2-2000

Item Selection Counts: A Comparison of Empirical Key and Rational Scale Validities in Theory-Based and Non-Theory-Based Item Pools

Roni Reiter-Palmon
University of Nebraska at Omaha, rreiter-palmon@unomaha.edu

American Institutes for Research

Item Selection Counts : A Comparison of Empirical Key and Rational Scale Validities in Theory-Based and Non-Theory-Based Item Pools

Roni Reiter-Palmon

University of Nebraska at Omaha

Mary Shane Connelly

American Institutes for Research

Roni Reiter-Palmon, Department of Psychology, University of Nebraska at Omaha; Mary Shane Connelly, American Institutes for Research, Washington, DC.

Mary Shane Connelly is now at the Department of Psychology, University of Oklahoma.

An earlier version of this article was presented at the meetings of the Society for Industrial/Organizational Psychology, Dallas, Texas, 1998.

Correspondence concerning this article should be addressed to Roni Reiter-Palmon, Department of Psychology, University of Nebraska, Omaha, Nebraska 68182. Electronic mail may be sent to roni@unomaha.edu.

Abstract: Little explicit attention has been given to the impact of item pools on the validities and cross-validities of different background data scoring approaches. This study tests the idea that pools of items theoretically related to the performance of interest will outperform pools of items with no hypothesized relationship with the criterion. Validities and cross-validities of rational scales and empirical keys created from theory- and non-theory-based item pools were compared for 3 criteria. When size of the item pools was held constant, theory-based empirical keys (correlational and vertical percent) and rational scales showed larger validities and cross-validities than non-theory-based empirical keys (correlational and vertical percent) and showed minimal shrinkage in cross-validities. Even when item pool for the non-theory-based keys was expanded to include all items in the instrument, the theory-based keys showed comparable or slightly better validities and cross-validities for 2 of the 3 criteria, including college GPA, which was separated from the predictors by 4 years.

The effectiveness of background data or life history measures has been demonstrated through their repeated use in the areas of selection, placement, and training (Asher, 1972; Brown, 1994; Hunter & Hunter, 1984; Owens, 1976; Reilly & Chao, 1982). Sharf (1994) noted that background-data instruments are good alternatives to general cognitive ability tests because they demonstrate comparable validities and show less adverse impact. The increased use of background data in industry testing underscores the need for continued research on item development, scaling, validation, theory development, and a host of other areas (Stokes, Mumford, & Owens, 1994). The present article focuses on the quality of background-data item pools and their impact on different scoring approaches.

Empirical keying is the most commonly used method of scoring background data in personnel selection (Hogan, 1994; Mitchell & Klimoski, 1982). Empirical keys are composite scores generated by weighting and combining items on the basis of their relationships with the criterion of interest. They capitalize on item-criterion relationships to maximize the size of the validity coefficients (Guilford, 1954; Mumford & Owens, 1987). For this reason, the criterion-related validity of empirical keys tends to be strong. Empirical keys can also identify nonobvious, subtle relationships between items and the criterion that other scoring techniques might miss (Paullin & Hanson, 1995).

Despite their ability to predict, empirical keying procedures have not been without criticism. Because empirical keys capitalize on chance relationships between items and the criterion in the sample used to develop the key, the size of initial validities is inflated. When keys are cross-validated or applied in other samples to ensure stability and generality, validities typically decrease, particularly over time (Mumford & Owens, 1987; Schmidt & Rothstein, 1994; Thayer, 1977). This reduction could be attributed to the use of a development sample that does not adequately reflect the samples to which a key is applied, to differences in the criterion used to develop the key and the ultimate criterion of interest, or to poor person-to-predictor ratios (Hogan, 1994). Decreases could also be due to one of the most significant problems with empirical keys—their atheoretical nature. It is often difficult to understand why keyed items are (or are not) related to the criterion. Several researchers have suggested that the lack of attention to content and construct validity issues in the development and application of background data limits its potential contributions to theory generation and testing (Dunnette, 1962; Henry, 1966; Mumford, Costanza, Connelly, & Johnson, 1996; Owens, 1976; Tenopyr, 1994).

Brown (1994) noted that “biodata measures are only valid to the extent that they adequately tap relevant life experiences” (p. 199). Rational approaches to developing and scoring background data have tried to ensure item relevance, generality, and long-term prediction by using theory to guide item development, selection, and scaling. Developing rational scales involves identifying psychological constructs that underlie the performance of interest and writing or selecting items to tap those constructs. It has been argued that rational scales and their resulting validities are only as good as the intuitive power of item developers (Paullin & Hanson, 1995). However, this concern is mitigated when item development is grounded in sound psychological theories where constructs and their interrelationships are defined with respect to the criterion and when developers receive training and guidance in item writing (Mumford et al., 1996).

