Methodology Review: Principles, Procedures, and Findings in the Application of Background Data Measures

Michael D. Mumford
Georgia Institute of Technology

William A. Owens
The University of Georgia

This paper provides a review and critique of the background data literature. As a life history measure, the effective application of background data items is based on a developmental strategy in which a pattern of prior behavior and experiences is related to certain forms of criterion performance. This principle provides a framework for discussing the various issues involved in generating an adequate pool of background data items. The four principal methods for scaling background data items are examined: rational scaling, factorial scaling, empirical keying, and subgrouping. The relative strengths and weaknesses of these four techniques are considered along with current research needs in each area. This review indicates that substantial progress has been made in the development and application of background data measures, but that alternatives to the traditional empirical keying strategy should receive more attention.

From the earliest days of psychological measurement, past behavior has been found to be an effective predictor of future behavior. Thus, psychologists have long displayed an interest in the potential applications of life history information (Dailey, 1960; Galton, 1883). Descriptions of an individual's life history may be cast in many forms, ranging from diaries and narrative biographies to quantified demographic data. While all these measures have substantial value, psychometricians have come to rely on a particular form of life history information, referred to as scored autobiographical data or background data, to address a variety of pragmatic problems arising in the areas of selection, placement, and differential diagnosis. The following discussion will review the available evidence pertaining to the development and scaling of background data items.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 1, March 1987, pp. 1–31

0146-6216/87/010001-31$2.80

Historic Considerations

The modern background data form represents an outgrowth of the job application blank. Ferguson (1962) reported that the first mention of standardized questions concerning an individual's life history occurred in an 1894 address by Colonel Thomas L. Peters. In this address Peters suggested that the selection of life insurance agents might be improved by presenting all applicants with a common set of questions concerning prior behavior and experiences. Shortly thereafter, the idea of formulating quantitative weights capable of discriminating good and poor performers appeared. Prior to World War I, Woods (1962) undertook an empirical analysis of the application blank responses characteristic of successful and unsuccessful salesmen. Over the next 10 years, a number of studies appeared in the literature demonstrating the utility of weighted application blanks in differentiating successful and unsuccessful employees in a variety of occupational fields (Goldsmith, 1922; Kenagy & Yoakum, 1925; Viteles, 1932).

The next major advance occurred during World War II when application blank responses were first cast in a multiple-choice format. An early study by Guilford and Lacey (1947) found that an empirically keyed set of multiple-choice background data items would predict success in Air Force training programs at the $r = .40$ level. In combining the predictive power of empirically keyed background data items with the scoring efficiency provided by the multiple-choice format, the modern background data form came into being. During the last 50 years, measures of this sort have been widely applied in personnel selection.

Obviously, the foregoing presents a limited historic overview. But it is important to recognize that the methodology commonly employed in constructing background data measures has been available for some time. In applications of the method individuals are presented with standardized, multiple-choice questions concerning their past behavior and experiences, and item responses are weighted on the basis of their ability to differentiate good and poor performers. While the predictive power of these empirically keyed background data measures is undoubted (Owens, 1976), the very success of this technique has become something of a liability. More specifically, in their satisfaction with the resulting validity coefficients, investigators have often lost sight of the psychological principles underlying the development and application of background data measures. Thus it would seem appropriate to begin this review by briefly reiterating a few of these principles.

Descriptive Considerations

As with any other life history measure, the use of background data is predicated on the hypothesis that an individual's past behavior and experiences are a potential predictor of future behavior and experiences. This is not to say that people will necessarily behave in the future as they have in the past. Rather, this statement implies that prior learning and heredity, along with the resources and limitations which are their result, will make some avenues of behavior more *likely* in new situations, thus allowing some prediction of future behavior through assessment of the individual's earlier behavior and experiences.

What distinguishes background data from other life history measures is the particular strategy used to capture this predictive information. Background data measures generate a description of an individual's life history through a quasi-longitudinal self-report format (Mikesell & Tesser, 1971). Essentially, this entails presenting all individuals with a common set of questions concerning their behavior and experiences in relatively discrete situations likely to have occurred earlier in their lives. In responding to these questions, individuals are asked to recall their typical behavior in or reactions to the referent situation, and then to select, from the available response options, the alternative or alternatives that best describe the overall pattern of their prior behavior and experiences.

Three advantageous outcomes are associated with the quasi-longitudinal strategy applied in background data measures. First, when multiple-choice items are in use, a comprehensive description of life history influences may be formulated through an economical paper-and-pencil format (Owens, 1976). Second, an operationally well-defined description of prior behavior and experiences may be obtained (Mumford & Owens, 1982). Finally, this item format yields a standardized description of life history influences which is amenable to quantitative analysis (Owens, 1976).

To further elucidate the characteristics of background data, it might be useful to consider its relationship to other types of measures. Clearly, background data can capture information similar to that obtained from interviews and demographic forms (McMurray, 1947; Mosel & Wade, 1951). However, background data measures differ from standard demographic forms by virtue of their ability to capture a wider range of prior behavior and experiences. Further, background data generally provides a more valid and economical measure of life history than interview data, which can be affected by observer attributions (Parsons & Liden, 1984).

It is also apparent that background data measures

have much in common with standard self-report personality measures (Rawls & Rawls, 1968). But unlike personality measures, background data items do not call for general descriptions of behavioral tendencies. Instead, they focus on prior behavior and experiences occurring in specified, real-life situations. This strategy has the advantage of minimizing self-report bias while carrying out assessment within the interactionist framework that appears to provide the best available vehicle for understanding the nature and ontogeny of differential characteristics (Magnusson & Endler, 1977).

Background data measures differ from standard aptitude, ability, and achievement tests in terms of the time frame and observational conditions employed in assessment. Aptitude, ability, and achievement test items present the individual with an immediate, somewhat artificial, situation calling for maximum cognitive performance (Wagner & Sternberg, 1985). The individual's success in this activity serves to predict performance potential on related tasks. With background data measures, a description of typical behavior and experiences in real-life situations is called forth from memory and used to predict performance. As a result, background data measures cannot provide upper bound descriptions of performance potential, although they may prove highly effective in predicting the performance levels actually observed.

Predictive Considerations

The focus of background data items on discrete elements of prior behavior and experience implies that whenever a background data item predicts performance, it must also represent a developmental antecedent or a sign for later performance. Cognizance of this fact led Owens (1976) to argue that performance prediction through background data items is based on a developmental strategy. This principle implies that optimal prediction will require a set of background data items capable of capturing the prior behaviors and experiences impinging on the later expression of criterion performance, and the subsequent weighting of these items in terms of their relative importance to the development of differential performance.

This description of the ideal background data measure has two important implications. First, the predictive power of any background data scale will depend on whether the item pool has provided a reasonably comprehensive description of the relevant behavioral and experiential antecedents. Second, because these behavioral and experiential antecedents must be identified and weighted with respect to their impact on criterion performance, it becomes apparent that effective application of background data items requires the quantitative definition of a developmental pattern. These observations indicate that item development and scaling methodology constitute the two most important topics in the background data literature.

In the ensuing discussion, an attempt will be made to delineate and clarify the many considerations entering into the effective construction and scaling of background data items. At this juncture, however, one caveat seems in order. Owens and Champagne (1965) and Brodie, Owens, and Britt (1968) reported that some 200 references could be found in the background data literature. During the intervening 20 years, this number has more than doubled. In view of the sheer volume of this literature and the availability of recent reviews on the criterion-related validity of background data measures (Asher, 1972; Ghiselli, 1973; Owens, 1976; Reilly & Chao, 1982), this paper will not attempt a detailed review of all relevant studies. Instead, the primary concern will be with evidence bearing on the principles and methodological procedures influencing effective application of background data measures.

Item Development

Principles

The preceding discussion indicates that the effectiveness of a background data measure will be conditioned by the characteristics of the item pool in use. Yet no attempt has been made to delineate the manner in which item development procedures influence the validity and utility of background data

measures. Historically, the validity of background data measures has been established primarily through criterion-related validity studies (Pace & Schoenfeldt, 1977). However, like any other measurement system, the meaningfulness of background data measures can be fully established only by considering evidence pointing to content, construct, and criterion-related validity (Messick, 1975).

Content validity. Item construction and sampling procedures will condition the content validity of any background data measure. A background data measure may be said to display content validity to the extent that items incorporated in the measure capture prior behavior and experiences contributing to the development of differential performance on the criterion of interest. Because the developmental antecedents of most real-world criterion performances are highly complex and are subject to variation by subpopulation (Ferguson, 1967; Lerner & Busch-Rossnagel, 1981), the content validity of a background data measure cannot be justified solely on the basis of overt similarity between item content and criterion behaviors. Rather, inferences concerning content validity are made possible by explicit hypotheses defining a domain of potentially significant developmental antecedents, and a systematic sampling of domain elements during item specification.

Criterion-related validity. Definition of an item pool in relation to well-formed developmental hypotheses not only provides a basis for drawing inferences concerning content validity, but also allows stronger inferences concerning the predictive power of background data measures. An item pool is likely to yield poor prediction and misleading inferences concerning individual performance when it is (1) deficient due to a failure to capture significant antecedents, (2) contaminated by the inclusion of irrelevant prior behavior and experiences, or (3) biased by failure to capture shifts in developmental patterns across age, race, or sex groups.

Under these conditions it will be difficult to draw firm conclusions concerning the predictive efficiency of background data measures. Here poor prediction might be attributed to either an inadequate item pool or an intrinsic failure of background data items to predict performance. One illustration of this principle may be found in a study by Schwab and Oliver (1974) examining the ability of background data measures to predict turnover among clerical workers, unskilled hourly employees, and machinists. They found that the empirical keys in use failed to cross-validate, and concluded that background data measures were not an effective predictor of turnover. Yet this result might just as well be attributed to their failure to employ items having some known relevance to turnover. It is of note here that studies by Williams (1961), Kavanagh and York (1972), and Quaintance (1981) found that items formulated on the basis of specific developmental hypotheses concerning the antecedents of performance are far more likely to yield significant relationships with an external criterion than items for which such hypotheses could not be formulated.

Construct validity. Because both content and criterion-related validity can be subsumed under the general rubric of construct validity (Cronbach, 1971), the foregoing discussion of the influence of item content on these validation strategies indicates the potential import of item content for the construct validity of background data measures. Owens and Schoenfeldt (1979) noted that in its broadest sense, the construct validity of a background measure will be reflected in the measure's ability to capture a psychologically meaningful pattern of individual development regardless of whether this pattern is referenced against a specific criterion measure. Of course, it will be possible for an investigator to establish a meaningful developmental pattern only when the background data items in use have some substantive psychological significance. Thus, systematic item development procedures may play an important role in establishing the construct validity of background data measures.

The influence of item content on the construct validity of background data measures has been illustrated in a number of studies. For instance, Malloy (1955), Place (1979), and Mumford and Owens (1982) showed that the construct validity of background data measures may be established through

the interpretability of a pattern of item correlates as it relates to the available evidence concerning the development of some psychological construct. Moreover, studies by Owens and Schoenfeldt (1979), Mumford (1981), and Mumford and Owens (1984) indicated that the interpretability of the internal relationships exhibited by a set of background data items may prove of great value in establishing the meaningfulness of a developmental pattern, and may thus provide evidence pointing to the construct validity of the measure. Owens and Schoenfeldt (1979) also showed that the relationship of background data items and scales to external measures will often prove useful in establishing construct validity.

Item Construction

Domain definition. Given the foregoing, it should be apparent that the first step in formulating a set of background data items involves defining a domain of antecedent behavior and experiences (Owens & Schoenfeldt, 1979). Definition of a developmental domain will generally begin with a clear, precise definition of the criterion performances of interest to the investigator. In generating this definition, some attention should be given to the environmental conditions under which these behaviors occur, and the processes likely to underlie effective performance (Fleishman & Quaintance, 1984). Additionally, it may prove useful to consider demographic attributes of the population at hand, such as its modal age, so that this information may be considered in evaluating performance and the likely antecedents of performance.

After an adequate understanding of criterion performance has been obtained, definition of the developmental domain may begin. Typically, a comprehensive knowledge of the differential attributes contributing to performance will not be available at the outset of a study. Three basic strategies for obtaining the descriptive information required for domain definition have appeared in the literature. First, a number of studies have collected job analysis data and used this information to develop hypotheses concerning the differential attributes underlying criterion performance (Hough, 1984; Levine & Zachert, 1951; Mumford, Cooper, & Schemmer, 1983). Second, investigators often identify salient attributes on the basis of the existing literature pertaining to the differential characteristics and background data items capable of predicting performance (Mitchell & Klimoski, 1982; Webb, 1960). Third, systematic life history interviews are sometimes used to formulate hypotheses concerning the life history influences relevant to performance (Myers & Fine, 1980). While all three of these strategies are likely to have some value, research contrasting their relative merits in domain definition is not available. This is unfortunate because these techniques appear to have some clearcut strengths and weaknesses. Of course, this suggests that a truly comprehensive domain definition will require the use of descriptive information drawn from multiple sources.

