

**Assessing Individual Managerial Skill Across Cultures: The Influence
of Language and Rating Source on 360-Degree Feedback**

Jean Brittain Leslie

Center for Creative Leadership

James Penny

CASTLE Worldwide, Inc.

Address: One Leadership Place
Center for Creative Leadership
Greensboro, NC, USA 27438-6300

Telephone: 336-288-7210

E-mail: [Lesliej @leaders.ccl.org](mailto:Lesliej@leaders.ccl.org)

This paper was presented at the 17th annual meeting of the Society for Industrial and Organizational Psychology, Toronto, Ontario, April 2002.

Abstract

The present study addresses the measurement equivalence of 98 items from a U.S.-based 360-degree managerial checklist, *SKILLSCOPE*®, which has been translated into Japanese for use in Japanese leadership development training. The basic question of interest was: How do the underlying managerial traits measured by rating source (boss, self, peer, direct report) relate to observed rating scale scores, and is that relationship the same across cultures? Logistic regression was used to model item responses by rater group and language of the survey (US English and Japanese). These results revealed forty items with differential item functioning (DIF) attributable to the language of the survey. Fifteen items exhibited DIF attributable to the rater group. Seven items exhibited DIF with respect to both language of the survey and rater group. All the items exhibited non-uniform DIF with magnitudes in the lower end of the small-to-medium range using Cohen's effect size for the Wald chi-square statistic. These results highlight the substantial influence of language and culture on assessing individual performance. Forty percent of the items were functioning differentially because of translation error, cultural differences in Japanese and US managerial style, or poorly constructed items. Fifteen percent of the items functioned differentially according to rater perspective. Rater perspective differences suggest certain aspects of the 360-degree feedback process may not be compatible with the Japanese collectivist culture. Further, these results provide modest evidence of an interaction between language and rating source suggesting that despite all efforts to produce a globally appropriate assessment, some items perform differently in the context of a particular language and a particular rater source.

The translation and use of U.S.-based 360-degree assessment for development instruments has increased rapidly over the past two decades yet little research has been conducted to substantiate their comparability (Leslie & Fleenor, 1998). Equivalence is particularly a concern when the instrument being used is based on a Western model of management and the administrative process employees a uniquely US model. As is the case with the use of the instrument being examined in this research, not only is the equivalence of the translated instrument a concern but also so is the equivalence of the feedback requested of various rating sources.

The use of 360-degree assessment for development instruments has remained popular among human resource professionals because they offer a unique multi-level perspective into managers' leadership styles. The target manager, the boss or superior, peers, direct reports, or in some case clients, typically complete 360-degree assessments. Inherent in the use of this methodology is the assumption that feedback generated from multiple perspectives is based on equivalent measures and the congruence or lack of congruence in ratings has some meaningful interpretation. In other words, does each type of rater use the same interpretation of item text? Murphy and Cleveland (1995), for example, indicate that subordinates have more opportunity to observe interpersonal behavior. Other generic managerial performance dimensions, such as communication and leadership, may be most likely observed by subordinates (Brutus, Fleenor, & London, 1998). Although there is little to no evidence to support cultural influences on subordinates' ability to observe and rate certain managerial performance dimensions, it is

reasonable to assume that cultural influences may account for equivalence or lack thereof.

This research seeks a better understanding of the application of a US-based 360-degree assessment in Japan through the examination of the influence of both language and ratings. Are these ratings equivalent in each rater source for each language, or are there some items that connote differently for particular rater sources and particular languages?

The fidelity of the translation

The process of test translation does not guarantee psychometric equivalence. Moreover, it is generally true that the translation of a survey from one language to another produces a survey that is different from the original (Beller, 1994; Foster, Olsen, Ford, & Sireci, 1997). That is, measurement equivalence does not necessarily exist between the two forms of the survey, regardless of the care taken in the translation of the items (Angoff & Cook, 1988; Brislin, 1980, 1986; Geisinger, 1994; Hambleton, 1993, 1994; Hulin, & Mayer, 1986; Olmedo, 1981; Prieto, 1992; Sireci, 1997; van der Vijver & Hambleton, 1996; van der Vijver & Tanzer, 1998). There are many reasons why measurement equivalence might not occur in cross-lingual assessments, including unintended effects such as non-standardized administration, poor item writing, poor translation, incomplete coverage of the constructs, and different levels of appropriateness of item content (van de Vijver & Leung, 1997). In the case of a 360 assessment, the

process itself may not be appropriate in non-US cultures (Leslie, Gyskiewicz & Dalton, 1998).

To have measurement equivalence between language versions of the same survey, the survey must measure the same construct, it must measure this construct in the same manner, and the resulting measures must lie on a common metric. Many researchers have used innovative statistical methods and research designs to assess the measurement equivalence of translated tests and surveys (Budgell, Raju, & Quartetti, 1995; Ellis, 1989, 1991; Ellis & Kimmel, 1992; Ellis, Minsel, & Becker, 1989; Hulin, Drasgow, & Komocar, 1982; Osberg, Scott, & Raju, 1985; Sireci, Fitzgerald, & Xing, 1998). These studies and more have contributed deeply to the development of a taxonomy of measurement bias and measurement equivalence presented in van der Vijer & Poortinga (1997) and van der Vijer & Tanzer (1998).