Mumford and Owens (1987) have pointed out that when background-data items capture significant antecedents of performance, they show larger validities across situations and time than items that don't. Items based on constructs reflecting individual attributes linked to performance are likely to be more generalizable than empirical keys for several reasons. First, these items are less sensitive to transient influences (Mumford & Stokes, 1992). Second, items reflecting underlying attributes are likely to show continuity over time because of the consistency of these attributes. People's choices regarding situational entry and behavior in those situations tend to be consistent with past choices, reinforcing and contributing to further development of existing attributes (Mumford & Stokes, 1992). Studies have shown the effectiveness of these items as meaningful predictors of performance (Mumford et al., 1996; Pannone, 1984; Russell, Mattson, Devlin, & Atwater, 1990; Schoenfeldt & Mendoza, 1988).

Empirical and rational scoring approaches have been compared in the past. Hough and Paullin (1994) presented findings from 21 studies that compared empirical keying (external), factorial scaling (inductive), and rational scaling (deductive). Median differences across studies in criterion-related validities showed no differences among scoring techniques. Two characteristics of Hough and Paullin's (1994) research make it difficult to determine whether their findings hold true for background data. First, item type was not a criterion for inclusion in the studies, so there is no indication of how many included background-data items. Second, although cross-validities were available for the studies in the comparison, they are not reported or compared, so there is no information bearing on the stability and generality of the scoring approaches. This information is critical given that cross-validities for empirical keys traditionally show meaningful shrinkage (Hogan, 1994; Mumford & Owens, 1987).

Other studies comparing empirical and rational approaches have included both validation and cross-validation evidence. Paullin and Hanson (1995) found that empirically keyed background-data items evidenced the same or higher validity coefficients as rationally scaled items in predicting a variety of job performance constructs and had comparable cross-validities. Mumford and Stokes (1992) described other studies comparing rational scaling and empirical keying approaches. Berkeley (1953) found that empirical keys resulted in larger criterion validities than rational scales but showed greater shrinkage in cross-validities. Hornick, James, and Jones (1977) and Mumford, Uhlman, and Kilcullen (1992) found that rational scales and empirical keys produced similar cross-validities and that the results of the rational scales were more interpretable. Finally, Mitchell and Klimoski (1982) found that empirical keys produced larger validities and cross-validities than rational scales.

One important factor that studies to date have not addressed is the quality of the item pools from which scales or keys are developed. Russell (1994), Mumford and Owens (1987), and others have suggested that quality of the item pool influences quality of an empirical key. Recent evidence also suggests that item development/selection (Hogan, 1994; Schmidt & Rothstein, 1994) and method of scaling (Devlin, Abrahams, & Edwards, 1992) are important determinants of the stability and generality of background data validities. Studies comparing empirical keying approaches have typically focused on differences resulting from how keys are constructed (i.e., criteria for item selection, weighting, scaling) and not on the initial quality of the item pool. High-quality item pools containing items theoretically related to the criteria of interest may result in more stable, meaningful empirical keys. The present study compares two empirical keying approaches to a rational scaling approach, varying the type of item pool used to generate the empirical keys. It is expected that validities and cross-validities for theory-based empirical keys will be as large as or larger than those for rational scales, which, in turn, will be larger than those for non-theory-based keys. Baseline analyses using all items in the instrument are also conducted, serving as a comparison to current industry practice.

Method

Participants

The sample used for this study was obtained as part of a longitudinal study described by Owens and Schoenfeldt (1979). Specifically, participants in this study consisted of the 1,969 college freshmen (915 women and 1,054 men) from a large southeastern university. This data set was randomly divided into a validation sample that was composed of approximately two-thirds of the participants (1,316 total, 613 women, 703 men) and a cross-validation sample that was composed of the remaining participants (653 total, 302 women, 351 men). Although these data have been used in a number of prior studies, comparisons of the scaling procedures used in the present study have not been addressed before. One advantage of using this data set in comparing methods of scaling background data is that other studies have demonstrated the stability of the item characteristics. Participants were given a 389-item background-data questionnaire (BQ389) during freshmen orientation, which was used to develop all rational scales and empirical keys used in this study. The questionnaire includes a variety of questions about life experiences in areas such as school, family, friends, and hobbies.

Criteria

In selecting the criteria for this study, we were guided primarily by a practical consideration: availability in the existing data set. Additionally, we wanted to work with familiar criterion and predictor domains so that the primary focus of the article would be on comparing scaling approaches rather than on theory development and testing. The main criterion for this study is college grade point average (GPA), which was collected from school records 4 years after the BQ389 was administered (Owens & Schoenfeldt, 1979). Two concurrent criteria are also included. High school GPA was obtained from school records, and a high-school-leadership scale was drawn from Mumford, O'Connor, Clifton, Connelly, and Zaccaro (1993). These authors developed rational scales and examined their relationships to adolescent leadership using the Owens and Schoenfeldt (1979) data. The leadership scale includes eight BQ389 items reflecting

leadership activities such as directing others in group activities, participating in school politics, influencing others, holding leadership positions, and being described by others as a leader. The internal consistency coefficient for this scale is .84.

