p < .05

TABLE 2

RELATIONSHIPS BETWEEN AGE, EXPERIENCE, TEST SCORES, AND JOB PERFORMANCE RATINGS

	Test Score (N=125)	Supervisory Ratings (N = 120)					
		Counseling	Caseload	Communication	Judgment	Interpersonal	Overall
Experience in Counseling	03	-14	-05	-23*	-12	-23*	-11
	(.142)	(-.090)	(-.005)	(-1.58)	(-.085)	(-.195)	(-.065)
Experience in Job Class	00	-10	-06	-11	-05	-13	-03
	(.086)	(-.058)	(-.024)	(-.048)	(-.023)	(-.099)	(-.009)
Experience in Position	01	-16	-06	-17	-10	-18*	-08
	(.116)	(-.111)	(-.019)	(-.104)	(-.062)	(-.144)	(-.041)
Age	-30**	-17	-13	-25*	-12	-15	-15

Note. Decimals omitted. The correlations, partialing out the effects of age, are indicated in parentheses.

** p < .01
* p < .05

Similarly the relationships between age, work experience, and job performance are either zero or negative, even when the effects of age are partialed out. We did not study the T and E rating procedure which had been used for selection of employees to these positions prior to this study, but as the data for another job to be presented shortly will suggest, traditional T and E ratings are strongly associated with these kinds of experience and training variables, and it is entirely possible that the T and E ratings used for probation and parole officers had a negative relationship with subsequent job performance. We were able to establish criterion validities for the 60-item test being studied for use in the selection of probation and parole officers, with validities against the criteria ranging from about .30 to .54 except for the interpersonal relations performance criterion.

With possible exceptions to be noted shortly, we suspect that broad indicants of training and prior work experience are likely to be appropriate in T and E rating schemes. The most plausible relationship likely to emerge between these kinds of indicants of training and experience are illustrated in Figure 6. Though the form of the relationships my differ for different kinds of jobs and different kinds of indicants, they are likely to be curvilinear, and in many instances ascend rapidly and then reach plateaus beyond which little or no increase in competence occurs. Logically, relationships such as these are best used as pass-fail criteria, rather than bases for rank-ordering candidates. In other words, the kinds of indicants typically found in traditional ratings may be more wisely used as minimum qualifications for admission to an examination rather than as scored indicants in the examination itself.

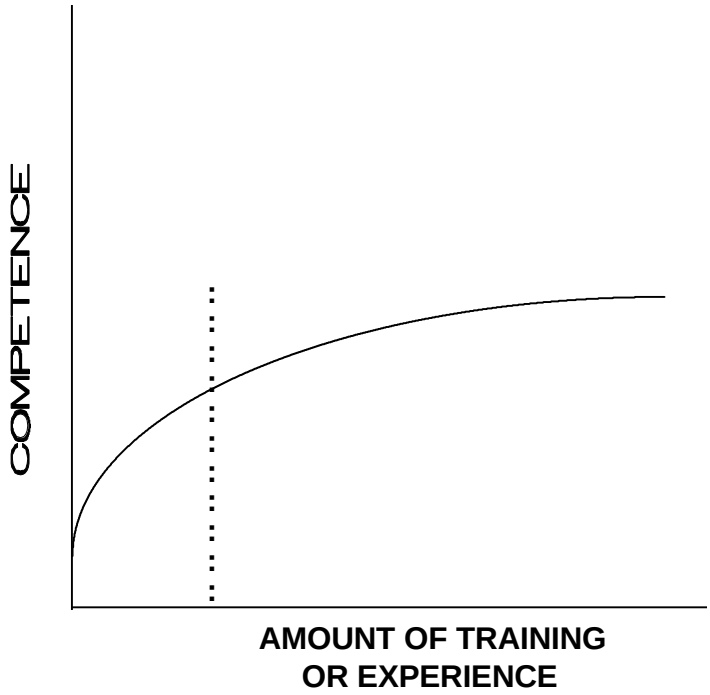

Figure 6. Hypothesized curvilinear relationship between amount of training or experience and job competence.

There are, of course, many other logical questions which emerge with respect to traditional ratings, including the weighting of individual indicants, the question of summing the points assigned to each indicant even if they are presumed to reflect the same attribute and perhaps instead should be substituted for one another, and so on.

