PERSONNEL PSYCHOLOGY
1977, 30

BEHAVIORAL OBSERVATION SCALES FOR PERFORMANCE APPRAISAL PURPOSES

GARY P. LATHAM[1]
University of Washington

KENNETH N. WEXLEY
University of Akron

Behavioral items ($N = 78$) critical to the job success of logging supervisors were developed from 1204 critical incidents. the frequency with which a supervisor ($N = 300$) engaged in each behavior was rated on a 5-point Likert type scale by two sets of observers. A factor analysis reduced the items to 38 and 33, respectively, for the two sets of observers which in turn constituted 10 and 11 factors or criteria for performance evaluation purposes. Multiple regression equations based on composite scores were used to predict cost-related measures of logging crew effectiveness. The shrinkage in *R*s after double cross-validation was moderately small. Moreover, the behavioral observation scales (BOS) that were developed by factor analyzing the observation ratings had moderately high reliability and accounted for more variance in the cost-related measures than did the BOS developed by traditional judgmental clustering techniques. The similarities and differences between BOS and BES procedures are discussed.

THERE are at least three approaches that a manager can take in evaluating the performance of an individual employee. He can examine cost-related variables, he can make judgments on traits or attitudes, or he can observe and record behavior.

Managers, stockholders, and consumers are generally concerned with cost-related variables such as profits, product quantity and quality, and returns on investment. These variables may serve as excellent indicators of an organization's effectiveness, but they are generally inadequate measures by themselves of an individual employee's job performance. This is because they do not inform the employee "how"

[1] Requests for reprints should be addressed to Gary Latham, Department of Psychology NI-25, University of Washington, Seattle, Washington 98195.

Acknowledgement is given to R. Alexander, D. Dossett, L. Kendall, and T. Mitchell for their constructive criticisms regarding this manuscript, and A. Zinsser for his assistance in conducting the statistical analyses.

or "why" he is effective or ineffective. Thus, he has minimum rather than maximum knowledge as to what he must do to maintain or improve his performance. Moreover, such measures are frequently excessive in that they are affected by factors over which the individual has little or no control (e.g., season, equipment, economic condition). Finally, they are often deficient in that they do not incorporate many of the key aspects of an employee's job. For example, assessing a typist's performance solely in terms of speed and accuracy would ignore such critical areas as setting priorities.

Performance appraisals based on traits and attitudes can foster misunderstanding and disagreement between the manager and his subordinates. A manager may tell an employee that he needs to show more initiative, become a better listener, and follow through on projects. This could be sage advice. But in its present form, the advice is not very helpful because it does not indicate *what* exactly the individual has to *do* differently. The employee may interpret the advice in ways in which the manager never intended or he may become hostile toward the manager, because he believes that he is already engaging in these behaviors.

As a result, psychologists have become increasingly vocal about the need to measure and evaluate the employee in terms of observable behaviors that are critical to job success or failure (Campbell, Dunnette, Lawler, and Weick, 1970). Behaviorally based measures are a more direct measure of what the employee actually does, and they are less influenced by factors not under the control of the employee than are cost-related indexes or managerial inferences as to a person's attitudes or traits.

The most frequently used procedure for developing behavioral criteria for performance appraisal purposes is the critical incident technique (Flannagan, 1954; Fivars, 1975). The critical incident technique (CIT) is a job analysis procedure whereby individuals who are aware of the aims and objectives of a given job, who frequently observe people performing that function, and who are capable of determining whether the job requirements are being performed satisfactorily are interviewed. Each interview focuses on the same three questions: (1) What were the circumstances surrounding this specific incident? In other words, what was the background or context? (2) What exactly did this individual do that was so effective or ineffective? If an individual is described as being aggressive, a team player, a self-starter, or the like, the question is asked as to what the individual actually did that led the observer to that conclusion. In short, what was the observable behavior? Descriptions of vague traits or attitudes are documented in terms of overt action. (3) How is this incident an example of effective or ineffective behavior? This is analogous to saying in a diplomatic manner, "So what? Tell me what this has to do with job performance."