Rational Scale Construction

Two sets of rational scales were developed for the present study, one for high school and college GPA and one for high school leadership. We expected that several variables would be positively related to GPA, such as math and verbal skills, motivation, and self-esteem. Thus, the following construct scales were generated for the GPA criteria: achievement motivation, quantitative skills/scientific interests, verbal skills, work ethic, institutional adaptation (adjustment to high school), and self-esteem. Given that grades in high school often correlate with grades in college, a high-school-achievement scale (i.e., grades in specific courses) was added for predicting college GPA. The Mumford et al. (1993) study provided some data regarding variables that predict high school leadership, college leadership, and success in leadership training courses. Their findings suggest that for high school leadership, motivational and social variables are important. The broader leadership literature has also shown consistent relationships of these types of variables to indexes of leadership (Bass, 1990). Accordingly, rational scales tapping verbal skills, achievement motivation, work ethic, self-esteem, institutional adaptation, persuasion/dominance, social adjustment, and independence were generated for the leadership criterion.

We developed rational scales by using a modified version of the procedures described in Mumford et al. (1996). A panel of three industrial/organizational psychologists with 5 or more years of experience working with background data was convened. For each construct, panel members examined a conceptual definition and came to a consensual understanding of the construct. Next, they were asked to independently select items from the BQ389 that would mark each construct. In conducting this item-selection task, panel members were asked to think about behavior, reactions, and outcomes in various situations that indicate direct manifestation of the construct at hand or that might contribute to the development of the construct at hand. The Mumford et al. (1993) study contained scales for some of the same constructs. Each person was given a list of items in these scales for reference. Once each person had selected items for each construct, the group reviewed the items for relevance to the construct at hand and jointly decided on the final set of items for the scale. This process was repeated for all scales. Reliabilities for the rational scales developed for the GPA criteria ranged from .56 to .85, averaging .70, an acceptable range for rational scales (Mumford et al., 1996; Mumford & Owens, 1987). Together, these scales contained a total of 71 background data items. The high-school-achievement scale was not used to predict high school GPA because the items in this scale asked directly about grades in high school, resulting in the use of 62 items for this criterion. Average reliability for the scales developed for the high-school-leadership criterion was .66, ranging from .54 to .80. These scales contained a total of 76 items. Table 1 lists the rational scales, providing the number of items in each scale, sample items, and Cronbach alphas.

Table 1
Rational Scales

Construct (no. of items in scale)	Sample items	Reliability
Achievement motivation (12)[a,b]	Set difficult goals; performed better under pressure; tried to achieve to the limits of ability; parents emphasized getting ahead	.67
Math and scientific interests (10)[a]	Enjoyed lab courses; enjoyed science courses; enjoyed math courses; conducted scientific experiments	.71
Verbal skills (13)[a,b]	Enjoyed discussion courses; read more books relative to others; read literary magazines; thought English classes were easy	.70
Work ethic/responsibility (8)[a,b]	Had a strong sense of responsibility; disturbed if a job was left unfinished; wished to become a benefit to society	.56
Institutional adaptation (7)[a,b]	Was liked and respected by teachers; liked high school; received consistent criticism	.60
Self-esteem (12)[a,b]	Satisfied with self in high school; did not feel self-conscious; was respected by classmates; effectively met demands of social situations	.80
Academic achievement (9)[a]	Made semester honor roll often; had high class standing; had high grades in English; had high grades in math	.85
Persuasion/dominance (4)[b]	Tried to make others see point of view; participated in small groups; said what he/she felt	.54
Social adjustment (14)[b]	Dealt effectively with the demands of social situations; did not have difficulty making friends; was invited to social activities; got along with people who were different from oneself	.80
Independence (6)[b]	Parents encouraged to explore new situations; parents allowed more independence in high school compared to friends; independent of others during high school	.62

[a] Scales developed for high school and college GPA criteria. [b] Scales developed for leadership criterion.

Empirical Key Construction

Researchers have used a number of different methods to construct empirical keys, such as correlation/regression, vertical percent, horizontal percent, mean criterion, and rare response (Devlin et al., 1992; Guion, 1965; Hogan, 1994). Variations within each of these approaches further increase the options available for developing an empirical key. Different keying methods use different procedures for selecting and weighting items; however, most methods use variance maximizing procedures to identify items or options that best discriminate low from high performers on the criterion of interest (Mumford & Owens, 1987). Cross-validation is essential in determining the stability over time and transportability of the key by examining its ability to differentiate other groups of people on the criterion. Details of the different types of keying approaches are discussed elsewhere in more depth (see Devlin et al., 1992; Hogan, 1994).

From the numerous key construction methods available, we selected two for this study—a correlation approach and a vertical-percent approach. Prior research recommends that when the criterion measure and the items to be keyed are both continuous, the correlation between each item and the criterion should

guide the development of the key (Hogan, 1994; Mumford & Owens, 1987). We selected pattern-of-response keying as the correlational keying method given that continuous items and criteria were used here. This keying procedure did not require establishing criterion groups. A comparison of other types of keying procedures in an analysis by Devlin et al. (1992) suggests that differentially weighted vertical-percent methods are preferred over other keying methods because they show less shrinkage when cross-validated. Thus, a vertical 5% strategy was selected as the second empirical keying method. We used three different types of item pools to generate six different empirical keys for each criterion (three using the correlational method, three using the vertical-percent method). A total of 18 different keys were developed.