Our conclusion with respect to the value of a traditional rating scheme is thus rather gloomy. We find absolutely no evidence for empirical validity, either in our own studies or in the studies of others we have thus far uncovered. But because assessment procedures, including more sophisticated forms of T and E ratings, are

substantially more costly to develop and often more costly to score, we have also considered the question, "when might a traditional rating be useful?" Under what conditions might it be acceptable to use a traditional rating scheme when resources are somewhat limited, even if other assessment procedures might be preferable? We suggest the following:

1. When the variation in competencies of applicants is very large.

2. When the indicants of training or of work experience are clearly and exclusively linked to job performance potential.

3. When a selection ratio (the ratio of applicants to vacant positions) is extremely large, or extremely small.

The rationale for the first condition stems directly from psychometric theory. If applicants are relatively homogeneous with respect to their competencies, a traditional rating scheme lacks sufficient validity, by its nature, to produce valid score differences. While score difference created by one versus two years of experience, for example, are not likely to be valid, nor is the difference between two years of college and three likely to produce discernible job performance differences, there may well be valid differences between persons having no courses in a particular field and those with substantial amount of training in that field, or no experience as compared to several years of experience.

The second condition for use of traditional ratings, that indicants of training or of experience be clearly and exclusively linked to job performance, stems from our consideration noted earlier of the discriminating power of indicants. For some kinds of competencies, there are many different ways in which the knowledges or other attributes can be acquired. In such cases, individual indicants of these competencies will be less valid than if the indicant is more clearly and exclusively the source of acquisition of the competence. For example, if the operation of a piece

of complex equipment can be learned only through completion of a particular training course, or through at least six months of on-the-job experience with the equipment, either of these indicants will be valid in predicting performance on the job. That is, if we compare persons who have the training or the on-the-job experience with those who possess neither the training nor the experience, the job performance of the former group will be clearly superior. On the other hand, if we compare the job performance of managers who have a college degree in business administration with those who have college degrees in other fields, we are unlikely to find much of a difference because of the diverse ways in which persons can acquire the relevant competencies: academic training in business administration is but one of many routes to managerial competence.

Finally, the conclusion that an extreme selection ratio (high or low) might appropriately affect a decision to use a traditional rating scheme is also derived from psychometric theory. When the selection ratio is very large, nearly all candidates will be employed and the examination serves, in fact, no useful purpose in rank-ordering candidates, regardless of the validity of the examination. The practical thing to do, of course, would be to use no examination at all, but where statutes or regulations prohibit such discretion, perhaps use of a traditional T and E rating is the next best thing! When the selection ration is very small, however, as would exist if there were one hundred applicants and two positions to be filled, even an examination with very modest validity will be of value. But even under these conditions, the examination must possess some validity to be of value. Moreover, it is often precisely under these conditions, when we have a relatively large number of

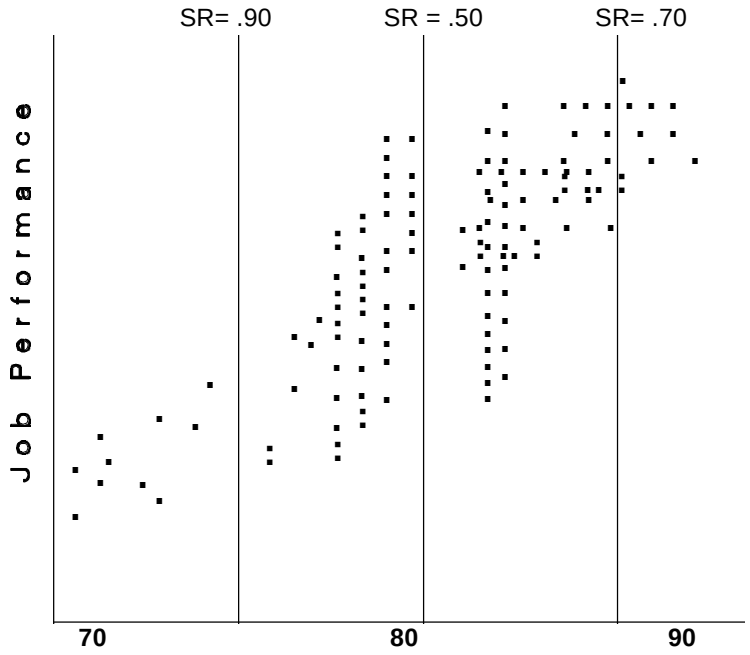

Figure 7. Illustration of the effects of increased selection ratios on the competencies of persons hired for a job, assuming 100 applicants and employment of 10, 50, or 90 persons.

applicants for the number of positions to be filled, that more objective assessment procedures are likely to be more practical.