When measurement inequivalence is present between the original [US English] form and the translated survey, we suggest two possible reasons. These reasons are: (1) the test items may not be equally culturally relevant for different groups, and (2) the meaning of the test items may have been changed in the process of the translation (Ellis, 1991).

Measurement equivalence across rating sources

Several studies have examined the degree to which 360-degree assessments support the assumption of measurement equivalence across rating sources. Maurer, Collins, & Raju (1998) used confirmatory factor analysis and item response theory (IRT) with data from an appraisal measurement to suggest that the ratings of peers and

subordinates are directly comparable. Collins, Raju, & Edwards (2000) used the DFIT framework (Raju, van der Linden, & Fler, 1995) to identify items in a satisfaction scale that exhibited a degree of between-group differences. Facticeau and Craig (2001) used both CFA and DFIT to examine the performance of items across rater sources. Other similar studies have also identified some differences between rater groups (Collins, Raju, & Edwards, 1997; Drasgow & Kanfer, 1985; Laffitte, Raju, Scott, & Fasolo, 1998; Mount, Judge, Scullen, Sysma, & Hexlett, 1997; Riordan & Vandenberg, 1994).

When an item functions differently for groups of raters, it suggests a degree of measurement inequivalence indicative of systematic bias. Several researchers have offered causes for this systematic bias. For instance, Reilly & Warech (1993) suggested that response distortion represents one cause of the measurement inequivalence between rater sources. Greguras & Robie (1998) and Smith, Kendall, and Hulin (1969) suggested that rating differences are attributable to different frames of reference in use by different rater groups. Flanagan (1997) suggested impression management as a potential source of rater differences. Yukl & van Fleet (1992) suggested that managers often change their behavior to fit particular situations, and, following this line of argument, managers who behave differently toward different groups of co-workers may receive different ratings from members of those groups.

To the translation fidelity hypotheses identified in the section above, we add another hypothesis, which concerns the question of whether it is culturally appropriate to use a US-based 360-degree process in Japan. These hypotheses will be tested using a

combination of quantitative (empirical) and qualitative methods (the judgment of language and cultural experts).

Assessing measurement equivalence: Differential item functioning

Differential item functioning is a statistical assessment arising from the study of test bias and fairness that has received substantial attention in the measurement community for the past 2 decades. DIF manifests in two types: uniform and nonuniform DIF (Hambleton, Swaminathan, & Rogers, 1990). Within nonuniform DIF, there are two sub-types, crossing and non-crossing (Penny & Johnson, 1999). With uniform DIF, the degree and direction of the influence of the differential functioning does not vary over the levels of the attribute measured by the survey. An item exhibiting DIF by consistently augmenting the ratings from the peer group over the ratings from direct reports, even though group members are otherwise giving equal ratings, is an item exhibiting uniform DIF. From the framework of item response theory, parallel item response functions represent uniform DIF. (Hambleton, Swaminathan, & Rogers, 1991). With crossing, non-uniform DIF, the direction of the influence on the ratings changes. For instance, this type of DIF may depress the ratings from peers when they are rating a manager who is performing below average for the given population, whereas this same item may increase the ratings from peers when they are rating a manager who is performing above the population average. From the framework of item response theory, this type of DIF produces item response functions that cross due to differing degrees of item discrimination. (Hambleton, Swaminathan, & Rogers, 1991). When the direction of the influence on the ratings is constant for all levels of standing on the attribute that the

survey measures, but the degree of that influence varies, the type of DIF is non-crossing, non-uniform. From the framework of item response theory, non-crossing non-parallel item response functions produced by differing lower asymptotes describe this type of DIF (Penny & Johnson, 1999).

DIF in management checklists

Management checklists are surveys that ask the respondent if a management strength is present or not. These checklists are often used in management training programs because they are easy to administer and can provide a great deal of information in a short period of time. These surveys provide dichotomous scores which can be either qualitative (quality present as 1, quality not present as 0) or ordinal (strength present as 1, strength absent as 0). As such, groups can be compared on their responses using 2 x 2 contingency table analysis with group membership in the columns and the group counts for 0 and 1 responses in the rows. For the purpose of this paper, the 360-degree management assessment tool used in the present study is classified as a management strength checklist in that raters are asked to give dichotomous responses regarding whether an item is a strength or not.

Research questions

This research sought to establish the degree to which differential item functioning attributable to language or rater source may influence the ratings of a 360-survey. Might an item in one language function differentially when translated to another language? Are there items that one rater source may interpret differently from the others? Is there

evidence to support the existence of an interaction between language and rater source? That is, do some items perform differently only in the context of a particular language and a particular rater group? A review of the literature did not reveal evidence of such an interaction. However, language is an important facet of culture, and, as such, may serve to produce detectable differences between some rater groups. Finally, what may be the subsequent implications for the interpretation of the 360-feedback?

