An Evaluation of the Quality of 360-degree Assessment Instruments

Froukje Jellema
Adrie Visscher
Martin Mulder
University of Twente, The Netherlands

Recently 360-degree assessment has become popular. It can be used as well for evaluating employee performance as for the evaluation of training transfer. The instruments used for this kind of assessment, where managers are evaluated by co-workers such as their supervisor, peers, subordinates and clients, should meet certain standards (e.g. psychometric and personnel evaluation standards). A checklist of standards will be presented on the basis of which 360-degree instruments are examined. In this study four 360-degree assessment instruments are examined. The outcomes of the study give insight into the qualities of a number of 360-degree instruments.

Keywords: 360-Degree Assessment, Multirater Assessment, Training Evaluation

Introduction and Problem Statement

In a multirater assessment, job-performance of an employee is evaluated by one or more co-workers, such as supervisors, subordinates, peers, customers, and suppliers. If all of these sources are being used, and the employee also evaluates his or her own job performance, this is called a 360-degree (full circle) assessment. The assumption is that with each additional rater source, the confidence that the reported results are an accurate reflection of what is happening is increased (Robinson and Robinson, 1989).

360-degree assessment can be used for several purposes. First, it can be used as an instrument for personnel development, for example for analyzing the strengths and weaknesses of an employee. It can also, as an element of a training or development program, provide a trainee with a clear picture of his or her performance and training priorities. Furthermore, 360-degree assessments can be instruments in formal performance appraisal, for example as a basis for making salary or promotion decisions. Finally, 360-degree assessments could be used in training evaluation. In that case, the trainee's co-workers provide feedback, both before and after the training, and/or are asked directly for perceived changes in their colleague's job-performance, as a result of the training he or she has undergone.

The 360-degree assessment usually concerns a questionnaire, by means of which raters give their opinion about the ratee's job-performance. This questionnaire generally consists of several competencies that are considered relevant for the ratee's job-performance. For example, an instrument may focus on five groups of competencies, i.e. leadership, communication, management, decision making, and personal behavior. Each of these groups is measured by more specific behavioral characteristics. For instance, 'communication' may be measured by 'listening skills', 'written communication', 'oral communication' and 'presentation skills'. Each category is usually measured by means of several items.

Ten years ago, only twenty to thirty instruments were on the US market, now over one hundred are available (Lepsinger and Lucia, 1997). The quality of these instruments should be ascertained since it will not matter what the data collection process reveals about an individual, if the instrument lacks validity, reliability and applicability for the organization (Church, 1999). Some instrument evaluation has already been carried out (Morical, 1999; Van Velsor and Leslie, 1991). Several studies have focused on their reliability, (Nijhof and Jager, 1995) and validity (Church, 1999), and on the effects of 360-degree instruments for evaluating management development (Rosti and Shipper, 1998; McLean et al., 1995; Hazucha et al., 1993). However, most of these studies focus on US-instruments.

The increasing use of 360-degree assessments in Dutch organizations has resulted in the development of a considerable number of Dutch instruments. This popularity, however, has not been supported by research on their quality. This study is therefore meant to reveal more about the extent to which Dutch instruments meet the standards included in the checklist.

Theoretical framework

A 360-degree assessment focuses on personnel evaluation. The Joint Committee on Standards for Educational Evaluation has developed standards for personnel evaluation (1988). These standards focus as well on the quality of the evaluation instruments used as on the process of evaluation, and fall into four categories: propriety, utility, feasibility, and accuracy.

The *propriety standards* reflect the fact that personnel evaluations may fail to address, or violate certain ethical and legal principles. They include recommendations that are meant to promote the accessibility of evaluation reports and guidelines for interactions with the person being evaluated. These standards appear to be very important for 360-degree assessment, especially for the assessment process. Ratees as well as raters should be approached with care, since giving and receiving feedback can be threatening. Issues like the purpose of the assessment, or confidentiality should be communicated clearly.