The traditional model, including variations more sophisticated than illustrated, use primarily job titles or academic course titles, majors, and minors, as predictors of job performance. Another approach, however, is to use a much more explicit description of the past experience or training of applicants than is afforded by means of a traditional scheme. It is usually necessary to design a supplementary application, tailored to the specific information needed to predict performance in a particular job class. We have devised three general "models" to characterize these approaches, one based on tasks, a second on KSAPs, and a third on

behavioral achievements, or what has been referred to as the behavioral consistency model. Each of these approaches will be briefly summarized, and then the results of one of our most extensive studies for one job class reviewed.

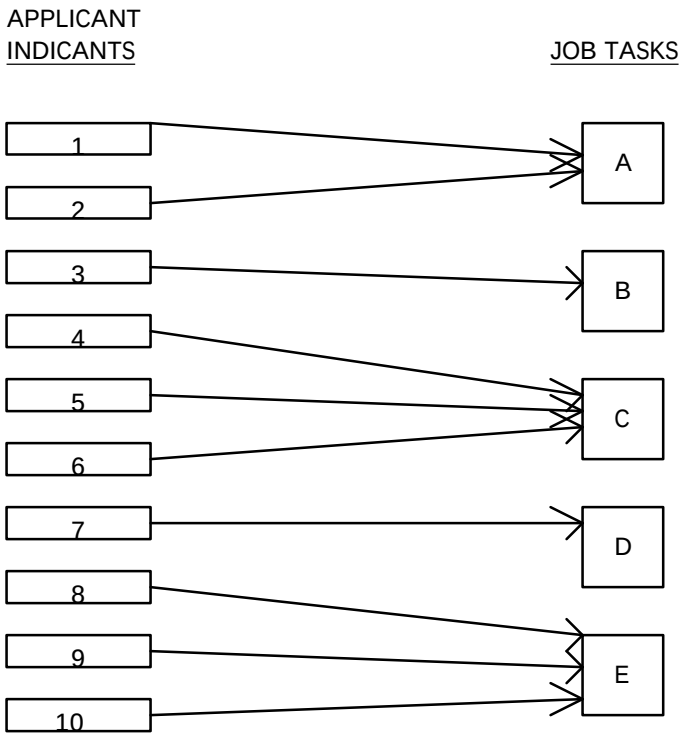

Figure 8. Task-based rating model.

A task-oriented procedure is based on the premise that greater validity can be achieved by obtaining detailed information about specific tasks which an individual has performed in the past, regardless of the job in which the task was performed.

Applicant History		Job Tasks
1. xxx	yes___ no ___	A
2. xxx	yes___ no ___	B
3. xxx	yes___ no ___	C
4. xxx	yes___ no ___	D
5. xxx	yes___ no ___	E
6. xxx	yes___ no ___	F

Figure 9. An application of the task model.

Job tasks, then, rather than job titles with their generalized descriptions, are indicants of past performance presumed to predict future performance. In one version of this model, the supplementary application is essentially an inventory in which individuals are asked to check tasks they have performed in the past.

Have you ever used a pipet to transfer fluids using any of the following techniques?

78. Mouth Suction
YES NO

☐ ☐

79. Bulb Suction
YES NO

☐ ☐

80. Have you ever mixed chemicals together to obtain a needed concentration of a sample?
YES NO

☐ ☐

81. Have you ever filtered a liquid using a vacuum pump?
YES NO

☐ ☐

82. Have you ever performed dilution procedure on specimen samples?
YES NO

☐ ☐

83. Have you ever injected substances under the skin of patients?
YES NO

☐ ☐

84. Did you fill out this application without any help?
YES NO

☐ ☐

I certify that all of the information given herein is true, complete, and correct to the best of my knowledge and belief and is given in good faith. I understand that if I knowingly make any misstatement of facts, I am subject to disqualification or dismissal and to such other penalties as may be prescribed by law or by the Department of Personnel Regulations.