Method

Instrument

SKILLSCOPE is a 98-item checklist of managerial skills grouped into 15 skill clusters. The conceptual basis for SKILLSCOPE is Mintzberg's (1973) research indicating that managerial work involves informational skills, interpersonal skills, and decisional skills. In the development of SKILLSCOPE, personal resources and motivation to make effective use of these resources ("use of self") were added to Mintzberg's skills (Kaplan & Ohlott, 1988). Items were then written to capture performance in these skill areas. Respondents indicate whether each item is a "strength" or a "development need." If it is neither a strength or development need, or if it does not apply to the person being rated, respondents are instructed to leave the item blank. From the five skill areas mentioned above, 15 clusters of items were developed. Informational Skills includes two clusters of items, Getting Information and Making Sense of It and Conveying Information. Decisional Skills is made up of four clusters: Taking Action, Making Decisions, Following Through; Risk-taking and Innovation; Administrative/Organizational Ability, and Managing Conflict. The Interpersonal Skills

category contains four clusters: Relationships; Selecting, Developing and Accepting People; Influencing, Leadership and Power; and Openness to Influence; Flexibility. The two clusters in Personal Resources are: Energy, Drive, and Ambition; and Knowledge of the Job and Business. The clusters in the Use of Self category are Time Management, Coping with Pressure and Adversity; Integrity, and Self-management, Self-insight, Self-development.

Insert Table 1 about here.

KR-20s (for dichotomous scales) are reported for each cluster, based on the responses of 4,953 observers and 2,364 participants. Because of the nature of the response scale, only responses indicating “Strengths” were included in this analysis. Values for the 15 clusters ranged from .66 (Communicating) to .83 (Taking Action).

Sample

The 1999 SKILLSCOPE database from the Center for Creative Leadership (CCL) was used for this research. A random sample of 8,000 US managers and their raters, peers, boss, and direct reports were included. The types of managers in the total US sample can be characterized as first line managers and middle-level managers. These data represented a broad range of organizations such as governmental agencies, manufacturing companies, and educational institutions.

The Japanese sample consisted of all available *SKILLSCOPE* data collected through The Japanese Management Association. The Japanese sample contained 5,000

cases (managers and their raters boss, direct reports, peers). Table 2 presents the breakdown of rater type for each language.

Insert Table 2 about here.

Computation of DIF: Logistic Regression Procedures

This research uses logistic regression to detect DIF. Swaminathan & Rogers (1990) first applied this methodology to dichotomous items; that is, items with two possible responses, usually 0 and 1. The logistic regression posits an item response function for each point on the scale according to

$$P(x = 1) = \frac{1}{1 + e^{(-a(\theta - b))}}$$

in which a is the discrimination parameter, b is the threshold parameter, and θ represents the standing of the manager on the measured trait. In this model, the threshold parameter, b , is the point on the θ -axis where the probability exceeds 50 percent that the response is in the next category. Researchers sometimes call the threshold parameter the “location parameter.”

Were DIF not existent,

$$z = \tau_0 + \tau_1\theta,$$

where θ represents the standing of the ratee on the attribute that the survey measures. The symbols τ_0 and τ_1 represent the intercept and the slope parameters of the logistic

regression model; these symbols also represent forms of the discrimination and location parameters of the item response theory 2-PL model (Lord, 1980). This model represents the situation where the rater source membership and the language of the survey do not influence the item response and where the only factor that does influence the response to the item is the standing of the ratee on the attribute the survey measures.

To expand the model to include components representing effects due to language and rater source membership,

$$z = \tau_0 + \tau_1\theta + \tau_2g + \tau_3l,$$

where g and l represent rater source membership and language, respectively, and τ_2 and τ_3 represent the logistic regression parameters for those two classifications. This model describes the instance where only uniform DIF exists.

This model has been expanded to accommodate the potential existence of nonuniform DIF by the addition of two more terms to produce

$$z = \tau_0 + \tau_1\theta + \tau_2g + \tau_3l + \tau_4g\theta + \tau_5l\theta,$$

where the two new terms indicate an interaction, respectively, between (a) rater source membership and standing on the attribute measured by the survey, and (b) survey language and standing. The symbols τ_4 and τ_5 represent the logistic regression parameters for these two interaction terms, respectively.

To complete the model for this research, two additional terms were included. One term is to indicate the possible interaction between rater source membership and survey language; the other term is to indicate the possible three-way interaction of rater source

membership, survey language, and standing. The types of DIF represented by these two terms are uniform and nonuniform, respectively.

$$z = \tau_0 + \tau_1\theta + \tau_2g + \tau_3l + \tau_4g\theta + \tau_5l\theta + \tau_6gl + \tau_7gl\theta,$$

The symbols τ_6 and τ_7 represent the logistic regression parameters for these two additional interaction terms, respectively.

The SAS[®] System produced Wald Chi Square statistics to test the null hypotheses that the parameter estimates of τ_1 through τ_7 were statistically significantly different from 0. It was our anticipation that τ_1 and τ_2 would consistently achieve statistical significance. It was also our anticipation that the parameter estimates of τ_3 through τ_7 would not routinely achieve statistical significance.

Unidimensionality

One of the basic assumptions underlying the logistical model is that the items in a scale form a unidimensional set. Reckase (1979) found that a scale is sufficiently unidimensional if the percent variance attributable to the first un-rotated principal component exceeds 20 percent. Though this rule-of-thumb applies directly only to dichotomous data, it has seen successful used with polytomous data (Oshima, Raju, & Flowers, 1997; Raju, 1999).