Utility standards are intended to guide evaluations, so that they will be informative, timely, and influential. The evaluation is only informative is the right questions are being asked. Therefore, ratings should focus on competencies that are relevant to the job and should be based on function or task analysis. Items and response wording should also be clear and easy to be understood. Preferably, there is room for recommendations to give raters the opportunity to explain themselves and to provide more useful information. Other important standards for 360-degree assessment are functional reporting and follow-up. Yukl and Lepsinger (1995) developed some guidelines for the display of feedback in the final report. In general, the report should clearly identify feedback from different perspectives, compare feedback from others with the manager's own perceptions, compare the manager's ratings with norms, display feedback for items as well as scales (mean score, range, and distribution), and should provide feedback on recommendations. Follow-up appears to be a crucial factor in a 360-degree assessment (Hazucha et al., 1993). The trainee should be helped to understand the results and pursue appropriate actions.

The *feasibility standards* promote evaluations that are efficient, easy to use, viable in the face of social, political, and governmental forces and constraints, and that will be adequately funded. These standards may be problematic in relation to 360-degree assessment, since this kind of assessment may not be very practical and/or easy to use.

Accuracy standards aim at determining whether an evaluation has produced sound information. Though the other categories mainly focus on the assessment process, the accuracy standards refer especially to the 360-degree instrument. The instrument should be valid, reliable and control bias. Van Velsor et al. (1997) indicate minimal levels of internal consistency, interrater reliability, and test-retest reliability (see checklist).

When studying 360-degree assessment, the performance appraisal literature can also provide valuable input. Performance appraisal concerns the process of identifying, observing, measuring, and developing human performance within organizations (Cardy and Dobbins, 1994). The purpose, in general, is to improve employees' performance and to provide information that can be used in making work-related decisions. Cascio has done important work regarding performance appraisal. According to Cascio (1995) appraisal systems have to meet five key requirements: they should be acceptable, practical, relevant, sensitive and reliable. In the view of Cascio, validity can not be measured directly, since it is unknown what 'truth' is in performance appraisal. By making appraisal systems relevant, sensitive, and reliable, it can be assumed that the resulting judgments are valid as well. One of the main implications of performance appraisal theory for 360-degree assessment is the importance of using behaviorally focused items in questionnaires (Murphy and Cleveland, 1995). Questionnaires should always be formulated in terms that are easy to interpret by raters. If the assessment is used to measure training transfer, only characteristics that can be controlled and developed by the trainees should be used. To warrant reliability and bias control, raters should be trained to use the instrument properly.

Many authors in the field of 360-degree assessment have established guidelines or recommendations to guarantee proper use (Dalessio, 1998; Tornow and London, 1998; Van Velsor et al., 1997).

On the basis of the aforementioned references, a checklist has been developed for judging the quality of 360-degree instruments (Table 1).

Method

Four instruments for 360-degree assessment have been scored on each element of the checklist. It was attempted to examine all 360-degree instruments developed and sold by Dutch organizations. It is difficult to say how many Dutch instruments are available, but we expect that there are considerably more instruments than the ones that have

been examined here. However, it is supposed that the examined instruments give a good picture of what is available on the Dutch market. The organizations approached to participate in the study delivered a sample of their instrument, a sample of the 360-degree feedback report that ratees receive after the assessment, and all other required information. Additionally, vendors of the instruments were interviewed.

The scoring is based on document analysis (the questionnaires and feedback samples), and on the information provided by the vendor. For some instruments, studies on their reliability and validity are available that provide interesting additional information. If relevant, the results of these studies are presented here.