Signature _____Date _____
Unsigned applications will not be processed

Figure 10. Illustration of task-based inventory.

They might also be asked, however, the amount of time they were responsible for performing the task, and the setting or job in which performance of the task occurred (for verification purposes). An example of one page from a different version of a task-based supplementary application, used for experimental purposes among applicants in a social services title in our public welfare agency, called Senior Eligibility Counselor, is illustrated.

I. INTERVIEWING

Senior Eligibility Counselors must schedule and interview a wide range of people (irate, talkative, uneducated) making application for social benefits. This involves controlling the course of the interview to maintain their schedule, gathering all relevant information from the client needed to determine eligibility and establish a budget of benefits, counseling and referring clients to other service agencies for additional benefits, and observing client's children for possible cases of neglect or abuse.

In the space provided below describe your interviewing experience. Describe in detail all relevant tasks you performed, your level of responsibility, and situations or problems you dealt with.

Job #'s _____ # Months Responsible _____

Who can verify this employment?
Name: _____
Address: _____
_____
Phone No. _____

Figure 11. Illustration of an alternative task-based method.

Lists of potentially relevant tasks are developed through meetings with subject matter experts who identify the specific activities which, if successfully performed, would distinguish more successful from less successful Eligibility Counselors. A guide is developed from these materials.

There are two particular advantages to use of tasks as indicants. One is that this approach provides a more direct description of the past behavior of an applicant which may be relevant to performance of the job than is either the traditional rating scheme or a KSAP-based rating scheme, to be discussed shortly. Second, it provides a close link to the job, from a common-sense perspective of content validity. Third, it is often easier to develop than alternative methods, and in our experience it is also easier for an applicant to understand and to complete. Fourth, the inventory version of this approach is easy to score and can readily be automated. Finally, it is possible to conduct studies of individual indicants in the same manner as items are studied in a paper and pencil test. Responses to indicants can easily be validated against external variables, including expert judgments concerning their validity, criteria of job performance, and so on.

Under what circumstances is a task-based model likely to be valid? Our data to date imply that it can be valid only when there is sufficient heterogeneity in the applicant population with respect to past experience to produce substantial variance. In addition, as in the case of test items, indicant "difficulties" must be reasonably high. This apparently is needed both to assure score variance and to reflect job task difficulty. A more critical job task may be more difficult because the base rate for success with respect to performance of the task on the job is lower. Our experience also suggests that the task model is most appropriate for jobs that are relatively structured rather than unstructured, and in which variation in performance of tasks is not as critical as whether an employee can or cannot perform the task; for example, a laboratory aid in contrast to a senior manager. It

appears to be less satisfactory than alternative methods when knowledges are critical factors distinguishing effective from less effective performance and when little or not prior experience is necessary for effective performance. Finally, it appears to be inappropriate when a vast range of possible past experiences are, in the opinion of experts, associated with job performance. The validities of individual indicants under these circumstances may be too low, and a logical scoring process is almost impossible to devise.