Statistical tests for DIF indexes

To compensate for the accumulated Type I error rate that could naturally occur in this research and to avoid the power analysis procedures (Hsieh, 1989; Whittemore, 1981) of logistic regression, we used effect size instead of statistical significance for

flagging anomalous items. This method (Penny & Johnson 1999) converts the Wald chi-square statistic to an effect size, w , described in Cohen (1988, ch. 7). The formula that relates the effect size to the sample size is

$$X^2 = nw^2$$

where X^2 is the chi-square statistic, n is the sample size, and w is the effect size. Cohen (1988, ch. 7) used the arbitrary values of .1, .3, and .5 to indicate small, medium, and large effects, respectively, and, although these values are indeed arbitrary, Penny & Johnson (1999) found those values to connote well when applied to the Mantel-Haenszel chi-square statistic. These three values were used to define four effect ranges to categorize the DIF. These ranges were nil-to-small, small-to-medium, medium-to-large, and large-to-extreme.

After classifying the items by the type of DIF, either uniform or nonuniform, they were classified according to the apparent source of the differential functioning. An item differentially influenced by rating source should produce parameter estimates that suggest a contrast of at least one rater source to all the others. Similarly, the parameter estimates for the terms indicating language should indicate the contrast of one language to one or more of the others when the item in that language functions to produce ratings differentially higher or lower than the same item in the other languages.

Results and Discussion

Unidimensionality

Together, all 98-items meet this condition for each rater source. The mean variance attributable to the first component for the self-raters was 50 percent. For the boss, peer, and direct report raters, the mean percent variances attributed to the first components were 58, 59, and 59 percent, respectively.

The influence of standing on the item responses

The main effect of standing on the attribute measured by the survey had by far the greatest influence on the item ratings of all the terms in the logistic model. The average effect was .53 with a standard deviation of .10. The range was from .14 to .70, indicating some skew toward higher values, and suggesting further that the logistic ogive was a suitable model for use with these data.

DIF attributable to language

Forty items exhibited DIF attributable to the language of the survey. Table 3 presents these items. All the items exhibiting DIF attributable to the language of the survey did so with magnitudes in the lower end of the Small-to-Medium range. The modal effect was .11 with a minimum of .10, and a maximum of .29. None of the items exhibited nonuniform DIF.

Insert Table 3 about here.

The items were submitted to a team of linguistic and cultural experts whose native language is Japanese but who are also fluent in English. The team of subject

matter experts (SMEs) consisted of psychologists with The Japanese Management Association (JMA), located in Tokyo, Japan. The SMEs were asked to suggest possible translation error and/or cultural explanations that could account for the DIF. It was our hypothesis that when DIF was present it may be due to: (1) changes in the meaning of the test items during the translation process, and/or (2) the items not being equally culturally relevant for different groups. The purpose of this analysis was to assess the degree to which DIF results could be explained.

It was the SMEs conclusion that translation error may have accounted for approximately 23% of the items showing DIF. Differences in cultural expressions or concepts that may not be equally relevant in the US and Japanese cultures may account for 50% of the discrepancy in the two language versions. Not surprisingly, it seems that expert judgment may not be sufficient enough to detect all of the subtleties of language or culture that can contribute to measurement nonequivalence of translated tests. About 27% of the source of the DIF items was a mystery to the expert reviewers.

One US English item posited as an example of probable translation error that could be corrected, “Can translate strategy into action over the long haul,” has been back-translated from Japanese to English by the expert reviewers to “can translate strategy into action to the long term.” The objective of the item, which is to convert strategy into long-term action, has been lost in the Japanese translation. This item, however, tends to favor Japanese managers.

In general, the results of the translation fidelity highlights the careful considerations needed to successfully translate instruments. Many items on the

SKILLSCOPE® check-list contain business jargon which makes translations and adaptations difficult. One such item that exhibited DIF in English, “Troubleshooter; enjoys solving problems” was back-translated from Japanese to read (in English), “Plays a role to solve the problem; enjoys solving the problems.” This is not an exact translation, but the apparent improvement seems to help in Japanese.

All fifteen skill clusters contained at least one item with DIF attributed to language. One whole skill cluster, Managing Conflict; Negotiation, showed DIF in favor of US managers. Similarly, all but one item in the Time Management skill cluster favored US managers. It was the SMEs conclusion that these skill differences are reflective of cultural differences. The Japanese value for harmony is well documented in the literature. In negotiations, Japanese “face” must not be lost and politeness must be maintained at all times. In Japanese business interactions, it is not only the manager’s face but that of his or her company as well. Western managers, on the other hand, tend to view conflict as something that can be managed individually through direct confrontation. In the US, negotiation and conflict are commodities to be gained and won, regardless of potential loss of group harmony. This particular skill is considered by Mintzberg (1994) to be an action role required of effective managers. That is, they are required to make decisions, resolve crises, seize opportunities, negotiate contracts, and manage conflict.

Time Management, also thought to be culturally relative, within US management refers to the ability to manage conflicting tensions inherent in managerial work while maintaining a high level of productivity. In many Western cultures, time is considered

monochronic, that is, time is experienced and used in a linear way (Hall & Hall, 1990). For many Americans time is something tangible that can be “spent,” “wasted,” “lost,” or “prioritized” but rarely interrupted. Within the workplace, US managers often treat time as the primary mechanism by which organizations are organized. The Japanese, on the other hand, are typically polychronic in their characterization of time. Polychronic systems are characterized by the simultaneous occurrence of many things and by a great involvement with people (Hall & Hall, 1990). Japanese managers, for example, might be late for a scheduled meeting to finish an interaction.