Item specification. Item specification will follow definition of the developmental domain. In the first step, dimensions identified in the initial domain definition that cannot be adequately assessed through background data items are eliminated. Next, item specifications capable of capturing prior behavior and experiences indicative of the attributes held to contribute to performance on the remaining dimensions are identified. Essentially, this entails the construction of hypotheses implying that a given form of prior behavior or experience influences the later expression of the attribute at hand (Morrison, 1977). Although this process requires substantial judgment, a number of pieces of information might be used as a guide, including (1) the developmental literature, (2) life history interviews with incumbents, (3) the known life history correlates of various characteristics, (4) the typical factor loadings of background data items, (5) the known predictive characteristics of various background data items, and (6) hypotheses formulated on the basis of general psychological knowledge.

In developing these item specifications, investigators must carefully attend to the age range of the target population as well as the behavior and experiences likely to have occurred in their lives. This will help ensure item specifications tied to the

situations to which members of this age group have been exposed (Ferguson, 1967; Kurtz, 1941; Peterson & Wallace, 1966). Moreover, because items may be lost in prescreening and because reliability will be enhanced when multiple items are available to cover each aspect of the developmental domain, it would seem prudent for investigators to develop at least 10 to 15 item specifications in each content area.

Item formats. Following construction of item specifications, these content specifications must be translated into multiple-choice items. Owens (1976) identified seven formats that are commonly employed in the construction of background data items: (1) yes-no; (2) non-continuum, single choice; (3) continuum, single choice; (4) non-continuum, multiple choice; (5) non-continuum, plus escape option; (6) continuum, plus escape option; and (7) common item, multiple continuum. Figure 1 presents examples of each item format. For the most part these examples are self-explanatory; however, one other point seems worthy of mention here. When non-continuum item formats are in use, each response option effectively constitutes an item unto itself with respect to scoring and weighting. Thus, the number of operational items must be considered to control for chance effects in assessing the proportion of items yielding significant differences.

A number of studies have examined the characteristics of effective background data items. For instance, Lecznar and Dailey (1950) contrasted background data inventories keyed through the response pattern or correlational technique with inventories keyed through conventional response option weighting. They found that these two methods yielded comparable initial validities, but that the pattern keys showed less shrinkage upon cross-validation. When this observation is coupled with the somewhat greater interpretability of continuum type item formats, it suggests that continuum items should be employed whenever possible.

In a comprehensive study of the characteristics of background data items contributing to retest reliability, Owens, Glennon, and Albright (1962) found that reliability was enhanced by (1) keeping questions as simple and brief as possible, (2) graduating the response options on a numerical continuum, (3) providing an escape option when all possible alternatives have not been covered by the response options, and (4) wording response options and questions in such a way as to provide a pleasant or neutral connotation. Again, these findings would seem to support the use of the continuum format modified to include escape options.

Asher (1972) suggested that the reliability and validity of background data items might also be improved by focusing item content on verifiable phenomena whenever possible, and making item content and response options sufficiently specific to avoid ambiguity in interpretation. Further, he noted that items intended to assess prior behavior and experiences in a manner congruent with colloquial usage may yield better results. Similarly, Mumford (1983) argued that the use of background data items requiring the recall of past behavior and experiences in relatively discrete situations will improve reliability and validity. Finally, Buel (personal communication, May 15, 1971) provided evidence suggesting that background data items phrased in a historical context, capturing the values, attitudes, and beliefs related to performance, will tend to be more effective predictors than demographic items. This finding may reflect the fact that the values, attitudes, and beliefs which individuals carry with them from past situational exposures are the characteristics most likely to determine future behavior and experiences (Tyler, 1965).

Because it is possible that the developmental patterns underlying performance vary with subpopulation status, some attention has been given to the need to develop separate response options or item weights within different subpopulations. In this regard, studies by Laurent (1962) and Ferguson (1967) have indicated that reliability and validity of background data items may vary with respondent age. Similarly, studies by Mumford, Shaffer, Jackson, Neiner, Denning, and Owens (1983), Nevo (1976), Webster, Booth, Graham, and Alf (1978), Ritchie and Boehm (1977), and Federico, Federico, and Lundquist (1976) showed that the predictive characteristics of background data items are

Figure 1
Types of Background Data Items

1. **Yes–No**
   Have you found your life to date to be pleasant and satisfying?

2. **Non-continuum, single choice**
   What was your marital status at college graduation?
   a) Single
   b) Married, no children
   c) Married, one or more children
   d) Widowed
   e) Separated or divorced

3. **Continuum, single choice**
   What is your weight?
   a) Under 135 pounds
   b) 136 to 155 pounds
   c) 156 to 175 pounds
   d) 176 to 195 pounds
   e) Over 195 pounds

4. **Non-continuum, multiple choice**
   Check each of the following from which you have suffered.
   a) Allergies
   b) Asthma
   c) Ulcers
   d) Epilepsy
   e) Headaches
   f) Arthritis
   g) Gastrointestinal upsets
   h) High blood pressure
   i) Loss of hearing

5. **Non-continuum, plus escape option**
   When are you most likely to have a headache?
   a) When I strain my eyes
   b) When I don't eat on schedule
   c) When I am under pressure
   d) January first
   e) Never have headaches

6. **Continuum, plus escape option**
   What was your length of service in your most recent job?
   a) Less than 6 months
   b) Between 6 months and 1 year
   c) 1 to 2 years
   d) 2 years or more
   e) No previous full time job

7. **Common stem, multiple continuum**
   Over the past 5 years, how much have you enjoyed the following? (use 1 to 4 below)
   a) Loafing or watching TV
   b) Reading
   c) Constructive hobbies
   d) Home improvement
   e) Outdoor recreation
      1) Very much
      2) Some
      3) Very little
      4) Not at all

different for men and women. More generally, these findings suggest that when there is reason to believe that normative age and sex role expectations may have led to different developmental patterns, empirical studies should be carried out to determine whether different items or item weights will be required to maximize within-group validity.

On the other hand, although one study by Toole, Gavin, Murdy, and Sells (1972) reported black-white differences in the predictive characteristics of background data items, significant ethnic group differences have not been obtained in most studies (Baehr, 1976; Cascio, 1976; Cherry, 1969; Lefkowitz, 1972; Malone, 1978; Mumford, Cooper, & Schemmer, 1983; Ritchie & Boehm, 1977; Sharf, undated). Thus it appears that there is sufficient similarity in the developmental patterns underlying performance across ethnic groups to permit application of a common item pool, provided that appropriate item construction procedures have been employed.

Item prescreening. Following construction of an item pool, it will generally prove necessary to conduct an item prescreening to ensure the appropriateness of item content with respect to the target population and the psychometric adequacy of the items. The first step in this prescreening is likely to entail a review of item content. A study by

Majesty (1967) is of some interest in this regard, because it implies that the rational screening of items can provide a vehicle for minimizing item bias and objectionability. In this study 1,036 students completed 295 background data items, and it was found that gender, race, and religion could be effectively determined by both rational and empirical keys. It was also found that demographic items were more likely to discriminate members of these subpopulations than self-description items. A later study by Mumford, Cooper, and Schemmer (1983) employed a rational screening technique in which two panels of subject matter experts were asked to review item content for objectionability and bias against certain subpopulations, such as females or minority group members. After the elimination of items judged to be biased, the resulting background data measure was found to be unbiased in a differential regression line analysis. If these results are coupled with Smart's (1968) observation that judges appear to employ a consistent set of rules in evaluating objectionability and potential bias in background data items, it seems that the rational screening of an initial item pool will serve to minimize the impact of these influences. Moreover, the Mumford, Cooper, and Schemmer (1983) study suggests that there may be some value in applying similar techniques to screen content for the clarity, appropriateness, and job relevance of both background data items and response options.

Once the initial pool of background data items has been constructed and their content reviewed, item tryout may begin. Here the item pool will be administered in a formal tryout employing 100 to 200 members of the target population. An excellent paradigm for item tryout may be found in Owens and Schoenfeldt (1979). Initially, the variability of item responses is reviewed, and items lacking sufficient variability are eliminated. Next, items with highly skewed or bimodal distributions are eliminated. Finally, the remaining items are intercorrelated and items are eliminated that fail to yield a significant and interpretable pattern of relationships with other items intended to capture the same or other theoretically relevant aspects of the developmental domain. While application of the foregoing procedures should ensure that the final item pool displays adequate psychometric characteristics, it should also be pointed out that, by conducting these analyses within subpopulations, marked shifts in response distributions or item correlations may be used to indicate the need for separate within-group scaling procedures.

Item Characteristics

Accuracy. Because background data items rely on self-reports of past behavior and experiences, there has been some concern that selective recall would distort the accuracy of item responses (Karla, 1981; Ross, McFarland, & Fletcher, 1981). Research examining verifiable background data items has produced a rather complex pattern of results. Studies by Keating, Paterson, and Stone (1950), Mosel and Cozan (1952), and Cascio (1975) found substantial agreement in comparing responses to background data items and objective information. However, studies by Goldstein (1971) and Weiss and Dawis (1960) found disagreement between self-report and verified data in up to 55% of the cases. The discrepancy between these two sets of findings may be attributed to the fact that the former set of studies employed correlational methodologies, while the latter relied on absolute levels of agreement. This observation suggests that, although there may be sufficient self-report bias to induce some response shifts, it is not sufficient to change the ordering of individuals. Although clear-cut empirical evidence is lacking, it seems likely that this effect represents a general self-presentation bias.

The correlational accuracy of background data measures does not appear to be limited solely to verifiable items. One study by Bronson, Katten, and Livson (1959) found correlations in the mid-40s between individuals' background data responses and psychologists' observations collected 16 years earlier. In a related investigation, Saunders (1983) obtained correlations ranging from .40 to .60 in contrasting parents' descriptions of their children during adolescence with their children's self-descriptions on a common set of background

data items. However, the degree of agreement observed here was higher for more objective items, such as high school grades. Nevertheless, when the limited reliability of item data is considered, the results obtained in both these investigations seem to support claims for the accuracy of item responses when there is no motive for faking.

Both Schrader and Osburn (1977) and Klein and Owens (1965) examined changes in item responses under honest and "fake good" conditions. Here it was found that individuals could improve their scores on items in the fake good condition. However, the Klein and Owens study also indicated that this faking will be far less effective when there is no apparent stereotype underlying the key employed in scaling background data items. This finding has been confirmed in a study by Lautenschlager (1985) which indicated that when individuals lacked a readily available stereotype, they had difficulty in faking their responses to background data items in such a way as to markedly improve their scores.

Under conditions where there is reason to believe that stereotypic faking might influence the effectiveness of background data measures, at least three control techniques might be employed. First, in item development investigators might systematically select items less sensitive to faking. It is of note here that Larson, Swarthout, and Wickert (1967) found that faking is most likely to occur on self-evaluation items with response options differing only in degree and on self-description items with response options differing in nature but not degree, while faking is least likely to occur when items focus on the description of past behavior in well-defined situations. Second, Cohen and Lefkowitz (1974) reported that background data keys can be developed to predict faking, and it is possible that such control keys might be incorporated into background data measures. Third, Norman's (1963) technique for superimposing faking keys on standard empirical keys might be applied in an attempt to control for faking.

An issue closely related to faking concerns the influence of various psychometric biases on item responses. Shaffer, Mumford, and Owens (1986), Mumford (1982), and French, Lewis, and Long (1976) correlated Crowne-Marlow Desirability scores and California F scale Acquiescence scores with responses to a set of background data items and factors. The results obtained in these studies indicated that Acquiescence produced near-zero relationships with both item responses and factor scores. Social Desirability produced relationships on the order of .20 with a sizable percentage of the items, but yielded near-zero relationships with factor scores. Taken as a whole, these observations suggest that response biases, such as social desirability and acquiescence, have remarkably little impact on background data measures. Yet the French, Lewis, and Long (1976) and Mumford (1982) studies suggest that the items most likely to yield significant correlations with social desirability scales typically examine success in social situations and traditional role expectations.

Reliability and validity. Aside from their freedom from bias, well-developed background data items also display a number of other advantageous psychometric characteristics. For example, Plag and Goffman (1967) presented evidence indicating that background data items display relatively low item intercorrelations, and this observation has been confirmed in a number of other studies (Kavanagh & York, 1972; Owens, 1976; Siegel, 1956). Lunneborg (1968) noted that when these low item intercorrelations are coupled with high retest reliability, they permit a few background data items to capture a great deal of descriptive information.