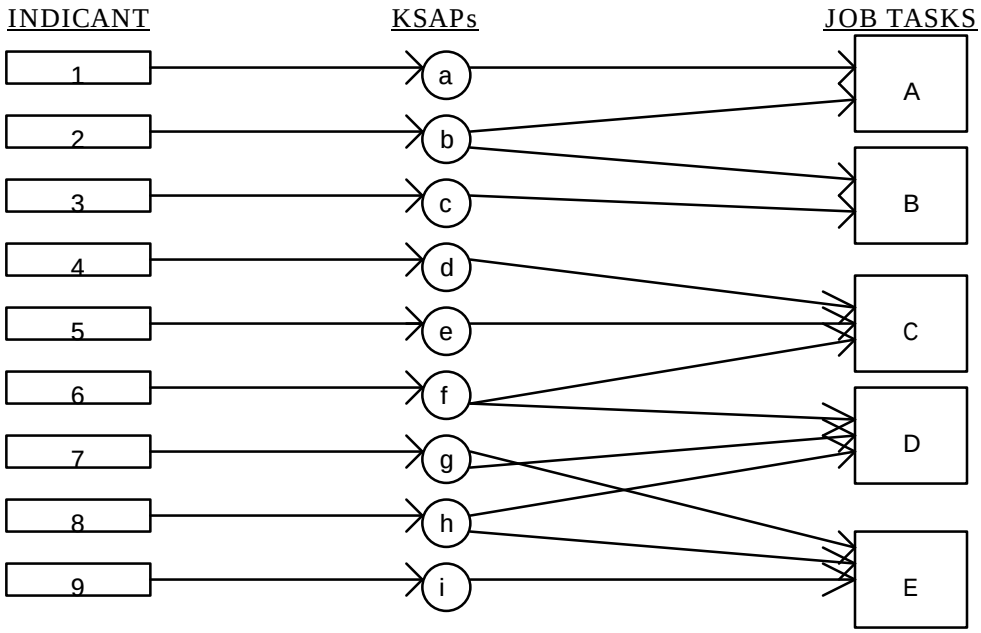

Figure 12. KSAP-based rating model.

The KSAP model represents a very different approach. Indicants are developed to reflect KSAPs rather than job tasks in a direct fashion. The job element approach is widely used and represents a major application of this model. In our experience, however, the job element method is not always easy to use and self-ratings which are frequently employed as part of the examination content have produced in some case marginal results at best. Primoff and others have reported

substantial evidence to support use of this approach for trades and blue-collar work, and there is some evidence to support its use in other occupations as well. Two other applications of the KSAP approach also appear to be useful, although our data on these methods to date are limited. One method is a simple checklist, similar to the task inventory mentioned earlier, in which the applicant simply checks whether he or she possesses the KSAP or not. This method has worked quite well for the selection of Homemakers in social service settings, for example, in which the traditional rating system produces scores which are inversely related to scores based on the KSAP methods. Another good example is entry-level Correctional Officer, for which a KSAP-based supplement is incorporated into part of the paper and pencil test. Another application is an inventory of very specific, well-defined knowledges for scientific and technical jobs.

The KSAP model is appropriate, we believe, in many circumstances in which the task model is not. It is more appropriate when expert judgment implies that experience is not necessarily important, when competence is acquired in a variety of different ways, and when knowledges or willingnesses clearly differentiate successful from less successful performance. It is sometimes more difficult to use this approach, however, in part because of the more indirect linkage to job performance in many cases, and because the indicants themselves are often associated only indirectly with the KSAPs.

Both in theory, and empirically, the behavioral model is most strongly supported. A job class for which we have considerable data concerning the behavioral consistency is titled, Senior Eligibility Counselor. An applicant for this class is typically a Junior Eligibility Counselor, and sometimes lower-level eligibility specialists within our Public Welfare agency. The examinations for the lower-level classes are paper and pencil tests; for the Senior Eligibility Counselor, the examination has been solely a rating of training and experience. We were able

to do a limited criterion validation study with this class because the bulk of the candidates are current employees performing virtually the same functions as the Senior Eligibility Counselor. The sample size in all analyses is 104.

The job performance criteria consist of performance ratings by supervisors on the dimensions "ability to learn," "quality of work," "use of working time," "ability to work with others," "quantity of work," "responsibility and initiative," "attendance and punctuality," "appearance," and a summary evaluation. (See Appendix) A group of applicants were examined by means of the traditional rating scheme, a task-based supplementary application, and a behavioral supplementary application.

TABLE 3

	Overall Job Performance
Traditional Rating	
Training Score	-.11
Experience Score	.06
Total Score	-.01
Task-based Supplement	
Training Score	-.07
Experience Score	.15
Total Score	.10
Behavioral Supplement	
Training Score	-.01
Experience Score	.25**
Self-Rating	-.02
Total Score	.26**

**$p < .01$

The concurrent validity coefficients indicate quite clearly that the traditional rating, including both the component training scores and the experience scores, bear no relationship whatever to the rating of job performance. Neither did the task-based supplement, although there are several interesting characteristics in

these data which cannot be reviewed here. In the behavioral supplement, the total score is significantly correlated to the criterion, as in the experience score, which is the component of this examination in which the behavioral consistency model is applied. Self-ratings using the behavior consistency model, however, are unrelated to the global criterion, as is the training score, which is essentially based on a KSAP model.