The incidents are then categorized (typically by the job incumbents working under the direction of a researcher familiar with the CIT). Incidents which describe essentially the same behavior are grouped into one cluster. A descriptive behavioral item is formulated on the basis of these incidents. Clusters which are similar are grouped together to form one overall criterion or behavioral observation scale (BOS). For example, two incidents which describe an individual's development of a thorough project plan might form a behavioral item such as "develops a project plan prior to conducting the project." This item, along with similar items might form an overall behavioral scale labeled, "Planning and Scheduling." Thus, similar incidents are grouped together to form a behavioral item, and behavioral items which are similar are grouped together to form a behavioral criterion.

The resulting performance appraisal instrument consists of one or more behavioral observation scales. These scales may take the form of a dichotomous (checked, unchecked) rating or a Likert-type rating (e.g., Kirchner and Dunnette, 1957). The advantage of a Likert-type rating with more than two points is that it allows a more precise assessment of the frequency with which an individual engages in a given behavior than does a dichotomous rating. An example of a behavioral item for appraising a salesman on a five-point rating scale is:

"Knows the price of competitive products."

Never	Seldom	Sometimes	Generally	Always
1	2	3	4	5

The manager simply records the frequency (0–19%, 20–39%, 40–59%, 60–79%, 80–100%) with which he has actually observed the employee demonstrate this behavior.

Analyses conducted on data from a sample of logging supervisors have shown that the BOS is both reliable and relevant. Reliability was measured in terms of *intra*observer and *inter*observer agreement.

In the first report, Ronan and Latham (1974) examined the ratings of logging supervisors by two sets of observers on each of 78 behavioral items. That is, correlations were made between a rating of a specific behavior observed one month and the rating recorded by the same observer in a second month. In addition, paired *t*-tests were conducted on the differences between the two pairs of ratings from each observer. Seventy-seven out of 78 *intra*observer correlations were .50 or higher for one sample of observers and 64 out of 78 *intra*observer correlations exceeded this value for a second sample of observers. Only 13 out of 78, and 11 out of 78 *t*-tests were significant for the two respective groups of observers. Chance factors plus real changes in an individual's behavior from month to month could easily explain the discrepancies between the two monthly observations. Thus, the *intra*observer reliability of the BOS items was considered satisfactory.

In the second report, Latham, Wexley, and Rand (1975) examined the *intra*observer reliability of the eight composite criterion scales that had been developed by subjectively clustering the 78 individual items. The correlations ranged from .64 to .82 for one group of observers and from .67 to .85 for the second group of observers.

Ronan and Latham (1974) found that the *inter*observer agreement on the ratings of the individual behavioral items was relatively low. However, the *inter*observer agreement on each of the eight composite criterion scores obtained by Latham et al. (1975) ranged from .43 to .67 and .44 to .65 for the two respective groups of raters. That the *inter*observer reliability coefficients for the composite scores were somewhat lower than the *intra*observer reliabilities may have been due to the fact that the two sets of observers were from two different occupational groups.

Since the purpose of collecting the critical incidents was to identify behaviors critical to excelling on cost-related measures, the relevance of the BOS was tested using a concurrent validity model. Specifically, Ronan and Latham (1974) examined the relationship between the behavioral items and measures of productivity, turnover, absenteeism, and injuries. Multiple *R*'s that were not cross-validated ranged from .16 to .27 and .19 to .31 for the two groups of observers ($p < .001$). Latham et al. (1975) used a multiple regression and a double cross-validation design in determining the predictability of the economic constructs from all 8 BOS composite scores. The average cross-validation coefficients were .37 and .37 for productivity; .31 and .32 for absenteeism; and .45 and .43 for attendance based on the two sets of observers' ratings.[2]

Latham et al. (1975) also examined the intercorrelations between the behavioral criteria. The respective mean *r*'s for the two sets of observers were .46 and .58. It would appear that halo error may have been committed to some extent. This may have been due to procedures

[2] The difference between measures of absenteeism and attendance have been discussed by Latham and Pursell (1975, 1977). In brief, the authors found that a measure of attendance had significantly higher reliability (stability) than a measure of absenteeism when the latter is not literally defined as the converse of attendance, but rather includes some subset of absenteeism.

used for developing the criteria. That is, the BOS were developed by means of a qualitative cluster analysis (Campbell, Dunnette, Arvey, and Hellervik, 1973) as is traditionally done in a critical incident study. Judges simply sorted the incidents into their respective category clusters.

