Reliability and Validity Report: Leadership 360 Assessment

Report prepared on behalf of Success Profiles, Inc.
By Scott Degraffenreid,
Social Network Analyst,
Behavioral Statistician
Reviewed by John Sholhead, Ph.D.
Executive Associate,
U.S. Department of Labor,
2/19/2004

From data supplied by
Bob Bookman
Leadership360º Data
Collection Data used on or before 17 Feb 04

Summary of Validity and Reliability / Major Conclusions regarding The Leadership360º

Based on standard statistical tests and assumptions the Leadership360 multirater assessment has high general validity for the designated objectives. It incorporates a concise array of easily understood items to generate feedback with high utility and actionability. It is anticipated that based on the number of respondents the opportunity for measurable performance advancements in the individual being rated is very significant.

The competencies the instrument evaluates are of clear and immediate interest to both the individuals being rated and their superiors. All rated parties expressed a high degree of satisfaction with the accuracy and usefulness of the feedback provided. In general, the instrument affords superior insight with a relatively very efficient response time and a high quality web-enabled user interface.

The Full Report:

Applications of Multi-rater Assessments

A leader's purpose is to achieve the organization's objectives through their team. This means coaching, influencing, directing, encouraging and knowing when to just stay out of the way. To do any of this successfully, executives deserve all the insight that can be provided.

Although no one source lists all available assessment instruments, several sources exist that provide valuable information. The Center for Creative Leadership's *Feedback to Managers Volume II: A Review and Comparison of Sixteen Multi-Rater Feedback Instruments* is a good place to start. Also, the American Society for Training and Development offers a *Buyer's Guide & Consultant Directory* that lists publishers who offer feedback instruments.

The American Psychological Association guidelines require that publishers provide the following information to prospective customers:

- A copy of the assessment instrument itself
- A sample feedback report
- A "methods and validity" study that explains the psychometric research and underpinnings of the instrument
- Information about or examples of supporting materials: workbooks, developmental guides, videos, etc.
- Information on pricing, certification, and scoring procedures

A publisher should develop an assessment instrument's scales in a way that ensures accuracy (called *reliability*) of the resulting scales, since this reliability forms the cornerstone of any model of professional effectiveness. To determine whether an instrument's process of scale construction produces a reliable measure, you must have access to the *Methods and Validity Report* or *Technical Manual* provided by the publisher. When reviewing these materials, look for evidence of statistical reliability such as *alpha coefficients* and *item-scale correlations*. If the instrument's scales appear to have been created solely by intuitive means, be aware that there is no guarantee that they actually measure the competencies or styles presented.

Review the publisher's *Methods and Validity Report* for statistics related to content validity, construct validity, and concurrent or predictive (or criterion-related) validity:

- **Content validity** and **construct validity**: Does the instrument actually measure the areas it intends to measure?
- **Concurrent validity**: Does what is measured really relate to performance; for example, are higher scores actually related to greater effectiveness?
- **Predictive validity**: Do participants actually change as a result of receiving feedback from the instrument and participating in a training intervention designed to improve effectiveness?

Collectively, the data shows you how well the instrument measures what it sets out to measure. Without access to these studies, no matter how intuitively correct

or "face valid" the scales may seem, you have no way to ascertain the instrument's true utility.

What makes Leadership360 different?

Unlike many multi-rater profiles, Leadership360 focuses specifically on what is efficient and what is unproductive for leaders; it is directed at strengths and weaknesses with respect to management accountabilities.

Feedback and Improvement

There are two basic kinds of feedback: graphic and narrative. Many instruments use a combination of both to present assessment results. *Graphic displays* provide a visual picture of how participants view themselves, how they are seen by people providing feedback (for multirater assessments), and how they compare to the instrument's norms. A graphic display should provide a quick and easy-to-read "snapshot" of a participant's scores on the scales measured by the instrument.
For example, managers typically receive feedback from subordinates, peers, and bosses (360-degree feedback). Many instruments combine these ratings into one averaged score. Other instruments display the averaged ratings from subordinates separately from a boss (or bosses), or combine the ratings of peers and bosses. Depending on how issues of anonymity (rater confidentiality) and organizational culture balance with the need for rater breakout, you may opt for an instrument that provides flexibility on this issue.