TABLE 4

Traditional Rating	Job Experience	Dept. Experience
Training Score	-.27**	-.67**
Experience Score	.41**	.74**
Total Score	.36**	.48**
Task-based Supplement		
Training Score	-.10	-.44**
Experience Score	.11	.07
Total Score	.11	.02
Behavioral Supplement		
Training Score	-.30**	-.51**
Experience Score	.05	-.10
Self-Rating	-.01	-.17
Total Score	.00	-.18
Overall Performance	-.02	.04

**p ⟨ .01

If we examine the relationship between these score components and experience, several interesting patterns emerge. In this table, job experience means the number of months the individual has been employed as a Junior Eligibility Counselor. Departmental experience is the number of months the candidates have worked within the Department of Human Services, regardless of job title. For the traditional rating scheme, the training component is related inversely to

experience, particularly departmental experience. This occurs among these applicants because the minimum qualifications provide for a substitution of work experience for training within this entire occupational series. The experience score component is, of course, strongly associated with job experience and especially departmental experience. The total score, using the traditional rating, thus bears a substantial relationship to experience.

In the case of the task-based supplementary application, using a KSAP-base subcomponent for training in which specific courses deemed by SME's to be relevant to job performance are scored, the relationships of training to job experience are again negative, but of much smaller magnitude than the correlations found in the traditional rating scheme. The experience score, however, is virtually unrelated to months of experience either in the directly comparable job of Junior Eligibility Counselor, or total departmental experience. Total scores bear a similarly insignificant relationship to amount of experience. In the case of the behavioral supplement, the training component, which is quite similar to the training component used in the task-based supplement, is again negatively associated with the two experience variables. The experience scores, the self-ratings, and the total scores bear zero relationships to job experience in the Junior Eligibility job class, and tend to be slightly negative with respect to experience in the agency. Length of experience is completely unrelated to overall performance.

TABLE 5

TRADITIONAL RATING

	Training Score	Experience Score	Total Score
Traditional Rating			
Training Score	- -	-.76**	-.19*
Experience Score	-.76**	- -	.78**
Mean	98.83	200.70	299.88
S. D.	52.03	81.53	53.73

**p ⟨ .01
* p ⟨ .05

If we examine the relationships between components of the traditional rating, the major portion of total score variance is due to variation reflected in the experience component. The training and experience components are strongly related inversely, and the training component thus bears a small negative relationship to the total score.

TABLE 6

TASK-BASED RATING

	Training Score	Experience Score	Total Score
Task-Based Rating			
Training Score	- -	.01	.24*
Experience Score	.01	- -	.93**
Mean	7.83	43.54	55.72
S. D.	6.74	11.88	13.87

**p ⟨ .01
*p ⟨ .05

In the task-based rating system, training and experience are independent components. Total score is almost entirely determined by experience; training has a very modest relationship.

TABLE 7

BEHAVIORAL SUPPLEMENT

	Training Score	Experience Score	Total Score	Self-Rating
Behavioral Supplement				
Training Score	- -	.01	.16	.06
Experience Score	.01	- -	.98	.61
Total Score	.16	.98	- -	.60
Mean	.96	5.17	8.96	14.77 = 5.94
S. D.	.91	1.16	1.97	2.41

**p ⟨ .01

In the behavioral supplement, total score is again determined almost entirely by the experience score, which is the component in which the behavioral consistency model was applied. This component bears no relationship to the training score, and its correlation with the self-rating is .61. The self-rating, which is a rating provided by each candidate to describe his or her own perception of competence in each component which was evaluated by an analyst to provide the experience score, is not associated with training, and is relates strongly to total score. Despite these relationships, and the relationship of the total score in this supplement to job performance, the self-rating is not related to supervisory rating of job performance as noted in Table 3.

TABLE 8

	Traditional Rating		
	Training Score	Experience Score	Total Score
Task-based Supplement			
Training Score	.59**	-.52**	-.22*
Experience Score	-.07	.10	.10
Total Score	.09	.01	.10

**p ⟨ .01
*p ⟨ .05

The behavioral supplement scores tend to be negatively associated with the traditional scores. These differing approaches clearly rank-order candidates in different ways. (See Table 8 and 9.) The behavioral approach may even tend to rank-order candidates in the reverse order of that produced by the traditional rating. Differences between the various T and E methods are clear in every comparison we have made, with intercorrelations rarely exceeding .25 to .30.