DIF attributable to rater source

Fifteen items exhibited DIF attributable the rater source. In every instance, the magnitude of the effect was in the lower end of the range. Modal effect was .11 and .12. The minimum was .11, and the maximum was .22. None of the items exhibited nonuniform DIF. Table 4 presents these items.

Insert Table 4 about here.

Items with DIF attributable to rater source were submitted to a SME team. Unfortunately, the task of constructing hypotheses about the causes of rater source DIF was not straightforward. The identification of patterns was difficult. The presence of rater source DIF, for example, was not consistent within any grouping of items. Further, very few items displayed DIF across all the comparisons examined (e.g., boss, peers, direct reports).

Two items which might have been expected to show DIF did, where as others did not. Higher ratings were found, for example, among direct reports on the item, “Has good relationships with subordinates” which one might expect to find given halo in direct report ratings. Bosses likewise, were more favorable than direct reports and peers when rating managers astute sense of “politics.” Of all the rater sources, direct reports, seem to more often define skills differently than the other groups (see Figure 1).

Figure 1
Managerial Skills Observed Differently By Rater Source

Managerial Skill	Direct Reports	Peers	Supervisors
Good public speaker; skilled at performing, being on stage.	✓ ↓		
Makes his or her point effectively to a resistant audience.	✓ ↓		
A team builder; brings people together successfully around tasks.	✓ ↑	✓ ↑	
Makes good use of people; doesn't exploit.	✓ ↑		
Has good relationships with subordinates.	✓ ↑		
Considers personalities when dealing with people.	✓ ↑	✓ ↑	
Possesses extensive network of contacts necessary to do the job.	✓ ↓	✓ ↓	
Astute sense of “politics.”	✓ ↓	✓ ↓	
Takes ideas different from own seriously.	✓ ↑	✓ ↑	

and from time to time changes mind.			
At home with graphs, charts, statistics, budgets.	✓↓	✓↓	
Understands cash flows, financial reports, corporate annual reports.	✓↓	✓↓	
Deals with interruptions appropriately; knows when to admit interruptions and when to screen them out.	✓↓	✓↓	
Avoids spreading self too thin.	✓↓	✓↓	✓↓
Strikes a reasonable balance between his/her work life and private life.		✓↓	✓↓
Takes good care of self; uses constructive outlets for tension and frustration.		✓↓	✓↓

↓ indicates lower ratings; ↑ indicates higher ratings

Direct report ratings are considered an important part of the 360-degree feedback process. There is however, recognition that upward feedback has potentially negative aspects. According to Brutus, Fleenor, and London (1998:16-17), “By shifting the traditional roles of the rater and the ratee, a new set of dynamics takes place. From the supervisor perspective, receiving upward ratings is potentially threatening... This represents a rare instance in which subordinates can directly and anonymously affect their supervisor.” They go on to add, “A potentially negative aspect of an employee rating a supervisor is the possibility of retaliation. Supervisors who are aware that subordinates have given them negative ratings may punish them by assigning undesirable tasks, withholding salary increases, or generally making the employees’ job more difficult. The fear of retaliation, real or imagined, may work to positively bias the ratings.” Clearly, asking for upward feedback is risky. There is some evidence that suggests that it is considered even more of a status violation [of the hierarchy] in non-

U.S. cultures. In certain Latin American cultures, for example, U.S. managers have reported difficulty in getting direct reports to give them negative feedback. In these cultures, conflict may be avoided for the sake of social harmony and asking direct reports for feedback may be considered shocking or even offensive (Leslie, Gyskiewicz & Dalton, 1998).

A cultures power distance may provide one explanation why feedback may not equally accepted across all cultures. Gerte Hofstede's research (1980) showed national cultures differ to the extent to which a society accepts the unequal distribution of power (power distance). People from larger power distance cultures tend to share values and beliefs that (1) dependence on higher-ups is accepted and desired, (2) directive and persuasive superiors are preferred, (3) authority is not to be questioned, (4) managers are entitled to privileges, and (5) the use of coercive and referent power is accepted (Hofstede, 1980). In cultures characterized by a high power distance (e.g., Philippines, Mexico, Venezuela, India, Yugoslavia, Singapore), bosses may feel threatened or fearful of direct reports data even though it is kept confidential. Japanese society has a hierarchical structure in which most relationships are unequal.

DIF attributable to the language by rater source interaction

Seven items in this study exhibited uniform DIF with respect to the interaction of rater source and language. The cut-values of effect that forms the ranges were .1, .3 and .5. All effects due to rater source were small-to-medium. Most notably the item, "Avoids spreading self too thin" had lower direct report, peer, and boss ratings than self-reported ratings. It also favored US managers. It is not surprising that Japanese

managers nor their co-workers see this as a developmental need. The Japanese enjoy one of the highest standards of living in the world but they also spend on average more hours at work per year than their Western counterparts.