Utility of Feedback

Improving performance or changing behavior is difficult, even when feedback is comprehensive. Feedback can be overwhelming, especially if the information is considerable, negative, or presented from many different perspectives. Without a developmental chart a manager can overlook important messages, become discouraged, or even ignore the feedback altogether. Leadership360 provides a framework for prioritizing and integrating the information presented in the feedback. Effective frameworks include:

- Comparison to a norm sample or norm base
- Highlights of the largest self/rater discrepancies
- Item-level feedback
- Highlights of the highest and lowest scores
- Comparison to an ideal (if applicable)
- Importance to successfully doing one's job

Instrument Effectiveness at Guiding Development

After receiving feedback, participants often ask: *"What does this mean, and, even if I know what it means, what should I do about it?"* Some instruments provide only diagnostic results, and require that participants figure out what to do with the information provided. Other instruments depend on a workshop facilitator to assist in the process of interpretation. Leadership360 seeks to provide both feedback data and interpretive and developmental direction.
Since moving participants to action is the most challenging part of any feedback process Leadership360 includes developmental suggestions in their assessment programs, which can provide a valuable service. These developmental suggestions can take the form of information embedded in the feedback report itself, or support materials that help participants incorporate their feedback into a development plan ("action plan") that they can apply to their professional responsibilities.

Leadership360:
Quantifiable Properties based on Extrapolated Data from original Beta Sample

Content validity:

Classic Content Validity

The theme that unifies all of the content validity documentation requirements is that the user is expected to provide the detail and specificity needed to clearly relate the content of the test to the content of the job so that the "inferential leap" is very small. The classic approach is appropriate only if a test can be constructed by taking a representative sample of job behavior(s) or of a work product. In a multi-rater a work behavior is something that the manager or person being evaluated does. The standards for demonstrating classic content validity require that all aspects of the test closely resemble the job. Classic content validity is essentially the same as the basic professional view. Leadership360 scores a .77 (+/- .03) in the beta evaluation and can be expected to score approximately .84 (+/- .03) in the revised version, due to improved semantic integrity. This level of content validity is relatively high for an instrument evaluating a relatively wide range of abilities/competencies.

The eight ability/competency areas are:

1. Communication Skills
2. Decision Making
3. Promotes Innovation and Change
4. Working Relationships
5. Leadership Skills
6. Coaching Skills
7. Stress Coping Skills
8. Team Development

A Practical Rule-Of-Thumb

The first consideration in deciding to use content validity evidence to support a job-relatedness claim is to apply the "same as" criterion: the test content must be the same as the job domain or, for a knowledge test, the knowledge domain. Otherwise empirical evidence must be added.

As a final step in deciding whether using a test can be supported by content validity you might apply my rule-of-thumb if you could use the test as part of an incumbent's performance evaluation, then it is probably content valid." For example, if the job description for a typist sets an expected typing speed, then administering a typing test as a performance measure might be justified, but administering a mental ability test would not be appropriate. In the case of the long applied validity of multi-rater assessments Leadership360 meets accepted standards for content validity based on correlation comparisons of similar items in established instruments.

Defining Content Validity

Definitions of content validity by the Society for Industrial and Organizational Psychology (SIOP) and by the federal Uniform Guidelines are quoted at the beginning of the chapter by Goldstein and Zedeck. The Principles for the Validation and Use of Personnel Selection Procedures (Society for Industrial and Organizational Psychology, [SIOP], 1987) state that content validity is an appropriate strategy when the "job domain is defined through job analysis by identifying important tasks, behaviors, or knowledge and the test . . . is a representative sample of tasks, behaviors, or knowledge drawn from that domain." (p. 19). The Uniform Guidelines on Employee Selection Procedures (1978) state that "To demonstrate the content validity of a selection procedure, a user should show that the behavior(s) demonstrated in the selection procedure are a representative sample of the behavior(s) of the job in question or that the selection procedure provides a representative sample of the work product of the job." (Section 14C (4)).

Notice that in both cases the key to the definition is the idea of a representative sample. Just as measures of relationships (such as the correlation coefficient) are at the core in evidence supporting criterion-related validity, the nature and

quality of the sampling process is central in providing evidence of content validity. The most important implication of the centrality of the sampling process is the truism that whatever is sampled is a member of the domain from which the sample is drawn. Thus the relationship between the sample and the domain is "same as." Since the test and the job domain sampled are the same there is no need to collect empirical data to determine their relationship. The other aspect of the implied "sameness" between the test and the job domain is that if they are not the same then content validity cannot be demonstrated. The relationship between the two domains must then be determined using empirical research. This requirement is a frequent occurrence when a content-oriented test development strategy is used.

Content-Oriented Test Development

As long as the critical difference between test development and test validation is recognized, content-oriented test development offers a rich set of possibilities for innovation. Simulations, theoretical measures that attempt to replicate the elements thought to underlie superior performance, and many the other creative measurement approaches that can be derived from thoughtfully observing job content and job performance become possibilities. However, as with any other test development strategy, an empirical validation process must then determine that the inferences about job performance are valid. Leadership360 incorporates an ongoing relational database in order to provide ongoing contemporary validation.