Items in all keys developed for this study were unit weighted for a number of reasons. Unit weighting is preferable if sample sizes are not large enough to develop stable differential weights (5 to 10 people per predictor; Dawes, 1971). The number of people in the present study would have been sufficient to apply differential weighting for some of the keys but not for others. Additionally, gains from differential weighting over unit weighting are more likely if the number of items in use is small and when those items have low intercorrelations (Guilford, 1954; Lawshe & Schucker, 1959). This situation was not the case in this study. Finally, McGrath (1960) found no differences in cross-validities of a unit-weighted key and a differentially weighted key where weights were developed on the basis of degree of statistical significance of item-criterion correlations (Hogan, 1994).

Correlational method

We used three item pools to develop keys using the correlational method. The first was a theory-based item pool containing the set of items used in the rational scales (62 items for high school GPA, 71 items for college GPA, 78 items for high school leadership). The second item pool contained the same number of items as the one used for the rational scales, but a random selection procedure was used to select items. The third item pool included all 389 items in the questionnaire, serving as a baseline comparison to common industry practice. However, it should be noted that it is uncommon to have 389 items in a background-data inventory. Typically, inventories contain about 100 to 200 items (Stokes, Mumford, & Owens, 1994). For keys developed using the correlational method, items were correlated with the criterion of interest in the validation sample. Items correlating .10 and above and −.10 and below were selected for inclusion in the key. Items were unit weighted and summed to form the keys.

Vertical-percent method

The same three item pools that we used to develop the correlational keys were used to develop the vertical-percent keys—a pool of theory-based items, a pool of randomly selected items, and a pool containing all 389 items. The following steps were followed to create keys for each item pool. First, criterion groups were defined (e.g., good vs. poor performers) in the validation sample. We used individuals whose scores fell in the top and bottom one-third of the distribution to define the high- and low-performing groups following common procedures for continuous criteria (Ghiselli, Campbell, & Zedeck, 1981). Different criterion groups of good and poor performers were created for each criterion of interest. Once the groups had been established, we identified percentages of individuals in each group who endorsed each response option for all items in the pool. We calculated differences in endorsement percentages by subtracting the percentage endorsed by poor performers from the percentage endorsed by

good performers. Only response options with a difference of at least 5% were assigned unit weights. When the poor-performing group's endorsement percentage was higher, a weight of −1 was assigned, and when the high-performing group's endorsement percentage was higher, a weight of 1 was assigned.

Validation and Cross-Validation Analyses

We compared scaling approaches by using multiple regression and correlational analyses to calculate validity and cross-validity coefficients. Longitudinal data enabled both predictive and concurrent analyses. Least squares multiple regression analyses were run separately on the validation sample for each of the three criteria to obtain rational scale validities. All rational scales were entered simultaneously into the regression equation for each analysis. We used regression weights obtained from the validation sample analyses to compute weighted composites for the cross-validation analyses. We correlated each composite with its corresponding criterion to obtain cross-validation coefficients. We computed validities and cross-validities for the empirical keys by correlating all keys with each of the criteria in the validation and cross-validation samples.

Results

Before comparing the validities and cross-validities of the different scoring methods, we provide some evidence regarding the meaningfulness of the rational scales and criterion measures. Table 2 includes the means, standard deviations, and correlations of the rational scales, high school GPA, college GPA, and the high-school-leadership scale. High school achievement (grades in specific classes) showed large positive correlations with high school GPA (r = .74) and college GPA (r = .53), moderate correlations with work ethic (r = .40) and achievement motivation (r = .36), and relatively small correlations with persuasive dominance (r = .08) and social adjustment (r = .06). Institutional adaptation showed moderate to large correlations with self-esteem (r = .44), achievement motivation (r = .41), and social adjustment (r = .31) but had smaller correlations with scientific/quantitative interests (r = .00) and independence (r = .20). High school leadership had the largest correlations with social adjustment (r = .60), work ethic (r = .57), persuasive dominance (r = .44), self-esteem (r = .44), and achievement motivation (r = .50) and showed smaller correlations with independence (r = .24) and scientific/quantitative interests (r = .11). It was somewhat surprising that the scientific/quantitative interests scale did not show larger correlations with high school or college GPA. Perhaps had we had items tapping scientific/quantitative skills rather than interests, the correlations would have been larger. Generally, however, the patterns of convergent and discriminant relationships among the rational scales and the criteria provide initial evidence that the scales are measuring what they were intended to measure.

Table 2
Means, Standard Deviations, and Correlations for Rational Scales

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Achievement motivation	43.2	5.07	—	.32	.49	.52	.41	.36	.25	.19	.35	.30	.35	.50	.29
2. Self-esteem	4.46	6.49		—	.36	.16	.44	.25	.27	.05[b]	.71	.15	.13	.44	.05[b]
3. Work ethic	27.7	3.84			—	.29	.40	.30	.34	.12	.43	.27	.21	.57	.13
4. High school achievement	31.88	6.87				—	.36	.08	.13	.20	.06	.23	.74	.30	.53
5. Institutional adaptation	26.2	3.51					—	.20	.20	.00[b]	.31	.28	.29	.31	.24
6. Persuasive dominance	14.5	2.33						—	.26	.09	.39	.41	−.03[b]	.44	.01[b]
7. Independence	25.0	3.86							—	.13	.23	.19	.07[a]	.24	.08
8. Quantitative ability	27.7	6.27								—	.03[b]	.04[b]	.01[b]	.11	−.04[b]
9. Social adjustment	52.2	8.35									—	.22	.01[b]	.60	−.05[b]
10. Verbal ability	43.3	6.84										—	.05[b]	.31	.12
11. High school GPA	3.12	0.54											—	.17	.53
12. High school leadership	25.3	5.88												—	.11
13. College GPA	2.54	0.67													—

Note. $N = 1{,}280$. GPA = grade point average.
[a] $p < .05$. [b] Not significant; all other correlations were significant at $p < .01$.