The relative independence of background data items has certain implications for the assessment of reliability. More specifically, the independence of these items makes it unlikely that the resulting scales will yield high internal consistency coefficients. Therefore, it is not surprising that the internal consistency coefficients obtained for rational background data scales lie between .40 and .80. Yet, as the veridicality studies would imply, background data items commonly yield substantial retest reliability coefficients. For instance, Bunch (1974) correlated item data collected over a two-month interval and obtained retest reliabilities of .60 to .80. Similarly, Saunders (1983) obtained retest coefficients near .60 in correlating item re-

sponses at age 18 and age 22. Thus, it appears that background data items provide an unusually reliable description of differential behavior and experiences, even over relatively long intervals.

Obviously, the validity of any given background data item will depend on the particular criteria under consideration. Because most criterion performances are determined by a complex developmental pattern, it is unlikely that sizable criterion-related validity coefficients will be produced by individual items. Nevertheless, studies by Plag and Goffman (1967) and Peterson and Wallace (1966) showed that background data items capturing salient antecedents of performance can yield sizable predictive validity coefficients. Of somewhat greater interest are the findings by Peterson and Wallace (1966) and Brown (1978) indicating that the predictive power of these items may be maintained over substantial periods of time.

Scaling Procedures

Once an adequate pool of background data items has been established, it will be necessary to obtain a summarization of the individual's item responses that has some predictive value. Four basic scaling techniques have been used to generate these summary descriptions; these techniques have been labeled (1) empirical keying, (2) rational scaling, (3) factorial scaling, and (4) subgrouping. The following sections will examine the literature pertaining to each of these scaling techniques and attempt to delineate their relative strengths and weaknesses.

Empirical Keying

Criterion issues. Historically, empirical keys have proven to be the most popular vehicle for scaling background data items. Essentially, empirical keys select and weight items on the basis of their ability to discriminate the members of a criterion group from some reference group. For instance, in personnel selection a key might be developed to discriminate high performers, or criterion group members, from the members of a reference group consisting of the current pool of job applicants. This statement implies that the nature and effectiveness of any empirical key will depend on the adequacy of the criterion measure used to define group membership, as well as the appropriateness of the reference group with respect to the intended application of the background data measure.

Because empirical keys are referenced against criterion group membership, a background data key will be no more effective than the definition of performance provided by the criterion measure. Thus, the first major concern in empirical keying efforts must be the development of an appropriate criterion measure. If a criterion measure is contaminated by spurious influences (such as geographic area in the assessment of sales performance), or if it is subject to bias in assessing the performance of certain subpopulations, these factors will affect criterion group membership. As a result, the selection and weighting of items on the basis of their predictive power will necessarily reflect the operation of these factors, and thereby yield a selection measure of limited utility. The potential impact of deficiency also argues for systematic criterion development efforts. An item reflecting an antecedent of some aspect of performance that is not captured by a criterion measure will not be retained in an empirical key. This may in turn result in misleading inferences concerning *individual* performance due to exclusion of relevant developmental influences. Finally, the referencing of background data items against a specific criterion implies that (1) the generality of an empirical key will depend on the generality of the criterion measure, and (2) the predictive power of the key will decline with changes in the nature of adequate criterion performance (Thayer, 1977).

On the basis of these considerations, Thayer (1977) argued that more attention should be given to criterion development issues in empirical keying efforts. Unfortunately, relatively few studies in the background data literature have devoted any substantial effort to criterion development because they have tended to employ hard criteria, such as turnover. However, the ambiguities arising in the Telenson, Alexander, and Barrett (1983) study, as a

result of the contaminating influence of geographic area, point to the importance of this issue. Moreover, studies by Malloy (1955), Laurent (1962), and Klein and Owens (1965) point to the enhanced efficiency of empirical keys formulated on the basis of well-constructed criterion measures.

Not only will the predictive power of an empirical key depend on an adequate definition of the criterion group, it will also depend on an adequate definition of the reference group. Traditionally, the issue of reference group definition has not received a great deal of attention, perhaps as a result of the generic concern with job applicants in selection studies. Yet it should be apparent that the generality of the reference group will influence the generality of an empirical key, just as shifts in reference group composition will cause changes in a key's predictive power (Strong, 1941). If a reference group does not constitute a reasonably well-defined and homogeneous sample of individuals, extreme variation within this group will make it difficult to produce the significant group differences required in any empirical keying effort. Additionally, different kinds of reference groups may produce substantial differences in the nature of the resulting keys. For instance, different keys may be obtained when the reference group consists of job incumbents rather than job applicants. Because these considerations suggest that reference group composition may be of some import, there would seem to be a need for systematic empirical investigations along these lines.

Keying procedures. In turning to the more specific methodological issues arising after these groups have been defined, the first concern likely to emerge will be sample size requirements within the criterion and reference groups. Because empirical keying requires demonstration of a significant mean or percent difference, it would seem that a minimum of 20 criterion group and 20 reference group members would be sufficient to produce stable results (Winer, 1971). However, it is important to note that when non-continuum items are in use, this requirement applies not to the item as a whole but to the response options per se. Operationally, this implies that (1) this minimum sample size requirement must be increased as a function of the number of response options, and (2) it will often be necessary to eliminate options with extremely low response frequencies, unless sample size is large due to the number of individuals required to produce stable differences.

Of course, it should be apparent that this minimum sample size applies only to the identification of absolute mean or percent differences. When investigators wish to formulate differential item weights reflecting the magnitude of the observed mean differences, these relative mean or percent differences will have larger sampling errors. Hence this minimum sample size must be increased to obtain stable item weights. The nature of this increase will depend on the magnitude of the mean or percent differences to be considered in assigning weights. It should be pointed out that when correlational statistics are in use, the relatively large sampling error associated with covariance statistics will require larger sample sizes. As a rule of thumb, between 200 and 400 individuals will be required in both the criterion and reference groups for correlational analyses.

The first step in key construction is the selection of items for inclusion in the final key. Clark and Gee (1954) and Nach (1965) showed that the inclusion of items yielding weak, albeit significant, discrimination will result in a disproportionate increase in error variance, and thus a net loss in reliability and criterion-related validity. Additionally, with respect to efficiency and economy, there is little point in scoring items yielding weak discrimination. Thus, in empirical keying efforts, an attempt should be made to eliminate items that fail to yield sizable differences. Commonly, items failing to yield a difference in a t test, correlational, or chi-square analysis significant at the .05 or .01 level are eliminated (Mumford, 1981; Place, 1979). However, this rule of thumb should be modified to take into account multiple comparisons and the relative contribution of error and true variance. Further, when a sizable sample is available, it may prove necessary to consider the magnitude of these differences. In past studies, a correlation of .10 to .15 has been found to provide an appropriate cut

point (Mumford, 1981). Of course, investigators may find it necessary to modify this rule of thumb to ensure adequate coverage of the antecedent domain.

Once a pool of items has been selected for inclusion in a key, attention will turn to the weighting of item responses. It should be noted that the differential weighting of items beyond simple unit weights will be most profitable when the intercorrelations among items are low and relatively few items are in use (Guilford, 1954). Because most background data measures contain a number of items and because item reliabilities will limit the stability of scoring weights, little may be gained from the application of more complex weighting schemes. Having stated this concern, the various weighting procedures employed in the development of empirical keys will be considered.

The weighted application blank is the device most frequently employed in the construction of background data keys. According to England (1961), this method entails determining (1) the percentage of individuals in the criterion and reference groups choosing an alternative, and (2) the difference between these two percentages. Strong's (1926) tables are then used to assign weights to each response option on the basis of the observed differences. In the correlational or pattern scoring method (Lecznar & Dailey, 1950), weights are assigned to items based on the magnitude of the observed relationship between each item and the criterion, using the appropriate correlational statistics. The differential regression techniques are similar to the correlational techniques, except that standard regression analysis procedures are employed to select and weight items on the basis of their ability to add to the prediction obtained from items already in the regression equation (Malone, 1978).

Beyond these three basic weighting techniques, a number of alternative weighting strategies have appeared in the literature. One of these is a variant on the weighted application blank (i.e., vertical percent method) that is referred to as the horizontal percent method (Stead & Shartle, 1940). In this method each item is weighted by dividing the number of criterion group members selecting a response option by the total number of criterion and reference group members. The deviant response scoring technique (Malloy, 1955; Webb, 1960) employs a correlational or weighted application blank strategy to construct item weights for criterion groups consisting of individuals above or below the regression line, after partialling out the variance attributable to an initial predictor set. Finally, Telenson et al. (1983) proposed a deviant response scoring technique where item responses are weighted on the basis of how few individuals select a given response option.

Studies comparing these weighting techniques have yielded some important findings. For instance, England (1961) suggested that the percent scoring methods will generally yield more stable weights than the available alternatives. However, studies by Lecznar and Dailey (1950) and Malone (1978) indicated that, although the weighted application blank procedure may produce higher initial validities, the correlational and regression procedures will yield equal or somewhat better cross-validated coefficients and show less shrinkage upon cross-validation when continuous items and criteria are employed with an appropriate sample. The implication here is that the weighted application blank procedure may be especially sensitive to capitalization on chance, and when this effect is partialed out, it yields results comparable to those obtained from the regression and correlational strategies. Unfortunately, because these studies did not systematically vary sample size, some ambiguities remain in this regard.

Both Webb (1960) and Malloy (1955) contrasted the weighted application blank approach with the deviant response procedure. In the initial keying sample, it was found that the deviant response procedure would add more to prediction than a standard empirical key. However, the Webb (1960) study suggests that the predictive power of deviant response keys may not hold up under cross-validation, due to the poor reliability of residual scores. Finally, Telenson et al. (1983) found that the vertical and horizontal methods yield similar results, and that the rare-response weighting procedure may occasionally produce significant validity coeffi-

cients. It should be noted here that the magnitude of the relationships obtained in this study was not great, and that application of rare-response weighting procedures may often be difficult to justify on the basis of content validity.

Reliability and validity. Relatively little information is available concerning the reliability of empirically keyed background data scales, although the available evidence for items keyed using the weighted application blank method is quite favorable. For instance, the ALPHA Biographical Inventory manual (Institute for Behavioral Research in Creativity, 1968) reports retest reliabilities of .82 to .88 on scales keyed for the prediction of creativity and academic performance. Similarly, Chaney (1964) administered a key containing 82 items and obtained a reliability of .85 over an 18-month interval. As might be expected, given the relative independence of background data items, internal consistency estimates for empirically keyed background data scales, such as split-half reliabilities, tend to be lower. For instance, Hinrichs, Haanperä, and Sonkin (1976) reported reliabilities of .65 to .78 on a 104-item form administered in five samples.

The question most frequently addressed with respect to empirical keys concerns their validity in predicting various criteria. Reviews of the validity of empirically keyed background data items may be found in Reilly and Chao (1982), Owens (1976), Asher (1972), Schuh (1967), and Henry (1966). Additionally, Ghiselli (1973) examined the criterion-related validity of background data measures with respect to other criteria and potential predictors. A summary of the data obtained in the Owens (1976) and Reilly and Chao (1982) reviews is given in Tables 1 and 2, covering the periods from 1940 to 1975 and from 1976 to 1982, respectively. It is of note that the Owens review made no attempt to control for cross-validation, while the Reilly and Chao review considered only cross-validated correlations.

Overall, the data presented in Tables 1 and 2 attest to the criterion-related validity of empirically keyed background data measures. In these two reviews, the data yielded validity coefficients of .30 to .50 against a variety of criteria in a number of different occupational fields. These observations agree with the findings of Asher (1972) and Ghiselli (1973). Moreover, the cross-validated results presented by Reilly and Chao (1982) contradict Schwab and Oliver's (1974) conclusion that the apparent validity of empirically keyed background data scales is due merely to capitalization on chance. Yet it should be recognized that the lower overall relationship reported by Reilly and Chao with respect to that reported by Owens (1976) may reflect the focus of the Reilly and Chao investigation on cross-validated coefficients. One other difference between the findings obtained in these two reviews should be mentioned; that is, the marked difference between the results Reilly and Chao reported for nonmanagement jobs and those Owens reported for factory workers. Here it seems that the smaller number of studies examined by Reilly and Chao and the methodological quality of these studies may account for the observed discrepancy.