By most accounts, Japanese is one of the world's hardest languages to learn and understand. Japanese is often described as a vague or ambiguous language sometimes deliberately to absolve blame and demonstrate politeness. Within the language itself, there are many rules on how to address parties particularly within a superior and subordinate relationship. There are, for example, at least four ways of saying thank you in Japanese. The one chosen reflects the relative status of the individuals. Perhaps these results reflect this subtlety in both Japan and The US.

Concluding Remarks

Does language and rating source influence 360-degree feedback? These results suggest that they do. Nonetheless, we cannot overlook the fact that that 50% of the items examined in this research exhibited measurement equivalence. Most notably, Mintzberg's (1973) informational skill (Getting Information, Making Sense of It; Problem Identification), decisional skills (Taking Action, Making Decisions & Administrative/Organizational Ability) interpersonal skills (Relationships; Selecting/Developing People; Influencing, Leadership, Power; & Openness to Influences; Flexibility), personal skills (Energy, Drive, Ambition) and, effective use of self (Coping with Pressure, Adversity, Integrity & Self-management, Self-insight, Self-development) are multicultural relevant to managerial work.

It is possible that the degree of DIF revealed by the research was due in part to the methodology. Effect size was used to flag the items, not statistical significance. Had statistical significance been used to flag anomalous items, many more would have been identified. However, those additional items exhibited degrees of DIF well under the cut-value for a small effect, indicating that using statistical significance, at least with this sample, would have created a far too aggressive DIF identification procedure.

In addition, most of the DIF identified in this study was small, and the cumulative effect, at least for group feedback reports, was small. However, it is likely that the effect of the DIF identified herein could be substantial on some individual feedback reports. At present, there is no methodology to apportion survey ratings into various sources, say, standing, rater type, language, gender, and other attributes. Such methodology would likely be of value in 360-reporting, although additional research will be necessary to determine the usefulness of apportioned 360-feedback ratings. Moreover, additional research could address the influence of multi-faceted items that may produce anomalous functioning in particular languages.

Finally, most researchers tend to see DIF as an unwanted quantity in most assessments. However, this research identified some items exhibiting DIF attributable to rater type that may be a naturally occurring phenomenon. It seems quite reasonable to anticipate that some raters will interpret some items differently simply because of the environment of the rater, and it is without doubt that bosses often see a more complex and ambiguous world than do direct reports. However, it may be that such an environmental influence varies to some degree with the level of the manager, and only

additional study that controls both the level and position of the rater is likely separate those influences.

References

Angoff, W. H., & Cook, L. L. (1988). Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test (Report No. 88-2). New York: College Entrance Examination Board.

Beller, M. (1994). Normative testing and bilingual populations. Journal of Multilingual and Multicultural Development, 9, 399-409.

Brislin, R. (1980). Translation and content analysis of oral and written material. In H. C. Triandis & J. W. Berry (Eds.), Handbook of cross-cultural psychology (Vol. 2; pp. 389-444). Boston: Allyn and Bacon.

Brislin, R. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), Field methods in cross-cultural psychology (pp. 137-164). Beverly Hills: Sage.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. Applied Psychological Measurement, 19, 309-321.

Candell, G. L., & Hulin, C. L. (1987). Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. Journal of Cross-Cultural Psychology, 17, 417-440.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. Educational Measurement: Issues and Practices, 2, 31-44.

- Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Collins, W.C., Raju, N.S., & Edwards, J. (1997, April). Assessing DIF in a polytomously-scored satisfaction scale. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.
- Craig, B., Nambury, R.S., Zieleskiewicz, J.R., & Formen, A. Measurement Equivalence of the Benchmarks ratings Across Four Rating Sources. Paper presented at 1999 SIOP Conference, Atlanta Georgia.
- Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston.
- Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. Journal of Applied Psychology, 70, 662-680.
- Ellis, B. (1989). Differential item functioning: Implications for test translators. Journal of Applied Psychology, 74, 912-921.
- Ellis, B. (1991). Item response theory: A tool for assessing the equivalence of translated tests. International Test Bulletin, 32, 33-51.
- Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. Journal of Applied Psychology, 77, 177-184.
- Ellis B. B., Minsel, B., & Becker, P. (1989). Evaluation of attitude survey translations: An investigation using item response theory. International Journal of Psychology, 24, 661-684.

Foster, D., Olsen, J. B., Ford, J., & Sireci, S. G. (1997). Administering computerized certification exams in multiple languages: Lessons learned from the international marketplace. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychological Assessment, 6, 304-312.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. European Journal of Psychological Assessment, 9, 57-68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. European Journal of Psychological Assessment, 10, 229-244.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage Publications.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to the analysis of attitude scale translations. Journal of Applied Psychology, 67, 818-825.

Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different sources comparable? Journal of Applied Psychology, 86, 215-227.

Flanagan, W. J. (1997). Measurement equivalence between high and average impression management groups: An IRT analysis of personality factors. Unpublished doctoral dissertation, Georgia Institute of Technology, Atlanta, GA.

Fiedler, F.E. (1978) The contingency model and the dynamics of the leadership process, In Advances in experimental social psychology, ed. L. Berowitz. New York: Academic Press.

Fielder, F.E. & Chemers, M.M. (1982). Improving leadership effectiveness: The leader match concept. 2nd Ed. New York: Wiley.