Table 1
A standards checklist for examining 360-degree instruments

Competencies	Ratings are made on competencies relevant to the job Competencies are based on function or task analysis
Items	Every competency is measured by several items Items focus on behavior that can be observed easily The item wording is clear Only factors over which the ratee has control are included Item wording and qualification are free of irrelevant characteristics, such as race, age, sex, religion
Adaptation	The questionnaire can be adapted to a specific situation
Response	Response scales are clear There is room for recommendations There is a non-response option ('don't know', 'not applicable')
Feedback report	Language and graphics are understandable Feedback is displayed from different perspectives Self-scores are compared to scores of other rater groups Feedback is reported for items as well as competencies The mean score of each item and competency is displayed The range of each item and competency is displayed Scores are compared to norms
Development	The instrument is based on a combination of theory, research and experience
Reliability	All scales have internal consistency coefficients (alpha) of at least .6 Interrater (within-group) reliability is at least .4 All scales have test-retest coefficients greater than .4
Validity	Research should be done to establish validity

Results

In this section the features of the examined instruments are described and their scores with respect to the judgment on the basis of the checklist are presented.

Instrument 1 is a 360-degree instrument which has been available since 1993. It is being used in many Dutch organizations as well as in Japan, Belgium and Great Britain. It is mainly used for management development, though recently a pilot started in which it is used for measuring behavioral change as a result of training. To use this instrument, certification is required and a coaching and follow-up process should be developed. Van der Woude (1995) examined the reliability and validity of this instrument, while Van der Giessen (1997) used it to examine the influence of 360-degree feedback over time on 12 selected behavioral characteristics. The results of these studies can be found in Table 2.

Instrument 2 was developed in 1997 in co-operation with a Dutch university (Rietveld, 1997). The instrument is being used in the Netherlands only. It focuses mostly on high-level managers and is used for developmental, appraisal and communication purposes. The use of the instrument as a training tool is still in the experimental stage. Rietveld (1997) has done some research, concerning the internal consistency and interrater agreement (supervisor versus others and self versus others) of the instrument. Boers (1999) formulated some new competencies and examined their internal consistency. The results of these studies are included in Table 2.

Instrument 3 was developed in 1997 and is being used in many Dutch organizations. Although its main purpose is the development of (higher level) managers and management teams, the instrument is also used in a training context, either before training (to assess the strengths and weaknesses of trainees), or after training, (to

assess further development areas). The intention is to use the instrument for measuring training effects. No research has yet been done on the reliability or validity of this instrument.

Instrument 4 was developed in 1997 and has been used in many Dutch organizations. Free discussion of feedback results is considered the main advantage of 360-degree assessment. The instrument is mainly used prior to training, to assess training needs. Although the instrument now is only sold in combination with training, the intention is to sell it in other contexts than the training context. Groeneveld (1997) examined the internal consistency of the competencies in the instrument. The results can be found in Table 2.

Table 2
The results of the comparison of four 360-degree instruments with the checklist

	Instrument 1	Instrument 2
Competencies	The user (organization/ratee) selects the relevant competencies Competencies are not based on function or task analysis	Ratee and supervisor select relevant competencies Competencies are based on experience with performance appraisal and function analysis
Items	All dimensions contain about 5 items Items are opposite statements about behavior Items focus on easy to be observed behavior Many items measure more than one skill/behavior Items focus on skills that can be developed Item wording does not explicitly contain bias characteristics	All dimensions are measured by 4-8 items Items do not focus on easy to observe behavior Items contain words as: often, sometimes, usually Statements are alternately positively, or negatively formulated Items focus on skills that can be developed Item wording does not explicitly contain bias characteristics
Adaptation	The instrument can not be adapted to a specific situation	The instrument can not be adapted to a specific situation
Response	1-5 scale, position between opposite statements No non-response option No room for recommendations	Response scales are clear 5-point agreement scales Non-response options No room for recommendations
Feedback report	Scores/graphics are explained and understandable Feedback is displayed from different perspectives Self-scores are compared to other-scores Feedback is reported for competencies and items The mean and range of each item and competency is displayed Scores are compared to norms The report includes a strength-weaknesses display (highest/lowest average scores) There are development directions	Graphics are explained and understandable Scores are percentages Feedback is displayed from different perspectives Self-scores are compared to other-scores Feedback is reported for competencies and items Item display is complex because of positively and negatively formulated items There is no mean score or range, since scores are presented as percentages Scores are not compared to norms There are some follow-up directions
Development	Developed on the basis of experience and research	Experience and some research
Reliability	Almost all alpha's are > 0.6 Interrater reliability is low Test-retest is moderate to low	Most alpha's are > .6 Interrater reliability is low Test-retest reliability has not been studied
Validity	Construct validity was studied and considered satisfactory	Validity has not been studied
Additional information		
Experience	six years of experience; been used in many Dutch organizations and in other countries	used for two years in many Dutch organizations
Purpose	Development	Development, communication, appraisal