Scaling approaches are compared with respect to several indicators—size of validities, size of cross-validities, and degree of shrinkage in cross-validities. Differences or decreases of .10 are typically considered meaningful (Hogan, 1994; Mumford & Owens, 1987). First, results for the predictive analyses involving college GPA are presented. Second, results for the concurrent analyses involving high school GPA and leadership are briefly described.

Predictive Criteria—College GPA

Table 3 presents the analyses for the college GPA criterion. As expected, the empirical keys developed from non-theory-based item pools had somewhat lower validities (a marked difference was found for the correlational key but not for the vertical-percent key) and substantially lower cross-validities than the rational scales. Second, these keys performed much more poorly than the theory-based keys, showing differences in validities and cross-validities of .34 and .32, respectively, for the correlational method and differences of .08 and .22 for the vertical percent method, again confirming expectations. These findings suggest that both theory-based empirical keys and rational scales are less susceptible to decays in validities over time than non-theory-based keys. Third, the theory-based empirical keys performed better than the rational scales in the validation and cross-validation samples by about .10. Given that empirical keys are designed to maximize prediction of the criterion, it is not surprising that when the same theory-based item pool is used, empirical keys perform better than rational scales. Although items are initially selected for the item pool on the basis of theoretical relevance to the criterion for both scaling procedures, an empirical key has the added advantage of only using items that demonstrate the strongest empirical relationships with the criterion.

Results from the full-instrument keys were mixed in terms of how well they performed relative to the other scoring methods. The full-instrument correlational key did not perform as well as the correlational theory-based key and the rational scales in the validation sample, although there was no shrinkage when it was cross-validated. Although the full-instrument vertical-percent key had a higher validity coefficient than the theory-based vertical-percent key, it showed a greater degree of shrinkage in the cross-validation

sample. The full-instrument keys showed higher validities and cross-validities than the non-theory-based keys.

Concurrent Criteria—High School GPA and High School Leadership

Table 4 presents results for high school GPA, showing mixed support for what was expected. Surprisingly, empirical keys developed from non-theory-based item pools produced validities and cross-validities comparable to the rational scales for the high-school-GPA criterion. However, theory-based empirical keys showed higher validities and cross-validities than both non-theory-based keys and rational scales, as anticipated. Findings for the correlational and vertical-percent keying methods were similar, and little shrinkage was observed across all scaling methods. Baseline analysis using the full instrument showed the largest validities and cross-validities across methods for the high-school-GPA criterion.

Table 4
Validities and Cross-Validities for High School Grade Point Average

Item pool	Scaling/keying method	No. of items in pool	No. of items in scales/key	Validity	Cross-validity
Theory based	Rational scales	62	62	.43	.41
Theory based	Correlational key	62	28	.51	.51
Theory based	Vertical 5%	62	42	.56	.52
Randomly selected	Correlational key	62	14	.40	.36
Randomly selected	Vertical 5%	62	36	.44	.43
Full instrument	Correlational key	380	116	.59	.58
Full instrument	Vertical 5%	380	239	.66	.61

Note. All validities and cross-validities were significant at $p < .01$. Items from the high-school-achievement scale were not included in the full-instrument item pool.

Examination of the results for the leadership criterion presented in Table 5 suggests that the validity and cross-validity coefficients for the rational scales are about .08 and .11 higher, respectively, than those produced by the two non-theory-based keys and showed half as much shrinkage. Findings for the leadership criterion were similar for both the correlational and vertical-percent keying methods (non-theory), with the correlational cross-validities showing slightly more shrinkage than those from the vertical-percent key. The relatively large validities across all scaling procedures for this criterion may be due in part to method bias because both the predictors and criterion were developed from the same background data instrument.

Table 5
Validities and Cross-Validities for High School Leadership

Item pool	Scaling/keying method	No. of items in pool	No. of items in scales/key	Validity	Cross-validity
Theory based	Rational scales	76	76	.71	.67
Theory based	Correlational key	76	57	.72	.68
Theory based	Vertical 5%	76	67	.71	.70
Randomly selected	Correlational key	76	43	.62	.54
Randomly selected	Vertical 5%	76	59	.64	.58
Full instrument	Correlational key	381	231	.69	.62
Full instrument	Vertical 5%	381	304	.70	.69

Note. All validities and cross-validities were significant at $p < .01$. Items from the leadership scale were not included in the full-instrument item pool.