The purpose of the present study was to determine whether BOS could be improved by developing them through quantitative methods. The underlying assumption of this research was that developing composite scales with greater internal consistency might improve their generalizability as evidenced by the cross-validation coefficients of scales based on factor analysis rather than subjective clustering. Specifically, the reliability and relevance (concurrent validity) of the quantitatively derived BOS composite scales were compared with the BOS developed subjectively by Latham et al. (1975).

Method

Subjects

The behavioral ratings on the supervisors (i.e., producers) of the 300 logging crews previously used by Latham et al. (1975) were employed in this study. Half of these producers had been randomly selected as a holdout group of cross-validation. Behavioral items ($N = 78$) critical to success or failure in logging had been developed from 1,204 critical incidents according to the procedures described in the introduction. The frequency with which the logging supervisor engaged in each of the 78 behaviors had been rated on a 5-point Likert type scale by observers from two distinctly different populations, namely, dealers (i.e., independent wholesalers) who purchase wood from the crews and foresters who procure wood from the dealers for wood products companies. Both dealers and foresters observed the logging crews in the woods at least once a week for three months. To minimize criterion contamination, the data on productivity (cords/manhour), absenteeism (number of men off job/number of men in crew), and attendance (number of men on job/number of men in crew) were collected for 12 consecutive weeks by a third group of individuals who operated the weighting scales in the company's woodyards.

Procedure

In order to develop quantitatively derived BOS, the ratings (always, generally, sometimes, seldom, never) for each of 78 job behaviors on the 300 producers were subjected to factor analysis. Since the dealers and foresters had made their observations independently once a month for three consecutive months, it was possible to do two separate factor analyses. The first factor analysis was based on the average response over the three months to each behavioral item completed by the dealers; the second analysis was done using forester observations.

Using Kaiser's criterion of an eigenvalue of 1.0 or greater, a principal components factor analysis yielded 18 factors. The resulting factor structure was too complex to permit clear interpretation. Since, to a certain extent, Kaiser's criterion is artificial (Zavala, 1971) the analysis was reduced to an 11-factor solution. The 11-factor limitation using an oblique rotation was chosen because it gave the fewest number of items having high loadings on more than one factor, and the greatest number of factors having as many high loading items as possible.[3]

An item was required to have a coefficient of ±.30 or greater before it was considered to load on that factor. Application of this requirement caused one of the factors from the analysis of the dealer's data to fall out since no items loaded on it. This resulted in 10 behavioral criteria derived from dealer observations and 11 behavioral criteria derived from forester observations. A composite score on a factorially based criterion was the algebraic sum (unit weights) of the scores of those items that loaded positively on a factor minus those items that loaded negatively.

Results

Factor Names

Upon reviewing the item content, the following nine names were applied to the factors developed from the dealer and forester data: Interaction with Associates and Crew; Responsible Behavior; Financial Equipment Management; Independent Business Behavior; Pay Methods; Safety; Manpower and Equipment Management: Management Initiative; and Equipment Usage. The remaining factor based on dealer observations was called Emotional Control while the two remaining forester factors were Supervision, and Developing Standards and Competition. It should be noted, however, that even where two criteria were given the same name (e.g., Safety), the items comprising the criteria based on dealer observations were not necessarily the same as those which constituted the criteria based on forester observations. However, there was no reason to believe that foresters and dealers would define safety or any other criterion in terms of the same set of

[3] An oblique rotation was used because there was no reason to believe that the criteria were conceptually independent. In fact, the dealers and the foresters believed that the criteria were logically related to one another. To use a traditional orthogonal rotation of factors would have been tantamount to forcing statistical independence and thus creating an artifact in the data (Cattell, 1966; Harmon, 1960).

behaviors because people's perceptions of behavior may vary as a function of their position in an organization (Miner, 1968).

Intraobserver Reliability[4]

Pearson product-moment coefficients were calculated between the dealer criterion scores reported during the first month versus the second month, the second month versus the third month, and the first month versus the third month. This procedure was repeated for the observations recorded by the foresters. The *r*'s were averaged using the Fisher *Z* transformation. The results are shown in Table 1.

It can be seen from the data that what an observer reported a producer doing one month had a high correlation with what he reported the other two months. Moreover, both the range (.66–.84) and average reliability (.78) of the dealers' observations are comparable to those found by Latham et al. (1975) with the qualitatively derived BOS (range = .64–.82, average = .74). Similarly, the range and average reliability of the observations made by the foresters in the present analysis was .72–.90 and .80, respectively. These correlations are also comparable to those obtained with the qualitatively derived BOS (range = .67–.85, average = .80).