Aside from these differences, two additional considerations are worthy of mention. First, when

Table 1
Criterion-Related Validity Coefficients (r) Reported by Owens (1976)

Criterion	r	N
Managerial performance	.35	21
Sales success	.35	17
Factory jobs	.46	14
Academic achievement	.45	14
Clerical performance	.48	13
Military criteria	.34	13
MBAs, leadership	*	
Credit risk	.62	3
Vehicle drivers	*	
Miscellaneous	*	
Overall	.42	112

Note. Data differ from Owens (1976) due to inclusion of academic performance and text citations along with elimination of a few additional studies.
N is number of studies.
*Data not appropriate due to analyses or methods.

Table 2
Cross-Validated Coefficients (r) Reported by Reilly and Chao (1982)

	Criterion											
	Tenure		Training		Ratings		Productivity		Salary		Average	
Occupation	r	N	r	N	r	N	r	N	r	N	r	N
Military	.30	3	.34	3	.25	3	*		*		.30	9
Clerical	.52	6	*		*		*		*		.52	6
Management	*		*		.40	4	*		.23	3	.38	7
Other non-management	.24	2	*		*		*		*		.14	2
Sales	*		*		.40	4	.62	1	*		.50	5
Scientific/Engineering	.50	2	*		.32	4	.43	5	.43	4	.41	15
Occupation Average	.32	13	.34	3	.36	15	.46	6	.37	7	.35	44

Note. N is number of studies.
*Data not appropriate or unavailable.

these results are considered in light of Ghiselli's (1973) findings, it is clear that empirically keyed background data measures are among the best available predictors of training and job performance criteria. Second, the results available in the broader literature suggest that empirical keys may be valid predictors of a variety of criteria not considered above, such as job satisfaction (Waters, Roach, & Waters, 1976), occupational choice (Albright & Glennon, 1961; Mumford, 1981), criminality (Frazier, 1981), absenteeism (Belohlau, 1982), personality attributes (Laughlin, 1984; Rawls & Rawls, 1968; Schuh, 1967), and the effects and use of clinical treatments (Carter & Fasullo, 1982; Porch, Collins, Wertz, & Friden, 1980). Although the general emphasis on industrial and academic criteria makes it difficult to obtain a comprehensive evaluation of the breadth of prediction, it appears that empirically keyed scales may have substantial value in addressing a variety of problems outside the area of industrial selection and placement.

Stability and generality. One question pertaining to the validity of empirical keys concerns the stability or generality of their predictive power over time, samples, cultures, and organizational environments. The issue of stability over samples has been addressed in the many discussions of cross-validation. As mentioned above, more recent investigations have taken this effect into account and have demonstrated that although there is some shrinkage in initial validity coefficients upon cross-validation, it is not sufficiently large to restrict application. Although the degree of shrinkage may vary with the particular keying system and sample in use, cross-validation of a well-developed key typically produces a 10-point loss in predictive power.

A few studies have also attempted to assess the generality of empirical keys across organizations. One study by Schmidt, Hunter, and Caplan (1981) found that the validity of an empirical key did not generalize across two petroleum industry jobs. In a related study, Brown (1981) examined situational moderation of an empirical key used in the selection of salesmen by 12 life insurance companies. He found that 38% of the variance in validity coefficients across companies was not artifactual, although the key displayed some validity in all 12 companies. This suggests that maximal prediction may require keying in the situation at hand, but also suggests that a key which is effective in predicting performance in one organizational setting is also likely to provide at least some prediction in other organizational settings.

Levine and Zachert (1951) obtained similar results in a study of the validity of five empirically keyed background data scales in 21 Air Force occupational specialties. Here it was found that a key designed to predict performance in mechanical occupations would yield significant validity coeffi-

cients for Machinists or Engineering Mechanics, while a key developed for electrical occupations would yield significant validity coefficients for Airplane Electrical Mechanics and Electricians. This suggests that the predictive power of empirical keys may generalize across groups of similar occupations when the initial criterion and reference groups are defined in a sufficiently broad manner (Levine & Zachert, 1951; Strong, 1941).

The generality of empirical keys across cultures has also been examined. For instance, Laurent (1970) attempted to revalidate a North American sales success key in three European countries. In the original United States sample a validity coefficient of .44 was obtained, while validity coefficients of .59, .61, and .55 were obtained in Norway, Denmark, and the Netherlands respectively. In a related study, Hinrichs et al. (1976) revalidated a sales success key developed in Finland and obtained coefficients of .38 to .72 in the United States, Sweden, Norway, France, and Portugal. Thus, within the same occupations in Western culture, empirically keyed background data measures appear to display some generality. At first glance this conclusion may seem somewhat surprising. Yet it is recognized that there is a great deal of cross-national consistency in occupational demands (Deeg & Paterson, 1947; Hall & Jones, 1950), and that significant ethnic group differences have not been obtained in test bias studies (Sparks, 1965). This conclusion would seem to agree with the broader literature.

Finally, the possibility of shifts in the nature of criterion and reference groups has engendered a concern with the stability of empirically keyed background data items over time (Thayer, 1977). Two studies have addressed this issue. Brown (1978) found that there was substantial stability in empirical keys designed to predict the tenure of life insurance salesmen over a 40-year interval. Similarly, Kirkpatrick (1968) found that an empirical key for predicting job performance maintained its validity over a 5-year interval, with only 4 of some 100 predictive items failing to revalidate. Overall, these studies indicate that under proper testing conditions, with well-developed items not markedly vulnerable to changing cultural influences, empirically keyed background scales may maintain their validity for some time.

Conclusions. There is a great deal of evidence concerning the characteristics of empirical keys that argues for their reliability and validity. However, two pervasive difficulties are associated with these scales. First, although the predictive power of empirical keys may display some generality over time, cultural groups, and occupational settings, it is apparent that their value is tied to the specific criteria employed in key development (Owens & Schoenfeldt, 1979). Because a large number of criterion performances are potentially of interest to psychometricians, this implies a lack of parsimony. Second, because empirical keys select items solely on the basis of their ability to predict complex criterion performances, item content tends to be complex and lack psychological meaningfulness (Guion, 1966). Taken together, these two observations imply that empirical keys may have limited value in more theoretically oriented measurement efforts. The following sections will examine some of the scaling techniques employed in attempts to ameliorate these difficulties.

Rational Scales

Approaches. One alternative to the traditional empirical key may be found in the various rational scaling procedures. Broadly speaking, two basic strategies have been employed in attempts to select items for inclusion in rational scales. For want of a better terminology, these strategies are referred to here as the direct and indirect approaches. In the direct approach, investigators obtain job analysis or basic descriptive information to define the behaviors related to differential performance (Myers & Fine, 1980; Schmidt, Caplan, Dunn, Bemis, Decuir, & Antone, 1979). Subsequently, items are developed to capture the expression of manifestly similar forms of behavior earlier in the individual's life. An attempt is made in item scaling to assess either the individual's quality of performance or his/her preference for these activities. The total number of activities engaged in, weighted by their assessed quality or preference, forms a basis for

describing the individual and predicting criterion performance (Myers & Fine, 1980). When the concern is quality, investigators typically use external evaluators to rate activity descriptions; when the concern is interest or liking, a standard self-report format is employed.

Like the direct approach, the indirect approach typically begins with an attempt to obtain information describing criterion performance. However, investigators subsequently summarize this information through the psychological constructs held to underlie performance, such as Fleishman's (1975) ability requirements taxonomy or Fine's (1974) functional job analysis system. Items and response options are then developed to reflect behavior or experiences believed to be capable of contributing to the development of these attributes (Mumford, Cooper, & Schemmer, 1983). Studies contrasting the relative effectiveness of the direct and indirect approach are not available. However, it should be apparent that the direct approach is likely to yield item content highly similar to manifest criterion performance, while items derived from the indirect approach may display only limited similarity to job behaviors. On the basis of this observation and the specified dimensional structure underlying the indirect approach, it appears that the direct approach will display greater face validity and that the indirect approach will display greater psychological meaningfulness or construct validity.

Traditionally, the selection of items for rational scales has been based on an internal consistency analysis. For instance, the procedure employed by DuBois, Loevinger, and Gleser (1952) involved grouping items into clusters on the basis of content similarity, and then intercorrelating these items. The three items yielding the highest within-cluster correlations are used to define a cluster, and the remaining items are added to these initial clusters on the basis of their correlations with the core items. This procedure is repeated in an iterative sequence until all items have been included in a cluster or dropped. A variation on this procedure, developed by Matteson, Osburn, and Sparks (1969), employs a computer algorithm to assign items to clusters so as to maximize cluster independence as well as homogeneity. Another approach to item selection on the basis of internal consistency was employed by Mumford, Cooper, and Schemmer (1983). Here all items were assigned to the underlying dimensions they were intended to capture, and on each dimension items failing to yield item-total correlations above .30 were eliminated. Finally, studies based on the direct approach, where external evaluators rate the quality of activity statements, have necessarily examined interrater agreement and have eliminated dimensions or statements yielding poor agreement.

Reliability and validity. In a study examining the predictive power of the indirect approach where items were grouped into homogeneous clusters, Matteson (1978) found that an optimal combination of 11 scales would yield multiple *R*s in predicting training performance, 6-month supervisory ratings, and job performance ratings of .41, .43, and .51, respectively, in a sample of production workers. Mumford, Cooper, and Schemmer (1983) found that a set of homogeneous rational scales yielded a cross-validated multiple *R* of .30 against performance ratings in a sample of law enforcement officers, while Turnage and Muchinsky (1984) found that such scales predicted management progress criteria roughly as well as assessment center ratings. Similarly, DuBois et al. (1952) showed that indirect rational scales display intercorrelations below .20, and are capable of predicting Air Force training criteria at the .30 level across a number of occupational fields.

Siegel (1956) reported that homogeneous clusters of background data items designed to predict academic performance are relatively independent. These scales were also found to yield internal consistency coefficients lying between .60 and .70, but produced 6-month retest reliability coefficients lying in the high .80s or low .90s. Overall, these studies suggest that rational scalings of homogeneous background data items will produce relatively independent scales displaying adequate reliability and substantial predictive validity. When these observations are coupled with the findings of Siegel (1956), Matteson (1978), and DuBois et al. (1952) pointing to the interpretability and psycho-

logical meaningfulness of the resulting dimensions, they appear to provide substantial evidence for the value of this approach.

The second line of investigation into the construction of rational background data scales derives from the behavioral consistency approach of Schmidt et al. (1979). Essentially, this approach specifies the major behavioral dimensions of job performance, and behaviorally anchored rating scales are developed for evaluating the quality of past activities falling under these dimensions. In operation, individuals are asked to write brief statements describing past activities relevant to each behavioral dimension. Judges then read these activity statements and evaluate the quality of the activities on the relevant behaviorally anchored rating scale. While the initial Schmidt et al. (1979) study did not provide detailed data concerning the psychometric characteristics of this technique, a later study by Hough (1984) provided some pertinent information in this regard. Briefly, her results indicate that these scales will produce reliabilities lying in the .70s and criterion-related validity coefficients in the mid-.20s against job performance rankings, job performance ratings, and task evaluations. She also found that they exhibited substantial construct validity. Thus, despite a need for further research, it appears that this approach is not without promise.

Comparative data. Sufficient information is not available to allow comparative statements concerning the generalizability of rational scales over time, culture, ethnic groups, and occupational settings. However, at least one study has examined the predictive efficiency of both rational and empirical scaling techniques. In this study, Berkeley (1953) found that (1) empirical keys showed higher criterion-related validities in the initial validation sample; (2) empirical keys showed greater shrinkage in cross-validation; (3) homogeneous rational scales and empirical keys yielded similar results in cross-validation; and (4) the homogeneous rational scales produced interpretable relationships while the empirical key did not. Similar results were observed by Hornick, James, and Jones (1977), who contrasted the utility of empirical and rational scaling procedures in the measurement of organizational climate.

Conclusions. It appears that the rational scaling of background data items presents a viable alternative to the traditional empirical keys with respect to criterion-related validity. It also seems that rational scaling procedures would prove especially attractive whenever content and construct validity are salient concerns. However, the number of studies cited above indicates that rational scaling procedures have not received much attention. Hence, there is a need for further research examining the fairness of such measures as well as the generalizability of their predictive power over time, culture, and organizational settings. Additionally, there would seem to be value in studies examining the knowledge requirements for effective rational scaling efforts, and the kinds of criterion performances that rational scales are best able to predict. Finally, the utility of these studies might be enhanced by contrasting the effectiveness of rational scales and alternative scaling techniques in this regard.