(Table 2 continued)

	Instrument 3	Instrument 4
Competencies	127 items of which 96 focus on management skills and 31 on personality traits It is a standard instrument (no selection)	39 competencies Ratee and supervisor select relevant competencies Competencies are based on literature research, interviews and expert knowledge
Items	Items do not always focus on easy to observe	Every competency is measured by 5 items

	behavior (personality traits) Not all items can be developed or controlled by the ratee Item wording is clear Item wording does not explicitly contain bias characteristics	Many items do not focus on behavior that is easy to observe, but on personality traits Not all items are easy to be developed by the ratee Item wording is clear Item wording does not explicitly contain bias characteristics
Adaptation	The only adaptation is that items can be left out.	The instrument can not be adapted to a specific situation
Response	Response scales are clear 9-point frequency scale Respondent indicate what the desired score is Non-response option No room for recommendations	Response scales are clear 6-point applicability scale Non-response option Room for recommendations
Report	Feedback is extensively explained Feedback display is complex Feedback is displayed from different perspectives Self-scores are compared to other-scores Mean and range is reported for items and for dimensions Scores are not compared to norms	No display from different perspectives (other-ratings are combined in one group) Self scores are compared to aggregated other scores In the paper-and-pencil version, feedback is only reported for competencies, the electronic version will report on both competencies and items Mean and range for each competency is displayed (in the electronic version also for items) Scores are not compared to norms
Background	Developed on basis of theory	Theory and expert opinions
Reliability	Internal consistency has not been studied Interrater reliability has not been studied Test-retest reliability has not been studied	Internal consistency is low Interrater reliability has not been studied (check) Test-retest reliability has not been studied
Validity	Validity has not been studied	Validity has not been studied
Additional information		
Experience	2 years, used in many organizations	2 years, used in many organizations
Purpose	Development	Input for training

Discussion

In this section each element of the standards checklist is discussed and similarities and differences between the four instruments are described. On the basis of the results presented, organizations and researchers can select the instrument that best meets their needs. For example, an organization that wants to use 360-degree assessment to stimulate communication about managers' job-performance will look for another instrument than a researcher that wants to measure training effects.

Competencies and items

- All instruments except one allow competency selection. Competency selection is necessary when the instrument is being used for training evaluation since only relevant competencies, i.e. those competencies that are developed in a specific training, should be included;
- Every competency usually is covered by several items;
- Though according to the literature it is important that items focus on easily observed behavior, most of the examined instruments contain one or more items that are not formulated into that direction. If items do not focus on observable behavior, raters need to interpret their meaning which probably will result in decreasing reliability. If the instrument is especially meant to be used for development purposes, this may not be a big problem. In contrast, if ratings are being used for training evaluation, reliability is of great importance;
- Some instruments focus on behavior that can be developed by the ratee (instrument 1 and 2) while others contain behavior that can not directly be influenced by the ratee. For example, personality traits are difficult to change and thus not easy to be developed by ratees. Likewise, items formulated as results may be strongly dependent on other factors than the ratee's performance;
- Whether the examined instruments do not cause bias is difficult to say on the basis of the available information. However, none of them does explicitly contain bias characteristics;

- Most instruments are characterized by clear item wording, though some of them contain items that focus on more than one thing at the time, e.g. 'this person communicates about important decisions *and* asks for input'. These kinds of items can not be answered on one response scale and therefore should be reformulated.