Intercategory Correlations

Using the average of the three monthly observations reported by the dealers, an intercorrelation matrix of the 10 criterion scores was developed. This same procedure was applied to the 11 forester criteria. For the sake of space, only the average intercorrelations of each BOS criterion with the remaining criteria are shown in Table 2.

With respect to the dealers' observations, Pay Methods (Criterion 5) had the lowest correlation with the other behavioral criteria, while the three management criteria (3, 7, and 8) and the Interaction criterion (1) had the highest correlations. It can be seen from Table 2 that the same pattern of intercorrelations did not necessarily hold for the foresters' observations.

Relevance

Using two groups of 150 producers each (designated as Samples A and B), separate multiple regression equations with stepwise inclusion were developed for each cost-related measure. The regression equations developed from Sample A (or B) were then cross-validated by applying them to Sample B (or A). This procedure was carried out separately on the dealers' and foresters' data. It can be seen from

[4] Since the factor based BOS frequently differed in item content for the foresters and dealers, the interobserver reliability of the criteria could not be measured.

TABLE 1

Intraobserver Reliability Coefficients of the Composite Scales Based on Factor Analysis

Cri- terion		Inter- action	Resp. Beh.	Finan. Equip. Mgt.	Indep. Busin. Beh.	Pay	Safety	Manpow. Equip. Mgt.	Mgt. Inita.	Equip. Mgt.	Emot. Control	Superv.	Devel. Stds. Compet.
Observer	Dealer	.79	.78	.84	.81	.74	.80	.73	.81	.84	.66	—	—
	Forester	.84	.79	.90	.81	.82	.80	.79	.77	.72	—	.76	.74

TABLE 2

The Average Intercorrelation of Each Behavioral Observation Scale with the Remaining Scales

				Finan.	Indep.			Manpow.					Devel.
Cri-		Inter-	Resp.	Equip.	Busin.			Equip.	Mgt.	Equip.	Emot.		Stds.
terion		action	Beh.	Mgt.	Beh.	Pay	Safety	Mgt.	Inita.	Mgt.	Control	Superv.	Compet.
Observer	Dealer	.59	.46	.54	.46	.25	.44	.54	.56	.43	.30	—	—
	Forester	.38	.41	.47	.40	.19	.37	.44	.28	.30	—	.20	.45

Table 3 that the significant relationships previously found between the qualitatively derived BOS and productivity, absenteeism, and attendance measures were also obtained in the present study. The small amount of shrinkage in the present validity coefficients is indicative of the generalizability of the regression equations.

A comparison was made between the results previously obtained with qualitative BOS and the results obtained in the present study with regard to the mean cross-validated correlations. The mean cross-validated coefficients based on the quantitative BOS using dealer observations were significantly greater ($p < .05$) than those obtained between all 8 qualitative BOS and productivity, absenteeism, or attendance. A significantly greater ($p < .05$) cross-validated correlation based on forester observations was also found in the present study for the measure of attendance than was the case with the 8 qualitatively based BOS.

The measures of productivity, absenteeism, and attendance were relatively independent of each other. Intercorrelations of −.50, .36, and .32 were obtained between attendance and absenteeism, productivity and attendance, and productivity and absenteeism, respectively.

Discussion

The gain in utility of the quantitative over the qualitative BOS is evident when one realizes that the multiple *R*s obtained by using the qualitative BOS were based on observations of all 78 behavioral items that constituted the 8 judgmental composite scales. On the basis of the factor analysis it was possible to use a much smaller number of items (see Table 3). This decrease in the number of items required for appraisal is extremely important for managers who often complain that they do not have enough time to conduct lengthy performance appraisal sessions.