Theory-based empirical keys and rational scaling methods do equally well for the leadership criterion. Again, results for the theory-based correlational and vertical-percent keys were similar, and little shrinkage was observed across all scaling approaches. Empirical keys developed from theory-based item pools showed larger validities and cross-validities and evidenced less shrinkage than the keys developed from a randomly generated item pool containing the same number of items. This pattern of findings was the same for both the correlational and vertical-percent keying methods. It appears that starting with items that are more theoretically relevant to the criterion results in stronger empirical keys.

Interestingly, comparison with baseline results revealed that theory-based empirical keys and rational scales perform equally well or even slightly better than the full-instrument keys, despite the fact that their item pools contained one-fifth as many items. The correlational full-instrument key showed more shrinkage (−.07) than the vertical-percent full-instrument key (−.01) and both of the theory-based keys (−.04, −.01).

Summary

Average results across all criteria are presented in Table 6. Overall, validities were strongest for the theory-based and full-instrument empirical keys, whereas cross-validities for the theory-based approaches (empirical keys and rational scales) were the largest and showed the least amount of shrinkage across criteria. On average, the rational scales and theory-based empirical keys showed little or no shrinkage in cross-validities for the high school GPA, high school leadership, and college GPA criteria. The non-theory-based and full-instrument keys showed greater shrinkage, especially for the leadership and college GPA criteria. Vertical-percent empirical keys did slightly better than correlational empirical keys in the non-theory-based item pools in overall size of validities but showed greater shrinkage in cross-validities. No differences between these types of empirical keying approaches were observed in the theory-based item pools. Validities and cross-validities were not a function of the number of items or options included in the item pools or the final keys. In addition to comparing the values of validities and cross-validities across different scaling methods and item pools, it is important to examine the items that were retained for keying in the non-theory-based keys. Examination of the meaning of item–criterion relationships for one of the non-theory-based keys reinforced prior findings that it is not always clear why items in an empirical key are correlated with the criterion. Although some items did have meaningful relationships

with the criterion, many items did not. Sample items demonstrating explainable and unexplainable relationships with the high-school-GPA criterion are presented in Table 7.

Table 6
Average Validities and Cross-Validities for All Criteria

Item pool	Scaling/keying method	Validity	Cross-validity
Theory based	Rational scales	.57	.56
Theory based	Correlational key	.64	.63
Theory based	Vertical 5%	.62	.60
Randomly selected	Correlational key	.46	.43
Randomly selected	Vertical 5%	.53	.46
Full instrument	Correlational key	.60	.59
Full instrument	Vertical 5%	.68	.53

Table 7
Sample of Randomly Selected Items Predictive of High School Grade Point Average

Item	Description
Explainable relationships	Tried to achieve to the limits of your ability; made semester honor roll in high school more often; parents often expressed interests in your activities; high grade in math; teachers were successful in arousing academic interest; liked school.
Unexplainable relationships	Read women's magazines; held no summer jobs; characterized oneself as difficult to get to know; was not punished by parents for acting irresponsibly; did not build things (models, furniture); did not repair electrical or mechanical devices; often took feelings out on parents.

Discussion

The patterns of results in this study generally supported our expectations regarding rational scales, theory-based empirical keys, and non-theory-based empirical keys. Comparisons of various methods for scaling background data using different types of item pools revealed several interesting findings. First, empirical keys generated from theory-based item pools resulted in comparable or larger validities and cross-validities than rational scales, performed better than non-theory-based keys, and showed little shrinkage in the cross-validation samples across the criteria. Second, rational scales produced larger validation and cross-validation coefficients than empirical keys developed from a non-theory-based item pool of the same size for two out of the three criteria in this study. They showed comparable validities and cross-validities for the high-school-GPA criterion, one finding that did not confirm expectations. This finding might suggest that when concurrent validation is of primary concern, empirical keys generated from non-theory-based item pools perform adequately if size of the validities and cross-validities is the primary concern. Third, the relationships of items to criteria appeared to be more interpretable for the theory-based empirical keys than for the non-theory-based and full-instrument keys. Fourth, differences in shrinkage between theory-based approaches (empirical keys and rational scales) were most apparent with college GPA, where the predictors and criterion were separated by 4 years. Specifically, empirical keys with no theoretical basis evidenced greater shrinkage than theory-based approaches. This finding was

particularly true when the number of items in the item pool was held constant. Perhaps the most interesting finding is that the theory-based keys performed as well as or better than the keys developed from a full-instrument item pool for two out of the three criteria used in this study, with an item pool containing one-fifth the number of items. The size of the validities and cross-validities was not a function of the number of items in the pool or the number of items/options included in the key/scales. Finally, vertical-percent empirical keys did slightly better than correlational empirical keys in the non-theory-based item pools in terms of overall size of validities and cross-validities (with the exception of the college-GPA criterion, where marked shrinkage was evidenced). This suggests that use of high-quality item pools is one factor that may minimize differences in results for various types of empirical keys (Devlin et al., 1992). No differences between these approaches were observed when theory-based item pools were used to generate the keys. Taken together, these results have a number of broader theoretical and practical implications for scaling background-data items.