Adaptations

- When the instrument is being used for training evaluation, it should be possible to select the competencies and items on which a specific training focuses. Most instruments do not allow further adaptation than selecting relevant competencies (instruments 1,2,4) or leaving out items (instrument 3). However, in most cases the database of competencies and items is large enough to enable the selection of relevant items.

Response scale

- Two instruments consist of 5-point response scales, one has a 6-point scale, and one a 9-point-scale. If the aim is to measure changes as a result of training this may ask for a, more specific, 9-point scale. However, 5-point scales also seem to provide enough opportunity to measure changes. In addition, most people are familiar with giving ratings on a 5-point scale. A 6-point scale could have the advantage that 'neutral' ratings are not possible since it does not have a central position, such as 5- and 9-point scales. It is difficult to say which response scale is the best option;
- Response scales are formulated as agreement, (for example: 'I agree/I do not agree with this statement'), frequency ('my supervisor often/never exhibits this behavior'), or applicability ('this statement is applicable/not applicable'). Instrument 1 is the only one that has a non-numbered scale where raters must choose between opposite statements;
- All instruments but one (1) have a non-response option;
- Only one instrument (4) contains room for recommendations;
- There is only one instrument (3) where raters also indicate the desired score on each item.

Feedback report

- All instruments, except the fourth instrument, display feedback from different perspectives. Instrument 4 only displays self-scores and aggregated other-scores. For developmental purposes, the feedback should be displayed from different perspectives so that ratees receive relevant information. If the purpose of the instrument usage is training evaluation, as in this case, feedback should also be provided for each different rater source since one of the aims is determining the extent of agreement between and within sources;
- All instruments compare self-scores to other-scores. Again, if the aim is development this is extremely important. This often is the main information of interest to ratees and a starting point for their development;
- All instruments display feedback for both competencies and items. This is important since a score on a competency is not specific enough. For example, if scores for one competency are low, the ratee may want to know which behavior is responsible for it;
- All instruments report the means and range, except instrument 3, which reports results as a percentage. The availability of means and ranges (and the standard deviation) is important if changes between two ratings, (before training and afterwards) are being measured;
- Instrument 1 is the only instrument that compares scores to norms;
- Instrument 1 is the only instrument that contains a strength-weakness display;
- Instrument 1 and 2 contain follow-up directions.

Instrument development

- The basis for the instrument consists usually some sort of experience and/or theory, however no research. Though instruments are used in many organizations, not much is known about their reliability and validity.

Reliability

- Some research concerning their reliability has been done for three of the instruments. Internal consistency appears to be satisfying (.6 or higher) for instrument 1 and 2, but low for instrument 4. Whenever interrater

reliability has been studied it appears to be low. Test-retest reliability has only been studied (in a limited way) for instrument 1 and appeared to be low.

Validity

- Validity has almost never been studied. Instrument 1 is the only instrument for which content validity was studied; it was found to be satisfactory.

Conclusion and limitations

Four Dutch instruments have been evaluated on the basis of a checklist containing important standards. The outcomes of this study give insight into the qualities of a number of 360-degree instruments in terms of their accordance with generally accepted personnel evaluation standards.

The examined instruments do not meet the standard that items should focus on easily observable behavior. Most of the examined instruments contain one or more items that are not formulated that way. Furthermore, only two of the examined instruments focus strictly on behavior that can be developed by the ratee. Regarding the feedback display, most instruments display feedback from different perspectives and all compare self-scores to other-scores and moreover, display feedback for competencies as well as items. The instruments are mainly used for personnel development, however vendors are more and more considering, or experimenting with, the use of 360-degree instruments for other purposes.

Depending on the specific purpose an organization has for a 360-degree instrument, it might prefer one or the other. On the basis of this study it is difficult to indicate one appropriate instrument to be used in training evaluation. Instrument 1 would probably be the most appropriate instrument for a research purpose. This is the only instrument that combines the possibility of competency selection, behavioral focus, focus on items that can be controlled and developed by the ratee, and feedback display of the mean and range for competencies as well as items, and from different perspectives. However, some items may need to be reformulated since they measure more than one behavior.