The results of this study indicated that the behavioral observation scales that were developed by factor analyzing observation-ratings had moderately higher reliability and accounted for as much if not more variance in the cost-related variables as the BOS developed by qualitative methods. This is probably due to the fact that the factor based scales made more efficient use of the information in the item pool which resulted in composite scores that had more stable relationships being used as predictors. The factor based development of composite scores is similar to the approach recommended by Gleason and Staelin (1973) for improving the quality of fallible data. In fact, the stability of the factor based composites was higher and the intercorrelations were lower than those obtained with the qualitatively

TABLE 3

Multiple Correlation and Cross-Validated Coefficients of Regression Equations Showing the Relationship of Quantitatively Derived BOS to Cost-Related Measures

Cost-Related Measures	Observer	Initial Sample		Cross-Validated Correlations[a]		Mean Cross-Validated r[b]	Latham et al. Mean Cross-Validated r[c]
		A	B	B	A		
Productivity	Dealer	.48***	.61***	.57***	.45**	.50***(5,30)	.37***(8,78)
	Forester	.55***	.40***	.36***	.37**	.36** (4,19)	.37***(8,78)
Absenteeism	Dealer	.48***	.56***	.49***	.41**	.46***(6,44)	.31***(8,78)
	Forester	.36**	.43**	.33**	.31*	.32** (6,44)	.30***(8,78)
Attendance	Dealer	.68***	.65***	.64***	.67***	.66***(8,63)	.45***(8,78)
	Forester	.66***	.68***	.59***	.65***	.62***(7,50)	.43***(8,78)

[a] The regression equation developed from the initial Sample A (or B) was cross-validated by applying it to Sample B (or A). $N = 150$ in each sample.

[b,c] The first number in the parentheses indicates the number of behavioral scales used as a predictor; the second number indicates the number of behavioral items on the scales.

* $p < .05$.

** $p < 0.1$.

*** $p < .001$.

based composites. These two conditions contributed to the magnitude of the cross-validated correlations.

Nevertheless, it could be argued that the superiority of the factor analytically derived BOS was a result of using the same sample of subjects for both the factor analysis and the multiple regression equations. This argument is weakened by the fact that the factor analysis involved the internal relationships among the predictor items without regard to their relationships to any external criterion. Thus, it doesn't seem likely that this biased the validity coefficients.

The problem of not having an additional sample on which to conduct the factor analysis was further minimized by computing composite scores using unit weights rather than using the factor loadings. This, in turn, increased the stability of the factors that were used in the multiple regression equations. Furthermore, we were not concerned here with the stability of the beta weights used in the regression equations, but rather the magnitude and stability of the cross-validation coefficients which were found to be comparable between the two subsamples of 150 producers each.

Two major disadvantages of quantitatively derived BOS are time constraints and sample size. Qualitatively derived criteria can be developed immediately after the critical incidents are collected. This can be done regardless of the size of the population that will ultimately be evaluated on the basis of these BOS. On the other hand, quantitatively derived BOS typically require a sample size of several hundred people. A survey instrument containing the individual behavioral items is sent to the future users of the final performance appraisal instrument. They complete the instrument in the field as was described in the present study. Finally, the results are collected for statistical analyses to determine (quantitatively) the clusters (BOS) in which the respective items belong. The increase in preparation time for following this approach would appear justified on the basis of the savings in time for the users of the final appraisal instrument who complete fewer items than those who use the qualitatively derived BOS. However, where the population size is small, qualitatively derived BOS should obviously be used.

A problem with using either type of BOS is gaining the acceptance of managers who question its relevance. Many managers argue that as results oriented people, they are only concerned with cost-related outcomes. Their resistance is based on the belief that, in general, a person can do well on a BOS and do poorly in terms of task outcomes. The present study has shown that this belief is unfounded. The correlations obtained between the behavioral and cost-related variables are high, particularly in light of the measurement problems inherent in cost-related outcomes. The advantage of the behavioral criteria is that it allows the manager to determine "how" and "why" performance on cost-related variables can be maintained or improved.

The BOS is similar to BES or behavioral expectation scales (Smith and Kendall, 1963; Zedeck and Baker, 1972) in that (a) both are variations of the critical incident technique (Flanagan, 1954); (b) both use rating forms that are worded in the terminology of the user; (c) both rating scales have face validity in that both are based on relatively observable job behaviors that have been seen by others as critical to task success, and (d) both take into account the multi-dimensionality or complexity of job performance.