The findings underscore the importance of high-quality item pools. Hogan (1994), Mumford and Owens (1987), and Russell (1994) noted that the quality of background-data scaling methods is likely to be improved by starting with a high-quality item pool. Theory-driven approaches to generating and scoring background data enhance the goals of prediction and stability in cross-validation over time. More importantly, they enable meaningful inferences to be drawn about why certain life history experiences are important antecedents or manifestations of job performance and how they might condition future performance. Additionally, problems of deficiency and contamination in the predictor domain are less likely when items are developed on the basis of underlying theory (Mumford & Stokes, 1992; Nickels, 1994).

Mumford and Stokes' (1992) description of how patterns of responding shape individual development may help to explain why theory-based approaches offer stable, meaningful prediction. Although changes occur during the course of individual development, there is a certain amount of continuity in the types of situations people choose to enter, in how people behave in or react to situations, or both. Patterns of responding in new situations tend to be consistent with or maximize fit with an individual's existing characteristics and capabilities, reinforcing their continued existence and development. These characteristics underlie behavior and performance in various types of situations, for better or worse. Background-data items that capture the range of responses reflective of these underlying characteristics, of the performance of interest, or of both are likely to show strong relationships with the criterion over time and will be less susceptible to transient influences on individual item responses. Schoenfeldt and Mendoza (1988) reinforced this point by emphasizing that theory-driven background-data items are required for establishing construct validity. Hypotheses about developmental determinants of performance must be developed to select or develop items with relevant content. Only then can background data hope to serve broader goals such as theory development and testing. Rather than debating whether a rational scaling or empirical keying approach to scaling background data best serves these interests, it may be more constructive to emphasize the selection or development of high-quality, theory-based item pools. Generating or selecting items on the basis of theories of performance will help to forward such interests.

As the world of work continues to change (Cascio, 1995), so must the nature of the predictors and criteria researchers use in employment settings. Mumford, Reiter-Palmon, and Snell (1994) suggested that scoring keys be reevaluated every 5 to 10 years as the nature and structure of situations contributing to the

development of individuals' characteristics changes. For example, items asking about the frequency and nature of computer usage in college students might produce very different relationships with GPA today than they would have 10 or 20 years ago. Likewise, as new ways of defining and measuring multiple facets of job performance emerge, our thinking about the predictor space must change. Hypotheses about why certain items underlie performance provide a basis for item evaluation and redefinition of predictor domains.

Theory-driven item pools and scaling approaches also have some practical advantages. Results from this study suggest that when a theory-based item pool is used, fewer items are required to achieve validities and cross-validities comparable to those obtained when a much larger number of items are used. Users of background data in organizational settings may be able to substantially reduce testing time, which is always at a premium. Linking background data items to theory is also likely to improve the job relatedness of the inventory, an absolute necessity for legal defensibility in selection and other employment contexts. Organizations using background-data inventories that have not been explicitly linked to a theory of performance could reexamine items in light of job analysis or other data that could be used to develop a model of job performance as a starting point for improving their instruments. Theory-based item pools and scaling approaches may also make it easier to develop parallel test forms (Mumford & Stokes, 1992).

Finally, several caveats should be addressed. First, gender differences were not investigated as a part of this study. Mumford and Stokes (1992) indicated that patterns of item responding can differ with respect to gender. We had some evidence to indicate that this was not the case for the high-school-leadership criterion from the Mumford et al. (1993) study. They found that similar sets of constructs measured with rational scales predicted leadership for both men and women. Additionally, similar percentages of men and women comprised the criterion groups used to develop the empirical keys for the three criteria. Second, this study did not examine the prevalence or impact of faking, an important issue in employment situations. Items with intuitively obvious relationships with the criterion may be more susceptible to faking because they may be more transparent. Investigations are needed to test whether items from theory-based pools are more susceptible to faking than items from item pools that have not been generated on the basis of underlying theory and to identify the impact of faking on the effectiveness of different scaling methods. Third, most of the item-response options used in this study were continuous. This type of response option is one of several that have been used with background data (Asher, 1972). Additional research is needed to better understand how different response options may impact the effectiveness of different scaling methods. Finally, although different types of item were used in this study (see Mael, 1991), differences that may result from using items with different attributes were not examined. Investigations of the effects of item type on the effectiveness of different scaling methods are also needed.

References

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

Bass, B. M. (1990). *Bass and Stogdill's handbook of leadership: Theory, research, and managerial applications* (3rd ed.). New York: Free Press.

Berkeley, M. H. (1953). *A comparison between the empirical and rational approaches for keying a heterogeneous test* (USAF Human Resources Research Bulletin No. 53–24). Lackland AFB, TX.

Brown, S. H. (1994). Validating biodata. In G. S.Stokes, M. D.Mumford, & W. A.Owens (Eds.) , *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 199– 236). Palo Alto, CA: CPP Books.

Cascio, W. F. (1995). Wither industrial and organizational psychology in a changing world of work?*American Psychologist*, *50*, 928– 939.

Dawes, R. (1971). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*, 571– 582.