Factorial Scales

Procedures. Because of their close tie to a particular form of criterion performance, new rational scales and empirical keys must be built for each form of criterion performance of interest to some investigator. While this strategy may optimize prediction in a particular situation, the lack of parsimony inherent in these criterion-specific scaling procedures limits researchers' efficiency and ability to develop general understandings. As a result, there has been some interest in employing factor-analytic techniques to identify psychologically meaningful summary dimensions. Here exploratory factor-analytic or principal components procedures are employed to identify those solutions that will yield the smallest number of dimensions accounting for the largest proportion of item variance. Application of the criteria commonly yields two or three potential solutions when background data measures are in use (Mumford, 1986). Each of these solutions is then subjected to rotation, generally a varimax procedure, and the rotated solutions are evaluated with respect to simple structure

and psychological meaningfulness. The solution that appears to optimize both these criteria is generally retained as the basis for scale development.

When employing this final dimensional structure in scaling, items yielding loadings below .30 on any given dimension are eliminated. Subsequently, dimensional scales are formulated through one of two basic procedures. First, the scoring coefficient matrix associated with the solution may be used to provide item weights. Second, items yielding loadings above .30 may be assigned unit weights in accordance with the direction of the loading.

Solution characteristics. It is unusual to find factor-analytic solutions that account for a large proportion of the total item variance, due to (1) the internal consistency of item data, (2) the relative independence of background data items, and (3) the heterogeneity of most item pools. Typically, retained solutions account for 30% to 50% of the total variance (Baehr & Williams, 1967; Mumford, 1986; Owens, 1976), although even with these percent variance figures, 70% to 90% of the items commonly yield loadings above .30 on one or more dimensions. As might be expected given the independence of homogeneous scales, Baehr and Williams (1967) found that orthogonal and oblique solutions yield roughly equivalent results.

Even given investigators' preference for orthogonal component analyses, background data factorings are not especially likely to yield general factors. Because some analyses have identified general factors (Mumford, 1986) while others have not (Owens, 1976), it appears that the emergence of general factors depends on the particular item base in use. Therefore, it may not be possible to characterize background data or life history in terms of a hierarchical model. A final characteristic of these factor analyses is that dimensions are generally found to be defined by items of homogeneous content; for example, items reflecting attitude toward school will typically produce high loadings on one dimension while items reflecting athletic involvement will fall together under another dimension (Owens, 1976).

Stability and generality. Beyond these basic findings, a number of studies have attempted to examine the stability or generality of the dimensional structures obtained in factor or component analyses of background data items. Mumford, Shaffer, Jackson, et al. (1983) contrasted the dimensional structure obtained in three item pools when male and female item responses were factored separately and when they were factored together. It was found that a joint factoring of men's and women's item responses reduced the percentage of total variance accounted for, and yielded dimensions that proved more difficult to interpret than those obtained in separate male and female analyses. It was also found that there were marked differences in the content and nature of the dimensions emerging in the various male and female analyses. These findings indicate that it may not be appropriate to apply a common dimensional structure in describing the life history of men and women.

Gonter (1979) examined the potential differences in dimensional structure across ethnic groups. He obtained similar dimensional structures for black and white females but not for black and white males. On the other hand, a study by Cassens (1966) argues for the stability of dimensional structures across Western cultural groups. Cassens factor-analyzed a 62-item background data scale in a sample of 105 United States managers, 74 United States managers working in Latin America, and 382 Latin Americans working in five countries. He found that of the 10 factors identified, 9 emerged in all three samples. Although further research is needed, these findings suggest that among cultural groups having similar life histories, similar dimensional structures are likely to emerge, but when cultural and economic differences are pervasive, different factor structures may well be observed.

A few studies have examined the stability of dimensional structures over time, samples, and age groups. For instance, the results reported by Owens and Schoenfeldt (1979) indicate that the predictive characteristics of orthogonal principal components, derived in samples of some 1,000 males and females, are sufficiently stable to be maintained in independent samples with no more than a 10% loss in predictive power. Further, Mumford and Owens

(1984) did not observe cohort effects on the components derived in an analysis of three background data questionnaires administered to seven cohorts over a 10-year interval. Finally, Ames (1983) found that, even with the addition of a substantial number of new items to the pool used by Owens and Schoenfeldt, roughly the same male and female dimensions would emerge in an orthogonal components analysis. These findings provide a compelling argument for the stability of the dimensional structures obtained in components analyses of background data items across samples of individuals over time. Nevertheless, it should be noted that data pertaining to other analytic techniques are lacking, and that there is a need for more formal confirmatory studies.

Two studies have examined the stability of the dimensional structures obtained from background data items over substantial periods of time. Eberhardt and Muchinsky (1982) readministered the original Owens and Schoenfeldt (1979) background data form to a sample of midwestern freshmen some 10 years later and carried out another orthogonal components analysis. They obtained roughly the same dimensional structure in the male sample, but found a markedly different dimensional structure in the female sample. Because the nature of these shifts indicated increasing similarity among men and women, they concluded that changes in the nature of women's roles during this period had induced some instability in the dimensional structure. While this seems a reasonable interpretation, it should be noted that a study by Lautenschlager and Shaffer (in press) employing data gathered 8 years after the Owens and Schoenfeldt study in the same southeastern population failed to find any shifts in dimensional structure. It is possible that the use of different item pools, sample differences, or an interaction between sample characteristics and cultural change might account for this discrepancy. As a result, it seems that any firm conclusions concerning the dimensional stability of background data items over long periods of time must await further studies controlling for sample differences and employing more sophisticated analytic techniques.

Another set of studies has examined age effects on the dimensional structures obtained in factorings of background data items. One study by Mumford, Shaffer, Ames, and Owens (1983) examined the stability of the male and female dimensional structure produced by a common item pool in two age groups known to be facing different developmental tasks. Highly different dimensional structures were obtained in these two groups among both men and women. Further, attempts to apply a common dimensional structure resulted in substantial losses with respect to the percentage of variance accounted for and the interpretability of the dimensional structure. Thus, it appears that when the age grading of behavior and experiences has resulted in marked shifts in the nature of individuals' lives, different dimensional structures will be required in different age groups. This conclusion was supported in a study by Mumford (1986), who found that very different kinds of background data dimensions emerged from item pools tailored to capture developmentally significant behaviors and experiences occurring in adolescence, youth, and young adulthood. Moreover, it was found that the background data dimensions required to describe individuals over all three periods were qualitatively different from those required to describe individuals within each of these developmental periods.

Dimensional content. Overall, 21 factor and principal components analyses of background data items were found in the literature. To summarize these data, the content of each dimension was reviewed, and a judgment was made as to which of 26 generic, conceptual dimensions (such as Social Leadership or Parental Warmth) was best reflected by each dimension. Although an attempt was made to assign each factor to only one of these generic categories, if a dimension appeared to capture significant elements of two or more categories it was considered a member of each one.

Table 3 presents the results of this analysis. As may be seen, dimensions concerned with General Adjustment, Maturity, Academic Achievement, and Social Effectiveness were especially likely to emerge. Dimensions concerned with Independence, Self-Esteem, Achievement Motivation, Conformity, and

Table 3
Number of Factor Analytic Studies (N=21)
Identifying Generic Dimensions

Dimension		Number
1.	Introversion vs. Extroversion	13
2.	Social Leadership	12
3.	Independence	8
4.	Self Esteem	7
5.	Achievement Motivation	7
6.	Social Conformity	8
7.	Personal Conservatism	6
8.	Maturity	11
9.	Adjustment	18
10.	Academic Achievement	16
11.	Parental Warmth	7
12.	Parental Control	6
13.	Parental SES	8
14.	Sibling Relationships	3
15.	Religious Involvement	9
16.	Family Commitment	4
17.	Urban/Rural Background	1
18.	Health	3
19.	Athletic Pursuits	8
20.	Scientific/Engineering Pursuits	7
21.	Intellectual/Cultural Pursuits	13
22.	Work Values	4
23.	Organizational Commitment	6
24.	Professional Skills	7
25.	Trade Skills	5
26.	Career Development	10

Conservatism also appeared relatively frequently, as did dimensions reflecting Religious Beliefs, Athletic Involvement, Scientific Interests, and Cultural Interests. Parental Warmth and Parental Socioeconomic Status also were identified as being important dimensions in a number of studies, along with dimensions reflecting Career Development, Professional Skills, Trade Skills, and Organizational Commitment. Overall, it appears that dimensions concerned with social, occupational, and academic effectiveness are those most likely to be identified in factor or principal components analyses of background data items, along with dimensions related to their development and expression, such as Parental Warmth, Independence, and Adjustment.

Reliability and validity. Studies by both Baehr and Williams (1967) and Owens (1976) obtained internal consistency coefficients lying in the mid-.70s, as might be expected given the relative independence of background data items. On the other hand, in an examination of short interval retest reliability, Owens (1976) obtained an average retest reliability near .90 over some 30 dimensional scales. Thus it appears that factorial scalings of background data items may yield measures displaying high reliability.

In examining the validity of factorial scales, it rapidly becomes apparent that the content and construct validity of these scales has not received sufficient attention in the literature. However, some compelling support for the criterion-related validity of five factorial background data scales has been

obtained by Morrison, Owens, Glennon, and Albright (1962). They found that these scales could yield significant prediction of creativity and performance criteria in a sample of 250 processing and heavy equipment operators. Cross-validated multiple *R*s of .35 and .53 were obtained from a weighted combination of scores on these scales. Similarly, Vandeventer, Taylor, Collins, and Boone (1983) found that scores on background data factors would predict success in Air Traffic Controller's school, while Engdahl (1980) found that they would predict occupational values. A series of studies by Lucchesi (1984), Neiner and Owens (1982), and Mumford (1986) found that scores on background data factors derived in earlier developmental periods would predict individuals' scores on background data factors derived in later developmental periods. Moreover, the Neiner and Owens study indicated that maximum prediction could be obtained only through an optimal combination of factors identified in each developmental period, while the Mumford (1986) study indicated that the effectiveness of factor scores in prediction would decrease as the time interval increased.

Obviously, the small number of studies cited above indicates a need for additional research examining the predictive power of factorial background data scales. On the other hand, the positive findings obtained in these studies might cause curiosity about how factorial scales compare with rational scales and empirical keys. While no studies are available explicitly contrasting factorial and rational scales, Mitchell and Klimoski (1982) contrasted the predictive power of empirical keys and factorial scales. It was found that the empirical keys predicted licensure of real estate salesmen better than a weighted combination of six factorial scales in both the validation and cross-validation samples. However, the factorial scales showed less shrinkage in cross-validation. This suggests that the factorial scales may display greater generality and stability but may not yield the predictive power equivalent to empirical keys. This conclusion finds some support in a comparison of the results obtained in the Morrison et al. (1962) factorial study of innovation among petroleum scientists with the results obtained in the Smith, Albright, Glennon, and Owens (1961) study of empirical keys in predicting the same criteria in the same sample.

Conclusions. Factorial scales may not be as effective as empirical keys with respect to criterion-related validity. Nevertheless, they may display greater stability and generality. Further, it appears that, like rational scales, the factorial scales can yield psychologically meaningful summary dimensions. Thus, when a reasonably general and stable scaling of life history items is to be constructed, there would seem to be some value in the factorial approach. This seems especially true when the concern is the specification of basic life history dimensions for theoretical purposes. However, when the concern is prediction in a specific situation, investigators are likely to find empirical keying a more attractive scaling strategy.

Subgrouping

Procedures. While the evidence examined earlier indicates that well-constructed empirical keys yield excellent prediction, it was also noted that the predictive power of empirical keys is tied to a specific form of criterion performance in a certain population. When investigators must predict performance on a number of criteria across a variety of situations or populations, the need to develop separate keys may become burdensome, thereby prohibiting the application of background data measures. Recognition of this problem led Owens (1976) to argue that techniques are required for constructing more general descriptive systems. Because empirically keyed background data measures are based on quantitative definition of the developmental pattern underlying a certain kind of criterion performance, Owens (1968, 1971, 1976) concluded that a more general descriptive system might be formulated by identifying modal patterns of individual development without reference to a particular criterion measure.

The basic methodology developed by Owens (1976) and Owens and Schoenfeldt (1979) to address this issue begins with the construction of a comprehensive set of background data items ca-

pable of capturing the developmentally significant behavior and experiences occurring in a certain developmental period. These items are then administered to a sample of 200 or more individuals. Item responses are subsequently summarized through an orthogonal components analysis carried out separately for men and women. An individual's profile of scores on each of the relevant components is then obtained, and the similarity between each person's profile and those of all other members of his/her gender group is assessed through the d^2 measure of profile similarity. These similarity data are then entered into a modified Ward and Hook (1963) cluster analysis to identify subgroups of similar individuals and the number of individuals who can be assigned to each subgroup (Feild & Schoenfeldt, 1975). Finally, the mean differences observed between subgroup members and the sample as a whole on the background data items and components are obtained. This information is then used to describe the subgroup members, and to generate a brief consensual label capable of summarizing each subgroup's differential characteristics.