TABLE 9

	Traditional Rating		
	Training Score	Experience Score	Total Score
Behavioral Supplement			
Training Score	.62**	-.56**	-.26**
Experience Score	.08	-.13	-.11
Self-Rating	.12	-.17	-.13
Total Score	.17	-.21**	-.16

**p ⟨ .01
*p ⟨ .05

The conclusion which must be drawn from these results is that the alternative methods do not represent different ways of examining candidates for

employment with respect to the same competencies. The differing methods are measuring different things, and it is obvious that some of these things are completely unrelated or perhaps even inversely associated with job performance. A great deal of empirical research is clearly needed to delineate more precisely the conditions under which a T and E rating will be a reasonably valid alternative to other examining strategies, and the kinds of methods which are useful. Some of our work suggests to us specific problems which we have begun to address.

First, use of experts to estimate the validities of individual indicants requires further exploration, possibly using methods which have been adopted in recent years to define critical job performance dimensions, job tasks, and KSAPs. Just as Primoff has successfully demonstrated that the critical KSAPs can be pinpointed by means of his rating system, we need to conduct comparable studies to find ways of pinpointing valid indicants.

Second, we need more research on response formats used in supplementary applications. We find that direct self-ratings by applicants are often of no value either as a consequence of insufficient variance or because they simply do not correlate with independent measures of job performance. Among the various response formats, some are easier for applicants than others, and some are more easily scored by analysts than others. But we don't know if the easiest approaches produce the most valid data.

Third, the scoring process itself is of major importance. The underlying psychometric theory supporting the simple scoring of paper and pencil tests is not applicable to T and E ratings, and in some instances is classic theory logically inappropriate. Alternative scoring methods often produce different results.

Finally, although the behavioral model appears., at present, to be strongly supported both theoretically and by existing data, we hope that other methods will be sufficiently studied to fully test our hypotheses. The behavioral model is difficult

and complex to use, and it may not be applicable for many jobs for which prior experience is not expected. Criterion validity studies are essential to this process. We suspect that the answer to the question, "Is a T and E rating valid?" will not be simple. It depends on which method, for what kind of applicant, and under many circumstances these methods may not be valid at all. The more appropriate question may thus be, "When is a T and E rating valid?"

REFERENCES

Beardsley, V. A. A Study of the Rating of Education and Experience as an Examination Method in the Pennsylvania State Civil Service Commission. Commonwealth of Pennsylvania, State Civil Service Commission, Bureau of Examinations, Research and Special Projects Division, 1976.

Johnson, J. C. The role of a state personnel psychologist. Paper presented at the symposium, The Role of Public Personnel Psychologists, the Southeastern Psychological Association, Atlanta, Georgia, 1977.

Johnson, J. C. and Hill, J. O. The Development and Concurrent Validation of a Written Test for Counselors 1 in the Tennessee Department of Correction. Technical Report, State of Tennessee, Department of Personnel, Division of Research, 1976.

Perry, R. A. Public Sector Selection Specialist: A Survey of State and Local Government Utilization and Training Needs, Public Personnel Management, 1980, 9, 86-93.

Tenopyr, M. Content-Construct Confusion. Personnel Psychology, 1977, 30, 47-54.

Tenopyr, M. USCSC, BRE. Exam Preparation Manual (Procedures for Job Analyses and Federal Examination Planning), June 1977.

Yost, E. L. Selection of Beginning Level Foresters in the Commonwealth of Pennsylvania. Unpublished, 1967. Cited in Beardsley (1976).

STATE OF TENNESSEE DEPARTMENT OF PERSONNEL

PERFORMANCE REVIEW FORM

NAME	POSITION NO.	DUE DATE	SOC. SEC. NO.

PNF 604
NON-SUPERVISOR

		RATER SOCIAL SECURITY NUMBER											

OBSERVATION FREQUENCY		EVALUATION TYPE		NO RATER SUPERVISES	
DAILY	☐	PROB.	☐	1 - 3	☐
2 - 3 DAYS	☐	ANNUAL	☐	4 7	☐
WEEKLY	☐	OTHER	☐	8 - 12	☐
				OVER 12	☐

INSTRUCTIONS: READ EACH DESCRIPTION CAREFULLY: THEY ARE NOT IN ORDER. SELECT ONE DESCRIPTION FOR EACH CONTENT AREA AND BLACKEN IN THE SPACE PROVIDED WITH A NUMBER 2 PENCIL. THE EMPLOYEE BEING RATED SHOULD THEN MARK EITHER THE AGREE OR DISAGREE BLOCK FOR EACH ARE. MAKE NO OTHER MARKS!