Devlin, S. E., Abrahams, N. M., & Edwards, J. E. (1992). Empirical keying of biographical data: Cross-validity as a function of scaling procedure and sample size. *Military Psychology*, *4*, 119– 136.

Dunnette, M. D. (1962). Personnel management. *Annual Review of Psychology*, *13*, 285– 314.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: Freeman.

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

Guion, R. M. (1965). *Personnel testing*. New York: McGraw-Hill.

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

Hogan, J. B. (1994). Empirical keying of background data measures. In G. S.Stokes, M. D.Mumford, & W. A.Owens (Eds.) , *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 69– 108). Palo Alto, CA: CPP Books.

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., & Paullin, C. (1994). Construct-oriented scale construction: The rational approach. In G. S.Stokes, M. D.Mumford, & W. A.Owens (Eds.) , *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 109– 146). Palo Alto, CA: CPP Books.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72– 98.

Lawshe, C. H., & Schucker, R. E. (1959). The relative efficiency of four test weight methods in multiple prediction. *Educational and Psychological Measurement*, *19*, 103– 114.

Mael, F. A. (1991). A conceptual rationale for the domain and attributes of biodata items. *Personnel Psychology*, *44*, 763– 793.

McGrath, J. J. (1960). Improving credit evaluation with a weighted application blank. *Journal of Applied Psychology*, *44*, 325– 328.

Mitchell, T. W., & Klimoski, R. J. (1982). Is it rational to be empirical? A test of methods for scoring biographical data. *Journal of Applied Psychology*, *67*, 411– 418.

Mumford, M. D., Costanza, D. P., Connelly, M. S., & Johnson, J. F. (1996). Item generation procedures and background data scales: Implications for construct and criterion-related validity. *Personnel Psychology*, *49*, 361– 398.

Mumford, M. D., O'Connor, J., Clifton, T. C., Connelly, M. S., & Zaccaro, S. J. (1993). Background data constructs as predictors of leadership behavior. *Human Performance*, *6*, 151– 195.

Mumford, M. D., & Owens, W. A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement*, *11*, 1– 31.

Mumford, M. D., Reiter-Palmon, R., & Snell, A. (1994). Background data and development: Structural issues in the application of life history measures. In G. S.Stokes, M. D.Mumford, & W. A.Owens (Eds.) , *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 555– 581). Palo Alto, CA: CPP Books.

Mumford, M. D., & Stokes, G. S. (1992). Developmental determinants of individual action: Theory and practice in applying background measures. In M. D.Dunnette (Ed.) , *Handbook of industrial and organizational psychology* (2nd ed., *Vol. 3*, pp. 61– 138). Palo Alto, CA: CPP Press.

Mumford, M. D., Uhlman, C. E., & Kilcullen, R. N. (1992). The structure of life history: Implications for the construct validity of background data scales. *Human Performance*, *5*, 109– 137.

Nickels, B. (1994). The nature of biodata. In G. S.Stokes, M. D.Mumford, & W. A.Owens (Eds.) , *Biodata handbook: Theory, research and use of biographical information in selection and performance prediction* (pp. 1– 16). Palo Alto, CA: CPP Books.

Owens, W. A. (1976). Background data. In M. D.Dunnette (Ed.) , *Handbook of industrial and organizational psychology* (1st ed., (pp. 609– 644). Chicago: Rand McNally.

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

Pannone, R. D. (1984). Predicting test performance: A content valid approach to screening applicants. *Personnel Psychology*, *30*, 159– 166.

Paullin, C., & Hanson, M. A. (1995, April). *Can the validity of rationally-derived inventories be improved by empirical keying procedures?*Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology*, *35*, 1– 62.

Russell, C. J. (1994). Generation procedures for biodata items: A point of departure. In G. S.Stokes, M. D.Mumford, & W. A.Owens (Eds.) , *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 17– 38). Palo Alto, CA: CPP Books.

Russell, C. J., Mattson, J., Devlin, S. E., & Atwater, D. (1990). Predictive validity of biodata items generated from retrospective life experience essays. *Journal of Applied Psychology, 75*, 511– 523.

Schmidt, F. L., & Rothstein, H. R. (1994). Application of validity generalization to biodata scales in employment selection. In G. S.Stokes, M. D.Mumford, & W. A.Owens (Eds.) , *Biodata handbook: Theory, research and use of biographical information in selection and performance prediction* (pp. 237– 261). Palo Alto, CA: CPP Books.

Schoenfeldt, L. F., & Mendoza, J. L. (1988, August). *The content and construct validation of a biographical questionnaire*. Paper presented at the annual meeting of the American Psychological Association, Atlanta.

Sharf, J. C. (1994). The impact of legal and equal opportunity issues on personal history inquiries. In G. S.Stokes, M. D.Mumford, & W. A.Owens (Eds.) , *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 351– 390). Palo Alto, CA: CPP Books.

Stokes, G. S., Mumford, M. D., & Owens, W. A. (1994). *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction*. Palo Alto, CA: CPP Books.

Tenopyr, M. L. (1994). Big five, structural modeling, and item response theory. In G. S.Stokes, M. D.Mumford, & W. A.Owens (Eds.) , *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 519– 534). Palo Alto, CA: CPP Books.

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