In the first implementation of this paradigm, Owens (1976) and Owens and Schoenfeldt (1979) administered a comprehensive set of 389 background data items concerned with significant behavior and experiences during childhood and adolescence to approximately 1,000 male and 1,000 female college freshmen. Item responses were factored and subgrouped in accordance with the procedures outlined above. This analysis yielded 23 male and 15 female subgroups. It was found that 73% of these individuals could be fitted to a single subgroup, while 20% of the sample consisted of "overlaps" who could be assigned to two or more subgroups and 7% were "isolates" who could not be assigned to any subgroup. These figures indicated that the classification structure was capable of describing most individuals, although the range of individuals fitted to a single subgroup indicated that some developmental pathways were followed more frequently than others. Examination of the subgroups' status on the background data items and components indicated that the subgroups displayed differences on a wide variety of indicators that were sufficiently cohesive and meaningful to allow each of these subgroups to be assigned a brief consensual label.

Of course, this utility of the subgrouping approach is contingent on its ability to capture meaningful developmental patterns that extend beyond those measures employed in initial subgrouping. This issue was also addressed by Owens and Schoenfeldt (1979). In this study a variety of standard psychometric measures, such as Rotter's I/E scale and the California F scale, were administered along with the background data items; these measures were found to yield a pattern of subgroup differences consistent with that observed on the background data measures. Additionally, an extensive set of field studies was conducted to determine whether subgroup status would predict performance in a variety of situations. In a series of some 50 studies, significant differences were observed between the subgroups in 80% of these investigations. Moreover, subgroup status was found to predict a remarkably wide range of criterion performances, such as academic over- and underachievement (Klein, 1970), vocational interests (Jones, 1970; Thomas, 1982), and drug use (Strimbu & Schoenfeldt, 1972). When these results were broken down by content areas, such as personality, social behavior, physiology, learning, motivation, and methodology, it was found that subgroup status was an effective predictor of performance in roughly 90% of the studies, excepting basic processes such as physiology and learning.

Stability and generality. A number of technical issues concerning application of this technique have arisen, one of which is the stability and generality of these developmental patterns. In part, this issue was addressed in the Owens and Schoenfeldt (1979) study, which indicated that the members of six freshman classes entering the university between 1970 and 1977 could be fitted into one of the initial subgroups with no more than a 10% loss in the number of individuals fitted to a single subgroup. At a broader level, Anderson (1972) readministered Owens' background data questionnaire at two southeastern universities and one midwestern university. She found that the same subgroup structure

could be applied in all three samples, although the proportion of individuals assigned to each subgroup might vary. Overall, then, the available evidence tends to argue for the stability of subgroup descriptions over time and their generalizability to new samples. On the other hand, it should be recognized that the Owens and Schoenfeldt (1979) study was not intended to provide a comprehensive definition of all subgroups, and that shifting developmental patterns could lead to changes in the nature of the resulting subgroups over substantial periods of time.

Some attempts have also been made to examine the generality of subgroup structures across various subpopulations. Mumford, Shaffer, Jackson, et al. (1983) contrasted a common subgrouping with separate male and female subgroupings on three separate background data questionnaires. They found that attempts to combine the male and female samples in subgrouping resulted in (1) a 20% loss in the number of individuals assigned to a single subgroup and (2) developmental patterns of limited interpretability. These results indicated that qualitative differences prohibit describing men and women in terms of a common set of developmental pathways. Although further research is needed, particularly studies focusing on ethnic and cross-cultural differences, these results suggested that the developmental patterns identified in a subgrouping should not be arbitrarily assumed to generalize to different subpopulations.

One obvious limitation of the research discussed above is that it has examined only individuals in late adolescence, in terms of a single item pool. Consequently, it would seem necessary to determine whether this approach can be extended to other item pools and other developmental periods. These issues have been examined in studies conducted by Mumford and Owens (1984) and Jackson, Mumford, Shaffer, and Owens (1986). In these studies, a 58-item questionnaire concerned with significant behavior and experiences during the college years and two 118-item questionnaires concerned with significant behavior and experiences during the first 2 to 4 and 5 to 10 years following graduation were administered to members of the original Owens and Schoenfeldt (1979) sample. Examination of the resulting subgrouping solutions indicated that effective solutions, in terms of interpretability and the number of individuals fitted to a single subgroup, could be obtained in later developmental periods. However, in terms of membership and content, there did not appear to be a point-to-point relationship between these earlier and later subgroups. Rather, it was found that there were systematic patterns of movement between the subgroups identified in earlier and later developmental periods, such that the members of a given adolescent subgroup were likely to enter only 2 or 3 of some 15 collegiate subgroups.

The existence of these systematic relationships led Mumford and Owens (1984) to investigate the feasibility of formulating developmental patterns extending across these developmental periods. To accomplish this, 417 males and 358 females who had responded to the adolescent and collegiate questionnaires and to a later post-college questionnaire were identified. Responses to the items contained in all three of these instruments were then factored by gender, and individuals were clustered on the basis of their factor scores. It was found that 84% of the males and 87% of the females could be fitted to a single one of the 15 male and 17 female subgroups. Examination of the posterior probabilities of subgroup membership generated by a discriminant analysis indicated that these subgroups were virtually independent and satisfied simple structure criteria. A review of relevant mean differences also indicated that they captured interpretable developmental patterns associated with meaningful differences in all three developmental periods.

Because the Mumford and Owens (1984) study formulated subgroups capturing antecedent/consequent relationships emerging over time, they provided an opportunity for examining the nature and meaning of subgroup membership. In this regard four points seem worthy of consideration. First, it was found that the subgroups displayed qualitative differences in their organization of behavior and experiences. Second, it appeared that subgroup membership and the developmental history it entailed would moderate the predictive implications

of various behaviors and experiences (Harding & Bottenberg, 1961; Owens, 1978). Third, while specific behavior and experiences changed over time, the subgroups appeared to capture an integrated pattern of behavior and experiences by which individuals attempted to adapt to their environment. Fourth, the predictive meaning of underlying differential characteristics often varied with the broader developmental context in which it was embedded.

Reliability and validity. Of the studies examining the reliability of subgroup assignments which have been conducted to date, only a few have examined the predictive power of life history subgroups against real-world criteria. Broadly speaking, these studies fall into two categories, as distinguished by their concern with educational as opposed to occupational criteria.

Two of the more significant studies in the educational arena are Schoenfeldt's (1972, 1974) studies of college dropout and freshman GPA. Using the initial Owens and Schoenfeldt (1979) subgroups as predictors, it was found that freshman GPA could be predicted at the .60 level, while dropout GPA could be predicted at the .50 level. Similarly, Lissitz and Schoenfeldt (1974) found that subgroup status would predict freshman GPA roughly as well as SAT scores, but that adding subgroup status to SAT scores produced a 10- to 15-point gain in prediction. Klein (1970) and Boardman, Calhoun, and Schiel (1972) showed that status on the initial Owens and Schoenfeldt subgroups would also predict academic over- and underachievement along with movement into campus leadership positions. In a somewhat different vein, Feild (1973) and Schoenfeldt (1974) showed that membership in the Owens and Schoenfeldt subgroups was a highly effective predictor of college major, while Ames (1983) showed that status on a different set of subgroups would predict both GPA and field of specialization in veterinary school. Finally, Feild, Lissitz, and Schoenfeldt (1975) found that although adolescent factor scores predict collegiate factor scores better than adolescent subgroup status, adolescent subgroup status would yield significant increments in prediction beyond that obtained from factor scores alone. This suggests that the subgroups captured some unique descriptive information.

The results obtained in the prediction of occupational criteria have also been promising. For instance, Taylor (1968) found that 84% of the successful salesmen in a sample of 222 came from 3 of the 9 subgroups identified on a clustering analysis of Form B of the Richardson, Bellows, and Henry application blank. In a similar Ward and Hook clustering, Pinto (1970) found that subgroup status would predict sales performance at the .60 level. As part of another line of investigation, Brush and Owens (1979) have shown that subgroup status was strongly related to entry into certain occupational categories within an organization, while Neiner and Owens (1985) and Eberhardt and Muchinsky (1984) found that subgroup status at age 18 was related to vocational interests and eventual vocational choice after college graduation as indexed by Holland's (1973) typology. More generally, Meacham, Shaffer, and Owens (in preparation) found that status on the Owens and Schoenfeldt (1979) subgroups was related to eventual entry into various occupational categories derived from the Position Analysis Questionnaire.

Conclusions. Overall, the studies discussed above indicate that membership in background data subgroups is an effective predictor of a variety of educational and occupational criteria, and thus may well provide the general predictive system that was desired. Further, the available evidence indicates that subgroup status may be as effective a predictor as well-developed ability tests and background data factors, while often yielding significant increments in prediction when added to these measures in a regression analysis. Although there have been few studies contrasting the predictive efficiency of subgroup status with rational scales and empirical keys, these results are promising. Nevertheless, the utility of the subgrouping approach is likely to be limited by the fact that it requires a sizable sample for implementation, and the fact that substantial utility gains will be observed only when these subgroups are used in a variety of different predictive situations. Of course, this suggests that there is a compelling need for a wider range of validation studies. Moreover, the exploratory nature of this

research indicates the need for some additional methodological work. However, the general predictive power evidenced by the subgrouping approach under these limitations suggests that it may someday provide one of the field's best available measurement techniques.

Discussion and Conclusions

In at least one sense, this review may be said to have reached a hopeful conclusion. Traditionally, background data measures have relied on empirical keying techniques. However, during the last decade a number of alternative techniques for scaling background data items have appeared in the literature. Although empirically keyed background data scales are still the most effective predictors of a particular criterion performance, each of these alternative techniques is associated with its own set of strengths and weaknesses. For instance, the predictive power of empirical keys is bought at the price of a close tie to criterion performance in a particular sample. Background data subgroups, on the other hand, appear to be reasonably effective predictors of a wide range of criterion performances, yet they require a sizable sample for scale development. Factorial scales seem to have some value in gaining an understanding of the basic dimensions underlying item responses. However, the practical value of factorial scaling may be limited by sample size requirements, population-specific effects, and predictive power. Finally, rational scales offer investigators the advantage of a close tie between item content and the requirements for adequate criterion performance, thereby ensuring the content validity and interpretability of background data measures. Yet it is also true that this technique requires a sophisticated understanding of the developmental antecedents of criterion performance, which is lacking in many fields.

Although the availability of these alternatives to the traditional empirical keying strategy represents an important advance, it is clear that these scaling techniques have not received sufficient attention in the research literature. In this regard three major categories of research studies seem to be called for. First, comparative studies examining the predictive characteristics of these new scaling techniques in a variety of criterion settings are required. Second, there would seem to be some value in research examining the psychometric properties of these scales, particularly with respect to their stability over time and their generality across subpopulations. Third, more detailed methodological research capable of delineating the item selection and weighting techniques most likely to enhance the effectiveness of these scaling procedures is needed.

While the need for further research is most obvious in reviewing the literature pertaining to these alternative scaling techniques, it should also be recognized that there are certain issues in the development and application of empirical keys that bear closer examination. For instance, it appears that the definition of criterion and reference groups constitutes an important determinant of the success of empirical keying efforts. Despite the recognized importance of content and construct validity in the design and application of any measurement system, these issues have received scant attention in the empirical keying literature. In part, this problem may be traced to the fact that most investigators have not paid sufficient attention to the need to engage in systematic item development efforts prior to the outset of empirical keying efforts.

In the course of this review, it was argued that the effective application of any background data measure depends on the quantitative definition of a developmental pattern reflecting antecedents of performance on the criteria of interest. The existence of this communality in the nature of all background data measures led to the conclusion that substantially more attention must be given to the development and application of systematic item specification procedures. More explicitly, it was argued that construction of any background data measure should begin with definition of a domain of antecedent behavior and experiences, which will then provide a basis for item development efforts. As was noted at the outset of this paper, such efforts may do much to enhance the content, construct, and criterion-related validity of background data measures.

It is hoped that this paper will serve as a useful guide in future research efforts intended to address these issues. However, it seems likely that this will come to pass only when researchers recognize that the effective application of background data measures is not simply a matter of the empirical keying of application blank responses. Rather, the full potential of this measurement format will be realized only when investigators attend to the varied applications of background data scales and carefully consider the many issues arising in the development and scaling of background data items with respect to the fundamental principles of psychometric science.

References

Albright, L. E., & Glennon, J. R. (1961). Personal history correlates of physical scientists' career aspirations. *Journal of Applied Psychology, 45,* 281–284.

Ames, S. D. (1983). *Prediction of research vs. applied interests in veterinarians: A biographical approach.* Unpublished master's thesis, University of Georgia, Athens GA.