ABILITY TO LEARN	☐ AGREE	☐ DISAGREE	☐ NOT SURE	
☐ Instructions must often be repeated; difficulty understanding directions	☐ Usually able to understand and follow directions; needs detailed instructions on some tasks.	☐ Able to understand and follow instructions; learns new ways with few difficulties.	☐ requires little instruction; sizes up most situations readily.	☐ Grasps and applies new ideas immediately; seldom has to be shown; excellent memory
QUALITY OF WORK				
☐ Job finished accurately; work may be relied on.	☐ Frequently makes mistakes; acceptable quality on repetitive work.	☐ Checks own work; seldom makes mistakes. Work is neat.	☐ Often makes mistakes; work below acceptable standards.	☐ Work needs comparatively little checking. Acceptable quality.
USE OF WORKING TIME				
☐ Wastes little time. Starts work immediately. Does not disturb other workers.	☐ Kills time. More conscious of time than in doing a good job.	☐ On job at all times; works up to the last minute; very efficient. Energetic anxious to get things done.	☐ Usually on job; makes adequate use of working time.	☐ Spends time away from work. Poor use of working time.
ABILITY TO WORK WITH OTHERS				
☐ Displays teamwork; secures good efficiency; uses tact and diplomacy; respected by others	☐ Arouses resentment. Ineffective when working with others; no use of tact and diplomacy.	☐ Average use of tact and diplomacy. Gets along with co-workers.	☐ Does not promote teamwork. Has difficulty making friends. Poor use of tact and diplomacy.	☐ Executive ability; handle difficult situations tactfully. Highly cooperative; cheerful. Can say "no" without giving offense.
QUALITY OF WORK				
☐ Generally fast and skillful; does work best and quickest way. High production record. When working with others, does own job and more.	Works at moderate speed. Needs some encouraging and urging. Good output on repetitive work.	☐ Output tends to be low. Needs improvement in skill, work knowledge. Has difficulty adapting to new work.	☐ Lacks basic skills for job. Little desire for improvement. Frequently slows down work.	☐ Works fast, skillfully. Understands method for turning out work. Does share when working with others.
RESPONSIBILITY AND INITIATIVE				
☐ Cheerfully takes any job offered; Never passes the buck or gives alibis. Always offers suggestions on better way of doing job.	☐ Goes out of way to accept responsibility; perfectly able to accept all given to him. Bright in doing things. Self-starter; always planning new ways of doing job.	☐ Avoids assuming responsibility; prefers to let things remain the same unless forced to change.	☐ Will not accept responsibility. Resents new methods; ambition is extremely limited.	☐ Will accept responsibility when necessary. Does not buck new methods, but does not offer many suggestions.
ATTENDANCE AND PUNCTUALITY				
☐ Never absent or late except for good reason. Reports absence or lateness in advance.	☐ Often absent or late; usually notifies supervisor.	☐ Continuously absent or late without notifying supervisor.	☐ Seldom absent or late. Does not always notify supervisor in advance.	☐ Always on time, good attendance record. Being on time is a matter of personal pride.
APPEARANCE				
☐ Appropriate dress and grooming. Makes good impression.	☐ Neat; makes good impression; makes certain clothes are clean and in good shape.	☐ Average appearance, bearing and poise.	☐ Appearance and poise need some improvement.	☐ Appearance is below acceptable quality; lacks concern for appearance.
OVERALL EVALUATION				
☐ Results achieved consistently; performance exceeds requirement of the job.	☐ Results achieved exceeded the requirements of the job at times.	☐ Results achieved meet the requirements of the job.	☐ Performance is marginal; needs improvement.	☐ Performance below acceptable standards.

IMMEDIATE SUPERVISOR AND I HAVE DISCUSSED THIS EVALUATION. I HAVE CHECKED MY AGREEMENT OR DISAGREEMENT WITH EACH AREA REVIEWED AND I UNDERSTAND THE AREAS IN WHICH I NEED IMPROVEMENT.

DATE IMMEDIATE SUPERVISOR DATE REVIEWED BY SECOND LEVEL SUPERVISOR