Flanagan, W.J., & Raju, N.S. (1997, April). Measurement equivalence between high and average impression management groups: An IRT analysis of performance dimensions. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Flower, C.P., Oshima, T.C., & Raju, N.S. (in press). A monte carlo assessment of DFIT with polytomously scored unidimensional tests. Applied Psychological Measurement.

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. Journal of Applied Psychology, 83, 960-968.

Hsieh, F. Y. (1989). Sample size tables for logistic regression. Statistics in Medicine, 8, 795-802.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Erlbaum.

Hulin, C. L., & Mayer, L. M. (1986). Psychometric equivalence of a translation of the JDI into Hebrew. Journal of Applied Psychology, 71, 83-94.

Laffitte, L.J., Raju, N.S., Scott, J.C., & Fasolo, P.M. (1998, April). Examination of the measurement equivalence of a 360° feedback assessment with confirmatory factor analysis and item response theory. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology in Dallas, TX.

Leslie, Jean B. & Fleenor, John (1998) Feedback to managers: A review and comparison of multi-rater feedback instruments. Greensboro, NC: Center For Creative Leadership.

Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 140, 44-53.

Lindsey, E., Homes, V., & McCall, M.W., Jr. (1987). Key events in executives' lives (Tech. Rep. No. 32). Greensboro, NC: Center for Creative Leadership.

Loehlin, J. C. (1992). Latent variable models: An introduction to factor, path and structural analysis. (2nd Ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Maurer, T.J., Raju, N.S., & Collins, W.C. (1998). Peer and subordinate performance appraisal measurement equivalence. Journal of Applied Psychology, 83, 693-702.

McCall, M.W., Jr., & Lombardo, M.M. (1983, February). What makes a top

executive? Psychology Today, 26-31.

McCall, M.W., Jr., Lombardo, M.M., & Morrison, A.M. (1988). The lessons of experience: How successful executives develop on the job. Lexington, MA: Lexington Books.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. Journal of Educational Measurement, 30, 107-122.

Mintzberg, H. (1973). The nature of managerial work. New York: Harper & Row.

Mount, M. K., Judge, T. A., Scullen, S. E., Systma, M. R., & Hexlett, S. A. (1997). Trait, rater, and level effects in 360-degree performance ratings. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Olmedo, E. L. (1981). Testing linguistic minorities. American Psychologist, 36, 1078-1085.

Osberg, D. W., Scott, J. C., & Raju, N. S. (1985). An analysis of the use of item response theory to investigate the fidelity of test translations. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Oshima, T.C., Raju, N.S., & Flowers, C. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. Journal of Educational Measurement, 34, 253-272.

Penny, J., & Johnson, R. L. (1999). How group differences in matching criterion distribution and IRT item difficulty can influence the magnitude of the Mantel-Haenszel chi-square DIF index. The Journal of Experimental Education, *67*, 343-366.

Prieto, A. J. (1992). A method for the translation of instruments to other languages. Adult Education Quarterly, *43*, 1-14.

Raju, N. S. (1999). Psychometric Analysis of the Benchmarks[®] Survey. Chicago, IL: The Center for Research and Service, Illinois Institute of Technology.

Raju, N.S., van der Linden, W., & Fler, P. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. Applied Psychological Measurement, *19*, 353-368.

Reckase, M. D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and Implications. Journal of Educational Statistics, *4*, 207-230.

Reilly, R. R., & Warech, M. A. (1993). The validity and fairness of alternatives to cognitive tests. In L. C. Wing & B. R. Gifford (EDS.), Policy issues in employment testing. (pp. 131-224). Norwell, MA: Kluwer Academic.

Riordan, C., & Vandenberg, R. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? Journal of Management, *20*, 643-671.

Rogers, H. J. (1989). A logistic regression procedure for detecting item bias. Dissertation Abstracts International, *50*, 3928A. (University Microfilms No. 90-11,788).

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, 17, 105-116.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores, Psychometrika Monograph, 17.

Samejima, F. (1979). A new family of models for the multiple choice item. Office of Naval Research Report 79-4. Knoxville, TN: University of Tennessee.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential item functioning. In Holland, P. W. & Wainer, H. (Eds.), Differential Item Functioning (pp. 197-240). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. Educational Measurement: Issues and Practices, 16, 12-19.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998). Adapting credentialing examinations for international uses. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures, 27, 361-370.

Terman, L. M. (1916). The measurement of intelligence. Boston: Houghton-Mifflin.

Thissen, D., & Wainer, H. (1990). Confidence envelopes for item response theory. Journal of Educational Statistics, 15, 113-128.

van der Vijver, F., & Hambleton, R. K. (1994). Translating tests: Some practical guidelines. European Psychologist, 1, 89-99.

van der Vijver, F., & Poortinga, Y. H. (1977). Towards an integrated analysis of bias in cross-cultural assessment. European Journal of Psychological Assessment, 13, 29-37.

van der Vijver, F., & Tanzer, N. K. (1998). Bias and equivalence in cross-cultural assessment. European Review of Applied Psychology, 47, 263-279.

Warm, T.A. (1978). A primer of item response theory (Tech. Rep. No. 941078). Washington, DC: U. S. Coast Guard.