Anderson, B. B. (1972). *An inter-institutional comparison on dimensions of student development: A step toward the goal of a comprehensive developmental-integrative model of human behavior.* Unpublished doctoral dissertation, University of Georgia, Athens GA.

Asher, J. J. (1972). The biographical item: Can it be improved? *Personnel Psychology, 25,* 251–269.

Baehr, M. E. (1976). *National validation of a selection test battery for male transit bus operators.* Washington DC: U.S. Department of Commerce, National Technical Information Service.

Baehr, M., & Williams, G. B. (1967). Underlying dimensions of personal background data and their relationship to occupational classification. *Journal of Applied Psychology, 51,* 481–490.

Belohlau, J. A. (1982). Absenteeism in a low status work environment. *Academy of Management Journal, 25,* 677–683.

Berkeley, M. H. (1953). *A comparison between the empirical and rational approaches for keying a heterogeneous test* (Bulletin No. 53, p. 54). Lackland Air Force Base TX: USAF Human Resources Research Center.

Boardman, W. K., Calhoun, L. G., & Schiel, J. H. (1972). Life experience patterns and development of college leadership roles. *Psychological Reports, 31,* 333–334.

Brodie, W. M., Owens, W. A., & Britt, M. F. (1968). *Annotated bibliography on biographical data.* Greensboro NC: Creativity Research Institute, The Richardson Foundation.

Bronson, W. C., Katten, E. S., & Livson, N. (1959). Patterns of authority and affection in two generations. *Journal of Abnormal and Social Psychology, 58,* 143–152.

Brown, S. H. (1978). Long-term validity of a personal history item scoring procedure. *Journal of Applied Psychology, 63,* 673–676.

Brown, S. H. (1981). Validity generalization and situational moderation in the life insurance industry. *Journal of Applied Psychology, 66,* 664–670.

Brush, D. H., & Owens, W. A. (1979). Implementation and evaluation of an assessment classification model for manpower allocation. *Personnel Psychology, 32,* 369–383.

Bunch, M. B. (1974). *Retest reliability of the University of Georgia biographical questionnaire.* Unpublished master's thesis, University of Georgia, Athens GA.

Carter, J. A., & Fasullo, B. B. (1982). A study of student utilization behavior at an urban university health center. *College Student Journal, 16,* 343–347.

Cascio, W. F. (1975). Accuracy of verifiable biographical information blank responses. *Journal of Applied Psychology, 60,* 767–769.

Cascio, W. F. (1976). Turnover, biographical data, and fair employment practice. *Journal of Applied Psychology, 61,* 576–580.

Cassens, F. P. (1966). *Cross-cultural dimensions of executive life history antecedents.* Greensboro NC: Creativity Research Institute, The Richardson Foundation.

Chaney, F. B. (1964). *The life history antecedents of selected vocational interests.* Unpublished doctoral dissertation, Purdue University, West Lafayette IN.

Cherry, R. L. (1969). *Socioeconomic level and race as biographical moderators.* Unpublished doctoral dissertation, Ohio State University, Columbus, OH.

Clark, K. E., & Gee, H. (1954). Selecting items for interest inventory keys. *Journal of Applied Psychology, 38,* 12–17.

Cohen, J., & Lefkowitz, J. (1974). Development of a biographical inventory blank to predict faking on personality tests. *Journal of Applied Psychology, 59,* 404–405.

Cronbach, L. J. (1971). Test validation. In R. E. Thorndike (Ed.), *Educational measurement.* New York: American Council on Education.

Dailey, C. A. (1960). The life history as a criterion of assessment. *Journal of Counseling Psychology, 7,* 20–23.

Deeg, M. E., & Paterson, D. G. (1947). Changes in the social states of occupations. *Occupations, 25,* 205–

208.

DuBois, P. H., Loevinger, J., & Gleser, G. C. (1952). *The construction of homogeneous keys for a biographical inventory*. San Antonio TX: USAF Human Resources Research Laboratory.

Eberhardt, B. J., & Muchinsky, P. M. (1982). An empirical investigation of the factor stability of Owens' biographical questionnaire. *Journal of Applied Psychology, 67,* 138–145.

Eberhardt, B. J., & Muchinsky, P. M. (1984). Structural validation of Holland's hexagonal model: Vocational classification through the use of Holland's model. *Journal of Applied Psychology, 69,* 174–181.

Engdahl, B. E. (1980). *The structure of biographical data and its relationship to needs and values*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

England, G. W. (1961). *Development and use of weighted application blanks*. Dubuque IA: William C. Brown.

Federico, S. M., Federico, P. A., & Lundquist, G. W. (1976). Predicting women's turnover as a function of extent of net salary expectation and biodemographic data. *Personnel Psychology, 29,* 559–566.

Feild, H. S. (1973). *Subgroup and individual differences in the quasi-actuarial assessment of behavior: A longitudinal study*. Unpublished doctoral dissertation, University of Georgia, Athens GA.

Feild, H. S., Lissitz, R. W., & Schoenfeldt, L. F. (1975). The utility of homogeneous subgroups and individual information in prediction. *Multivariate Behavioral Research, 10,* 449–462.

Feild, H. S., & Schoenfeldt, L. F. (1975). Ward and Hook revisited: A two-part procedure for overcoming a deficiency in the grouping of persons. *Educational and Psychological Measurement, 35,* 171–173.

Ferguson, L. W. (1962). *The heritage of industrial psychology*. Hartford CT: Author.

Ferguson, L. W. (1967). Economic maturity. *Personnel Journal, 46,* 22–26.

Fine, S. A. (1974). Functional job analysis: An approach to a technology for manpower planning. *Personnel Psychology, 27,* 843–858.

Fleishman, E. A. (1975). Toward a taxonomy of human performance. *American Psychologist, 30,* 1127–1149.

Fleishman, E. A., & Quaintance, M. K. (1984). *Taxonomies of human performance: The description of human tasks*. New York: Academic Press.

Frazier, C. E. (1981). Personal documents and life histories in criminological research and practice: An exploration. *Case Analysis, 1,* 265–290.

French, N. R., Lewis, M. A., & Long, R. E. (1976, March). *Development and validation of a methodology to assess social desirability responding in a multiple choice biographical questionnaire*. Paper presented at the meeting of the Southeastern Psychological Association, New Orleans.

Galton, F. (1883). *Inquiries into human faculty and its development*. London: MacMillan & Co.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26,* 461–477.

Goldsmith, D. B. (1922). The use of the personal history blank as a salesmanship test. *Journal of Applied Psychology, 6,* 149–155.

Goldstein, I. L. (1971). The application blank: How honest are the responses? *Journal of Applied Psychology, 55,* 491–492.

Gonter, D. (1979). *Comparison of blacks and whites on background data measures*. Athens GA: Institute for Behavioral Research.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Guilford, J. P., & Lacey, J. I. (1947). Printed classification tests. *AAF Aviation Psychology Research Program Reports*. Washington DC: U.S. Government Printing Office.

Guion, R. M. (1966). *Personnel testing*. New York: McGraw-Hill.

Hall, J., & Jones, D. C. (1950). Social grading of occupations. *British Journal of Sociology, 1,* 31–35.

Harding, F. D., & Bottenberg, R. A. (1961). Effect of personal characteristics on relationships between attitudes and job performance. *Journal of Applied Psychology, 45,* 428–430.

Henry, E. R. (1966). *Research conference on the use of autobiographical data as psychological predictors*. Greensboro NC: Creativity Research Institute.

Hinrichs, J. R., Haanperä, S., & Sonkin, L. (1976). Validity of a biographical information blank across national boundaries. *Personnel Psychology, 29,* 417–421.

Holland, J. C. (1973). *Making vocational choices: A theory of careers*. Englewood Cliffs NJ: Prentice-Hall.

Hornick, C. W., James, L. R., & Jones, A. P. (1977). Empirical item keying versus a rational approach to analyzing a psychological climate questionnaire. *Applied Psychological Measurement, 1,* 489–500.

Hough, L. M. (1984). Development and evaluation of the accomplishment record method of selecting and promoting professionals. *Journal of Applied Psychology, 69,* 135–146.

Institute for Behavioral Research in Creativity (1968). *Manual for Alpha biographical inventory*. Greensboro NC: Predictions Press.

Jackson, K. E., Mumford, M. D., Shaffer, G. S., & Owens, W. A. (1986). *The durability of a classification of persons*. Manuscript submitted for publication.

Jones, E. L. (1970). *The affinity of biodata subgroups for vocational interest*. Paper presented at the meeting

of the Georgia Psychological Association, Atlanta.

Karla, S. K. (1981). Differences in recall of adolescent socialization among high and low achievers in business. *Journal of Applied Psychology, 66,* 562–568.

Kavanagh, M. J., & York, D. R. (1972). Biographical correlates of middle managers' performance. *Personnel Psychology, 25,* 319–332.

Keating, E., Paterson, D. G., & Stone, C. H. (1950). Validity of work histories obtained by interview. *Journal of Applied Psychology, 34,* 1–5.

Kenagy, H. G., & Yoakum, C. S. (1925). *The selection and training of salesmen.* New York: McGraw-Hill.

Kirkpatrick, J. J. (1968). A second cross-validation of an executive weighted application blank. *Proceedings of the 76th Annual Convention of the American Psychological Association, 3,* 581–582.

Klein, H. A. (1970). *Personality characteristics of discrepant academic achievers.* Unpublished doctoral dissertation, University of Georgia, Athens GA.

Klein, S. P., & Owens, W. A. (1965). Faking of a scored life history as a function of criterion objectivity. *Journal of Applied Psychology, 49,* 451–454.

Kurtz, A. K. (1941). Recent research in the selection of life insurance salesmen. *Journal of Applied Psychology, 25,* 11–17.

Larson, R. H., Swarthout, D. M., & Wickert, F. R. (1967). *Objectionability and fakeability of biographical inventory items.* Paper presented at the meeting of the Midwestern Psychological Association, Chicago.

Laughlin, A. (1984). Teacher stress in an Australian setting: The role of biographical mediators. *Educational Studies, 10,* 7–22.

Laurent, H. (1962). Early identification of management talent. *Management Record, 24,* 33–38.

Laurent, H. (1970). Cross-cultural cross-validation of empirically validated tests. *Journal of Applied Psychology, 54,* 417–423.

Lautenschlager, G. (1985). Within subject measures for the assessment of individual differences in faking. *Educational and Psychological Measurement, 46,* 309–316.

Lautenschlager, G., & Shaffer, G. S. (in press). Re-examining the component stability of Owens Biographical Questionnaire. *Journal of Applied Psychology.*

Lecznar, W. B., & Dailey, J. T. (1950). Keying biographical inventories in classification test batteries. *American Psychologist, 5,* 279.

Lefkowitz, J. (1972). Differential validity: Ethnic group as a moderator in predicting tenure. *Personnel Psychology, 25,* 223–240.

Lerner, R. M., & Busch-Rossnagel, N. A. (1981). Individuals as producers of their own development. In R. M. Lerner & N. A. Busch-Rossnagel (Eds.), *Individuals as producers of their own development: A life span perspective.* New York: Academic Press.

Levine, A. S., & Zachert, V. (1951). Use of biographical inventory in the Air Force classification program. *Journal of Applied Psychology, 35,* 241–244.

Lissitz, R. W., & Schoenfeldt, L. F. (1974). Moderator subgroups for the estimation of educational performance. *American Educational Research Journal, 11,* 63–75.

Lucchesi, C. Y. (1984). *The prediction of job satisfaction, life satisfaction and job level from autobiographical dimensions: A longitudinal application of structural equation modeling.* Unpublished doctoral dissertation, University of Georgia, Athens GA.

Lunneborg, C. E. (1968). Biographic variables in differential versus absolute prediction. *Proceedings of the 76th Annual Convention of the American Psychological Association, 3,* 233–234.

Magnusson, D., & Endler, N. S. (1977). *Personality at the crossroads: Current issues in interactional psychology.* Hillsdale NJ: Lawrence Erlbaum.

Majesty, M. S. (1967). *Identification of race, sex, and religion through life history.* Springfield VA: Clearinghouse for Federal Scientific and Technical Information.

Malloy, J. (1955). The prediction of college achievement with the life experience inventory. *Educational and Psychological Measurement, 15,* 170–180.

Malone, M. P. (1978). *Predictive efficiency and discriminatory impact of verifiable biographical data as a function of data analysis procedure.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Matteson, M. T. (1978). An alternative approach to using biographical data for predicting job success. *Journal of Occupational Psychology, 51,* 155–162.

Matteson, M. T., Osburn, H. G., & Sparks, C. P. (1969). *A computer based methodology for constructing homogeneous keys with applications to biographical data.* Houston TX: Personnel Psychology Services Center, University of Houston.