The BES differ in at least two important ways from the BOS. First, BES typically require that each criterion be arranged on a continuous vertical graphic rating scale with a behavioral incident or anchor listed beside each of 7–9 points ranging from ineffective to effective behavior (e.g., Campbell, Dunnette, Arvey, and Hellervik, 1973). Second, the manager simply examines the respective criterion and places a check mark beside the one behavioral anchor that he believes best describes the behavior that the employee could be *expected* to demonstrate. This expectation presumably is based on what the manager has seen the employee do over a given period of time (e.g., 3–6 months). A potential problem is that the behaviors that the manager has seen the employee demonstrate may not resemble the specific anchors on any of the scales. Thus, the manager is required to extrapolate from observed behaviors to those which could be "expected" as defined by the scale anchors.

The BOS requires no such extrapolation. Each critical behavior is listed in a questionnaire format and the manager indicates the frequency with which he has observed the behavior. However, a potential problem with BOS is the time required for its completion. Eight criteria on which an individual is to be evaluated using a BES requires only eight check marks. Eight criteria on a BOS could conceivably require as many as 80 or more ratings (e.g., if there were 10 or more behaviors listed under each criterion). This level of detail required for appraising the employee, however, may be the major strength of this method.

Kendall (personal communication) has argued that the primary distinction between the BES and the BOS is two-fold. First, only the BES is concerned with the scaling of an observed behavior. The anchors are simply examples that aid the observer in weighting (scaling) *observed* behaviors on a given criterion. Second, the BES procedures, unlike the BOS, do not specify what specific behaviors are to be observed.

The limitation of the BES regarding the actual behaviors that are sampled has been commented upon by Schwab, Heneman, and De

Cotiis (1975). The limitation of the BOS regarding its failure to explicity focus on the quality or value of given behaviors would appear somewhat minor in light of the correlations between the BOS and the cost-related measures for which the former were explicitly developed to predict.

Future research needs to determine whether BOS are more reliable and relevant than BES. It is likely that the answers will vary depending upon the amount of contact between a manager and his subordinate. The BOS would seem preferable in instances where there is a high degree of contact between the manager and his subordinates. The BES may be preferable when there is minimal opportunity for the manager to observe a subordinate.

REFERENCES

Campbell, J. P., Dunnette, M. D., Arvey, R. D., and Hellervik, L. W. The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology*, 1973, 57, 15–22.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., and Weick, K. E. *Managerial behavior, performance, and effectiveness*. New York: McGraw-Hill, 1970.

Cattell, R. B. *Handbook of multivariate experimental psychology*. Chicago: Rand McNally, 1966.

Fivars, G. The critical incident technique: A bibliography. *JSAS Catalog of Selected Documents in Psychology*, 1975, 5, 210.

Flanagan, J. C. The critical incident technique. *Psychological Bulletin*, 1954, 51, 327–358.

Gleason, T. X. and Staelin, R. Improving the metric quality of questionnaire data. *Psychometrika*, 1973, 38, 393–410.

Harmon, H. H. *Modern factor analysis*. Chicago: University of Chicago Press, 1960.

Kirchner, W. K. and Dunnette, M.D. Identifying the critical factors in successful salesmanship. *Personnel*, 1957, 34, 54–57.

Latham, G. P. and Pursell, E. D. Measuring absenteeism from the opposite side of the coin. *Journal of Applied Psychology*, 1975, 60, 369–379.

Latham, G. P. and Pursell, E. D. Measuring attendance: A reply to Ilgen. *Journal of Applied Psychology*, 1977, 62, 234–236.

Latham, G. P., Wexley, K. N., and Rand, T. M. The relevance of behavioral criteria developed from the critical incident technique. *Canadian Journal of Behavioural Science*, 1975, 7, 349–358.

Miner, J. B. Management appraisal: A capsule review and current references. *Business Horizons*, 1968, October, 83–96.

Ronan, W. W. and Latham, G. P. The reliability and validity of the critical incident technique: A closer look. *Studies in Personnel Psychology*, 1974, 6, 53–64.

Schwab, D. P., Heneman, H. G., and De Cotiis, T. A. Behaviorally anchored rating scales: A review of the literature. PERSONNEL PSYCHOLOGY, 1975, 28, 549–562.

Smith, P. C. and Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 1963, 47, 149–155.

Zavala, A. Determining the hierarchical structure of a multi-dimensional body of information. *Perceptual and Motor Skills*, 1971, 32, 735–746.

Zedeck, S. and Baker, H. T. Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. *Organizational Behavior and Human Performance*, 1972, 7, 457–466.