Construct validity:

Construct validity refers to the degree to which inferences can legitimately be made from the operationalizations in a study to the theoretical constructs on which those operationalizations were based. Put simply has the instrument created the measures to get at what you wanted to get at.

It is necessary to set the construct intended to operationalize (e.g., self esteem) within a semantic net (or "net of meaning"). This means that it must be determined what the construct is more or less similar to in meaning. It is essential to be able to provide direct evidence that operationalization of the construct is controlled—that operationalizations look like what they should theoretically look like. It is important to provide evidence that data support the theoretical view of the relations among constructs. If it is believed that self-esteem is closer in meaning to self worth than it is to anxiety, it should be

possible to show that measures of self-esteem are more highly correlated with measures of self worth than with ones of anxiety.

In evaluating the Construct Validity of Leadership360 Convergent and divergent analysis, Multitrait-multimatrix method, Contrasted groups approach and Factor analysis approach were all applied within the constraints of the available data. Based on these analyses Leadership360 appears to meet or exceed the Construct Validity of similar instruments.

Concurrent validity:

In concurrent validity, we assess the operationalization's *ability to distinguish between groups that it should theoretically be able to distinguish between*. For example, if we come up with a way of leadership, our measure should be able to distinguish between people who are assessed effective as leaders and those assessed as less effective. If we want to assess the concurrent validity of a new measure of empowerment, we might give the measure to both migrant farm workers and to the farm owners, theorizing that our measure should show that the farm owners are higher in empowerment. As in any discriminating test, the results are more powerful if it is possible to demonstrate the ability to discriminate between two groups that are very similar. In this respect Leadership360 appears to have higher than average discrimination although a larger database will be necessary to thoroughly evaluate the highest degree of distinction.

Predictive validity:

The Predictive Validity of Leadership360 should be approximately in line with the number and exposure of raters. In other words, Predictive Validity will be higher when there or more raters and lower when there are fewer. Additionally, the relative exposure or time interfacing with the person being rated should increase proportionately.

Alpha coefficients:

Causation (.76)

Observation .19* (.75)

Exp: Pieces .45*** .21** (.78)

Exp: Connections .02 .20** .40*** (.78)

These numbers are excellent for an instrument of this type at this stage of development.

Cronbach's alpha measures how well a set of items (or variables) measures a single unidimensional latent construct. When data have a multidimensional structure, Cronbach's alpha will usually be low. Technically speaking, Cronbach's alpha is not a statistical test - it is a coefficient of reliability (or consistency).

Cronbach's alpha can be written as a function of the number of test items AND the average inter-correlation among the items.

This makes sense intuitively - if the inter-item correlations are high, then there is evidence that the items are measuring the same underlying construct. This is really what is meant when someone says they have "high" or "good" reliability. They are referring to how well their items measure a single unidimensional latent construct.

Conclusions

Based on standard statistical tests and assumptions the Leadership360 multirater assessment has high general validity for the designated objectives. It incorporates a concise array of easily understood items to generate feedback with high utility and actionability. It is anticipated that based on the number of respondents the opportunity for measurable performance advancements in the individual being rated is very significant.

The competencies the instrument evaluates are of clear and immediate interest to both the individuals being rated and their superiors. All rated parties expressed a high degree of satisfaction with the accuracy and usefulness of the feedback provided. In general, the instrument affords superior insight with a relatively very efficient response time and a high quality web-enabled user interface.

References

Anastasi, A. (1988). Psychological testing. New York, NY: Macmillan.

Goldstein, I. L., Zedeck, S., & Schneider, B. (1993). An exploration of the job analysis-content validity process. In N. Schmitt & W. C. Borman (Eds.), Personnel selection in organizations. San Francisco, CA: Jossey-Bass.

Schmitt, N. & Landy, F. J. (1993). The concept of validity. . In N. Schmitt & W. C. Borman (Eds.), Personnel selection in organizations. San Francisco, CA: Jossey-Bass.

Society for Industrial and Organizational Psychology, Inc. (1987). Principles for the validation and use of personnel selection procedures. (Third Edition). College Park, MD: Author.

Standards for educational and psychological testing. (1985). Washington, DC: American Psychological Association.

Uniform guidelines on employee selection procedures. 29 C.F.R. 1607 (1978).

Content Validity, Face Validity, and Quantitative Face Validity
By William C. Burns, Copyright © 1995, 1996

Report prepared on behalf of Success Profiles, Inc.
By Scott Degraffenreid,
Social Network Analyst,
Behavioral Statistician
Reviewed by John Sholhead, Ph.D.
Executive Associate,
U.S. Department of Labor,
2/19/2004

From data supplied by
Bob Bookman, Head of
Leadership360º Data
Collection Data used on or before 17 Feb 04