This study was limited by the fact that only a sample of Dutch instruments was involved. In the future, more instruments should be examined. This study can be used as a starting point for examining other 360-degree instruments.

Furthermore, the examination was based solely on document analysis and interviews with the developers and vendors of the instruments. Thus, the instruments have not been studied in practice and no raters and ratees have been interviewed.

Nevertheless, this paper provides a sound basis for developing better assessment instruments that can be used in the context of human resource development to accomplish various goals, one of them being the measurement of training transfer.

References

Boers, C. (1999) *Voor een spiegel: uitbreiding en aanpassing van een 360-graden feedback instrument.* [Looking in a mirror: expansion and adaptation of a 360-degree feedback instrument]. Thesis, Tilburg

Cardy, R. and G. Dobbins (1994) *Performance appraisal: alternative perspectives*. Cincinnati: South-Western Publishing Co

Cascio, W. (1995) *Managing human resources*. New York: McGraw-hill

Church, A. (1999) *Do higher performing managers actually receive better ratings? A validation of multirater assessment methodology*. Paper presented at the Annual Conference of the Academy of Human Resource Development

Dalessio, A. (1998) Using multisource feedback for employee development and personnel decisions. In: Smither, J.W. (ed.), *Performance Appraisal: state of the art in practice*, San Fransisco: Jossey-Bass

Giesen, R. van der (1997) *Top-performance door 360-graden feedback en coaching* [Top-performance by using 360-degree feedback and coaching]. Thesis, Leiden

Groeneveld, M. (1997) *Competentiegericht opleiden bij klanten van Boertien & Partners.* [Competency-based training of clients of Boertien & Partners]. Thesis, Enschede.

Hazucha, J., S. Hezlett and R. Schneider (1993) The impact of 360-degree feedback on management skills development. *Human Resource Development*, summer/fall, 325-351

The Joint Committee on standards for educational evaluation (1988) *The personnel evaluation standards: how to assess systems for evaluating educators.* Newbury Park: Sage

Lepsinger, R. and A. Lucia (1997) 360-degree feedback and performance appraisal. *Training*, September, 62-70

McLean, G., M. Sytsma and K. Kerwin-Ryberg (1995) *Using 360-degree feedback to evaluate management development: new data, new insights.* Paper presented at the Annual conference of The Academy of Human Resource Development

Morical, K. (1999) A product review: 360 assessments. *Training & Development,* April, 43-47

Murphy, K. and J. Cleveland (1995) *Understanding performance appraisal: social, organizational, and goal-based perspectives.* Thousand Oaks: Sage

Nijhof, W. and A. Jager (1995) *Reliability testing of multirater feedback.* Paper presented at the Annual Conference of the Academy of Human Resource Development

Rietveld, A. (1997) *De 360-graden methode als eye-opener: de ontwikkeling van een feedback instrument.* [360-degree assessment as an eye-opener: the development of a feedback instrument]. Thesis, Tilburg

Robinson, D. and J. Robinson (1989) Training for impact: how to link training to business needs and measure the results, San Fransisco: Jossey-Bass

Rosti, R. and F. Shipper (1998) A study of the impact of training in a management development program based on 360 feedback. *Journal of Managerial Psychology,* no1/2, 77-89

Tornow, W. and M. London (1998) *Maximizing the value of 360-degree feedback,* San Fransisco: Jossey-Bass

Velsor, E. van and J. Leslie (1991) *Feedback to managers: a review and comparison of sixteen multi-rater instruments.* Greensboro: Center for Creative Leadership

Velsor, E. van, J. Leslie and J. Fleenor (1997) *Choosing 360: a guide to evaluating multi-rater feedback instruments for management development.* Greensboro: Center for Creative Leadership

Woude, M. van der (1995) B*eoordelingsprocessen binnen 360-gradenfeedback instrumenten.* [Appraisal processes within 360-degree feedback instruments]. Thesis, Leiden.

Yukl, G. and R. Lepsinger (1995) How to get the most out of 360-degree feedback. *Training*, December, 45-50