McMurray, R. N. (1947). Validating the patterned interview. *Personnel, 23,* 263–272.

Meacham, R. C., Shaffer, G. S., & Owens, W. A. (in preparation). *The prediction of job family membership from life history subgroups.*

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30,* 955–966.

Mikesell, R. H., & Tesser, A. (1971). Life history antecedents of authoritarianism: A quasi-longitudinal approach. *Proceedings of the 74th Convention of the American Psychological Association, 6,* 136–137.

Mitchell, T. W., & Klimoski, R. J. (1982). Is it rational to be empirical? A test of methods for scoring bio-

graphical data. *Journal of Applied Psychology, 67,* 411–418.

Morrison, R. F. (1977). A multivariate model for the occupational placement decision. *Journal of Applied Psychology, 62,* 271–277.

Morrison, R. F., Owens, W. A., Glennon, J. R., & Albright, L. E. (1962). Factored life history antecedents of industrial research performance. *Journal of Applied Psychology, 46,* 281–284.

Mosel, J. L., & Cozan, L. W. (1952). The accuracy of application blank work histories. *Journal of Applied Psychology, 36,* 365–369.

Mosel, J. N., & Wade, R. R. (1951). A weighted application blank for reduction of turnover in department store sales clerks. *Personnel Psychology, 4,* 177–184.

Mumford, M. D. (1981). *Life history and vocational interests*. Unpublished master's thesis, University of Georgia, Athens GA.

Mumford, M. D. (1982). *Item correlates of acquiescence and social desirability*. Athens GA: Institute for Behavioral Research.

Mumford, M. D. (1983). *Life history dimensions and types between age 18 and 30*. Unpublished doctoral dissertation, University of Georgia, Athens GA.

Mumford, M. D. (1986). *The description of individual differences in development: Cross-time versus point-in-time summarizations*. Manuscript submitted for publication.

Mumford, M. D., Cooper, M., & Schemmer, F. M. (1983). *Development of a content valid set of background data measures*. Bethesda MD: Advanced Research Resources Organization.

Mumford, M. D., & Owens, W. A. (1982). Life history and vocational interests. *Journal of Vocational Behavior, 21,* 330–348.

Mumford, M. D., & Owens, W. A. (1984). Individuality in a developmental context: Some empirical and theoretical considerations. *Human Development, 27,* 84–108.

Mumford, M. D., Shaffer, G. S., Ames, S. P., & Owens, W. A. (1983). *Analysis of PCEI responses over time*. Athens GA: Institute for Behavorial Research.

Mumford, M. D., Shaffer, G. S., Jackson, K. E., Neiner, A., Denning, D., & Owens, W. A. (1983). *Male-female differences in the structure of background data measures*. Athens GA: Institute for Behavioral Research.

Myers, D. C., & Fine, F. A. (1980). *Development of pre-employment experience questionnaires*. Bethesda MD: Advance Research Resources Organization.

Nach, A. N. (1965). A study of item weights and scale length for the SVIB. *Journal of Applied Psychology, 49,* 12–17.

Neiner, A. G., & Owens, W. A. (1982). Relationships between two sets of biodata with 7 years separation. *Journal of Applied Psychology, 67,* 146–150.

Neiner, A. G., & Owens, W. A. (1985). Using biodata to predict job choice among college graduates. *Journal of Applied Psychology, 70,* 127–136.

Nevo, B. (1976). Using biographical information to predict success of men and women in the Army. *Journal of Applied Psychology, 61,* 106–108.

Norman, W. T. (1963). Personality measurement, faking and detection: An assessment of a method for use in personnel selection. *Journal of Applied Psychology, 47,* 317–324.

Owens, W. A. (1968). Toward one discipline of scientific psychology. *American Psychologist, 23,* 782–785.

Owens, W. A. (1971). A quasi-actuarial basis for individual assessment. *American Psychologist, 26,* 992–999.

Owens, W. A. (1976). Background data. In M. D. Dunnette (Ed.), *Handbook of industrial psychology*. New York: Rand-McNally.

Owens, W. A. (1978). Moderators and subgroups. *Personnel Psychology, 31,* 243–247.

Owens, W. A., & Champagne, J. E. (1965). *A selected bibliography on biographical data*. Greensboro NC: The Richardson Foundation.

Owens, W. A., Glennon, J. R., & Albright, L. E. (1962). Retest consistency and the writing of life history items: A first step. *Journal of Applied Psychology, 46,* 329–332.

Owens, W. A., & Schoenfeldt, L. F. (1979). Toward a classification of persons. *Journal of Applied Psychology, 64,* 569–607.

Pace, L. A., & Schoenfeldt, L. F. (1977). Legal concerns in the use of weighted applications. *Personnel Psychology, 30,* 159–166.

Parsons, C. K., & Liden, R. C. (1984). Interviewer perceptions of applicant qualifications: A multivariate field study of demographic characteristics and nonverbal cues. *Journal of Applied Psychology, 69,* 557–568.

Peterson, D. A., & Wallace, S. R. (1966). Validation and revalidation of a test in use. *Journal of Applied Psychology, 50,* 13–17.

Pinto, P. R. (1970). *Subgrouping in prediction: A comparison of moderator and actuarial approaches*. Unpublished doctoral dissertation, University of Georgia, Athens GA.

Place, H. (1979). A biographical profile of women in management. *Journal of Occupational Psychology, 52,* 267–276.

Plag, J. A., & Goffman, J. M. (1967). The armed forces qualification test: Its validity in predicting military effectiveness for Naval enlistees. *Personnel Psychology, 20,* 323–340.

Porch, B. E., Collins, M., Wertz, R. T., & Friden, T. P. (1980). Statistical prediction of change in aphasia.

Journal of Speech and Hearing Research, 23, 317–321.

Quaintance, M. K. (1981). *Development of a weighted application blank to predict managerial assessment center performance*. Unpublished doctoral dissertation, George Washington University, Washington DC.

Rawls, D., & Rawls, J. R. (1968). Personality characteristics and personal history data of successful and less successful executives. *Psychological Reports, 23,* 1032–1034.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35,* 1–63.

Ritchie, R. J., & Boehm, V. R. (1977). Biographical data as a predictor of women's and men's management potential. *Journal of Vocational Behavior, 11,* 363–368.

Ross, M., McFarland, C., & Fletcher, G. J. (1981). The effect of attitude on recall of personal histories. *Journal of Personality and Social Psychology, 40,* 627–634.

Saunders, V. C. (1983). *Verification of responses to an autobiographical data form*. Unpublished master's thesis, University of Georgia, Athens GA.

Schmidt, F. L., Caplan, J. R., Dunn, L., Bemis, S. E., Decuir, R., & Antone, L. (1979). *The behavioral consistency method for unassembled testing*. Washington DC: U.S. Office of Personnel Management.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981). Validity generalization results for two job groups in the insurance industry. *Journal of Applied Psychology, 66,* 261–273.

Schoenfeldt, L. F. (1972). *Maximum manpower utilization: Development, implementation, and evaluation of an assessment classification model*. Paper presented at the annual meeting of the American Psychological Association, Honolulu.

Schoenfeldt, L. F. (1974). Utilization of manpower: Development and evaluation of an assessment-classification model for matching individuals with jobs. *Journal of Applied Psychology, 59,* 583–585.

Schrader, A. D., & Osburn, H. G. (1977). Biodata faking: Effects of induced subtlety and position specificity. *Personnel Psychology, 30,* 395–404.

Schuh, A. L. (1967). The predictability of employee tenure: A review of the literature. *Personnel Psychology, 20,* 133–152.

Schwab, D. P., & Oliver, R. L. (1974). Predicting tenure with biographical data: Exhuming buried evidence. *Personnel Psychology, 27,* 125–128.

Shaffer, G. S., Mumford, M. D., & Owens, W. A. (1986). *On the validity of retrospective self-report data*. Manuscript submitted for publication.

Sharf, J. (undated). *The supervisory profile record: Differential validity—test fairness*. Unpublished manuscript, Richardson, Bellows, Henry & Co., Inc., Washington DC.

Siegel, L. (1956). A biographical inventory for students: I. Construction and standardization of the instrument. *Journal of Applied Psychology, 40,* 5–10.

Smart, B. D. (1968). Reducing offensiveness of biographical items in personnel selection: A first step. *Studies in Higher Education, 95,* 14–21.

Smith, W. J., Albright, L. E., Glennon, J. R., & Owens, W. A. (1961). The prediction of research competence and creativity from personal history. *Journal of Applied Psychology, 45,* 59–62.

Sparks, C. P. (1965). *Using life history items to predict cognitive test scores*. Houston TX: Author.

Stead, N. H., & Shartle, C. L. (1940). *Occupational counseling techniques*. New York: American Book Company.

Strimbu, J. L., & Schoenfeldt, L. F. (1972). Life history subgroups in the prediction of drug usage patterns and attitudes. *JSAS Catalog of Selected Documents in Psychology, 3,* 83.

Strong, E. K. (1926). An interest test for personnel managers. *Journal of Personnel Research, 5,* 194–203.

Strong, E. K. (1941). *The vocational interests of men and women*. Stanford CA: Stanford University Press.

Taylor, L. R. (1968). *A quasi-actuarial approach to assessment*. Unpublished doctoral dissertation, Purdue University, Lafayette IN.

Telenson, P. A., Alexander, R. A., & Barrett, G. V. (1983). Scoring the biographical information blank: A comparison of three weighting techniques. *Applied Psychological Measurement, 7,* 73–80.

Thayer, P. W. (1977). Somethings old, somethings new. *Personnel Psychology, 30,* 513–524.

Thomas, L. L. (1982). The biographical antecedents of vocational choice. *Dissertation Abstracts International, 46-B,* 2031.

Toole, D. L., Gavin, J. F., Murdy, L. B., & Sells, S. B. (1972). The differential validity of personality, personal history, and aptitude data for minority and nonminority employees. *Personnel Psychology, 25,* 661–672.

Turnage, J. J., & Muchinsky, P. M. (1984). A comparison of the predictive validity of assessment center evaluations versus traditional measures in forecasting supervisory job performance: Interpretive implications of criterion distortion for the assessment paradigm. *Journal of Applied Psychology, 69,* 595–602.

Tyler, L. E. (1965). *The psychology of human differences*. Englewood Cliffs NJ: Prentice-Hall.

Vandeventer, A. D., Taylor, D. K., Collins, W. E., & Boone, J. O. (1983). *Three studies of biographical factors associated with success in air traffic control specialist screening/training at the FAA Academy*. Washington DC: Federal Aviation Administration.

Viteles, M. (1932). *Industrial psychology*. New York: Norton.

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49,* 436–458.

Ward, J. H., & Hook, M. E. (1963). Application of an hierarchical grouping procedure to a problem of grouping profiles. *Educational and Psychological Measurement, 23,* 69–81.

Waters, L. K., Roach, D., & Waters, C. W. (1976). Estimates of future tenure, satisfaction, and biographical variables as predictors of termination. *Personnel Psychology, 29,* 57–60.

Webb, S. C. (1960). The comparative validity of two biographical inventory keys. *Journal of Applied Psychology, 44,* 177–183.

Webster, E. G., Booth, R. F., Graham, W. K., & Alf, E. F. (1978). A sex comparison of factors related to success in Naval Hospital Corps School. *Personnel Psychology, 31,* 95–106.

Weiss, D. J., & Dawis, R. V. (1960). An objective validation of factual interview data. *Journal of Applied Psychology, 44,* 381–385.

Williams, W. E. (1961). *Life history antecedents of volunteers versus nonvolunteers for an AFROTC program*. Paper presented at the meeting of the Midwestern Psychological Association, Chicago.

Winer, B. J. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill.

Woods, J. G. (1962). An empirical analysis of the life history of good and poor salesmen. In R. W. Ferguson (Ed.), *The heritage of industrial psychology*. Hartford CT: Author.

Acknowledgments

Parts of this paper were made possible by a research grant from the National Institute of Child Health and Human Development (No. 5R01-HD04135-10). The authors thank their colleagues in background data research and two anonymous reviewers for their insightful comments, as well as Aglaja Hartmann for her careful review of the manuscript and reference listings.

Author's Address

Send requests for further information to Michael D. Mumford, School of Psychology, Georgia Institute of Technology, Atlanta GA 30332, U.S.A.

Reprints

Reprints of this article may be obtained *prepaid* for $2.50 (U.S. delivery) or $3.00 (outside U.S.; payment in U.S. funds drawn on a U.S. bank) from Applied Psychological Measurement, N658 Elliott Hall, University of Minnesota, Minneapolis MN 55455, U.S.A.