Whittemore, A. (1981). Sample size for logistic regression with small response probability. Journal of the American Statistical Association, 42, 415-427.

Table 1
 Name and description of SKILLSCOPE skill clusters

Skill Cluster	Items	Sample Item
1. Getting Information, Making Sense of It; Problem Identification	7	Seeks information energetically.
2. Communicating Information, Ideas	5	Adept at disseminating information to others.
3. Taking Action, Making Decisions, Following Through	5	Action-oriented; presses for immediate results.
4. Risk-Taking, Innovation	5	Consistently generates new ideas.
5. Administrative/Organizational Ability	9	Establishes and conveys a sense of purpose.
6. Managing Conflict; Negotiation	3	Effective at managing conflict.
7. Relationships	10	Builds warm, cooperative relationships.
8. Selecting, Developing, Accepting People	7	Attracts talented people.
9. Influencing, Leadership, Power	6	Delegates effectively.
10. Openness to Influence; Flexibility	9	Listens well.
11. Knowledge of Job, Business	6	A good general manager.
12. Energy, Drive, Ambition	4	High energy level.
13. Time Management	4	Avoids spreading self too thin.
14. Coping with Pressure, Adversity; Integrity	8	Doesn't hide mistakes.
15. Self-Management, Self-Insight, Self-Development	7	Compensates for own weaknesses.

Table 2
Breakdown of rater type by survey language

Language	Total	Direct Report	Peer	Self	Boss
US	N=?				
English					
Japanese					

Table 3

Items exhibiting DIF attributable to language

Item text and item number	Effect size	Impact of DIF
1. Seeks information energetically.	Small-to-Medium	Higher ratings in Japanese
7. Logical, data-based, rational.	Small-to-Medium	Higher ratings in Japanese
12. Strong communicator on paper; good writing skills.	Small-to-Medium	Higher ratings in Japanese
15. Troubleshooter; enjoys solving problems.	Small-to-Medium	Higher ratings in Japanese
16. Implements decisions, follows through, follows up well; an expediter.	Small-to-Medium	Lower ratings in Japanese
18. Has vision; often brings up ideas about potentials and possibilities for the future.	Small-to-Medium	Higher ratings in Japanese
19. Entrepreneurial; seizes new opportunities.	Small-to-Medium	Higher ratings in Japanese
20. Consistently generates new ideas.	Small-to-Medium	Higher ratings in Japanese
26. Resourceful; can marshal people, funds, space required for projects.	Small-to-Medium	Higher ratings in Japanese
27. Can organize and manage big, long-term projects; good shepherding skills.	Small-to-Medium	Higher ratings in Japanese
30. Can easily handle situations where there is no pat answer, no prescribed method for processing.	Small-to-Medium	Lower ratings in Japanese
31. Can translate strategy into action over the long haul.	Small-to-Medium	Higher ratings in Japanese
32. Effective at managing conflict.	Small-to-Medium	Lower ratings in Japanese
33. Confronts others skillfully.	Small-to-Medium	Lower ratings in Japanese
34. Negotiates adeptly with individuals and groups over roles and resources.	Small-to-Medium	Lower ratings in Japanese

Item text and item number	Effect size	Impact of DIF
37. Makes good use of people; doesn't exploit.	Small-to-Medium	Higher ratings in Japanese
38. Has good relationships with subordinates.	Small-to-Medium	Lower ratings in Japanese
40. Has good relationships with peers.	Small-to-Medium	Lower ratings in Japanese
48. Tolerant of the foibles, idiosyncrasies of others.	Small-to-Medium	Lower ratings in Japanese
49. Good coach, counselor, mentor; patient with people as they learn.	Small-to-Medium	Lower ratings in Japanese
55. Astute sense of "politics."	Small-to-Medium	Higher ratings in Japanese
57. Comfortable with the power of the managerial role.	Small-to-Medium	Higher ratings in Japanese
61. Listens well.	Small-to-Medium	Lower ratings in Japanese
63. Accepts criticism well; easy to give feedback on his/her performance.	Small-to-Medium	Lower ratings in Japanese
66. Flexible; good at varying his or her approach with the situation.	Small-to-Medium	Lower ratings in Japanese
69. Doesn't let power or status go to his/her head.	Small-to-Medium	Lower ratings in Japanese
71. A good general manager	Small-to-Medium	Higher ratings in Japanese
72. Effective in a job with a big scope.	Small-to-Medium	Higher ratings in Japanese
73. In a new assignment, picks up knowledge and expertise easily; a quick study.	Small-to-Medium	Higher ratings in Japanese
75. Understands cash flows, financial reports, corporate annual reports.	Small-to-Medium	Higher ratings in Japanese
76. Good initiative; continually reaches	Small-to-	Higher ratings in Japanese

Item text and item number	Effect size	Impact of DIF
for more responsibility.	Medium	
78. Ambitious; highly motivated to advance his/her career	Small-to-Medium	Higher ratings in Japanese
80. Sets priorities well; distinguishes clearly between important and unimportant tasks.	Small-to-Medium	Lower ratings in Japanese
82. Deals with interruptions appropriately; knows when to admit interruptions and when to screen them out.	Small-to-Medium	Lower ratings in Japanese
83. Avoids spreading self too thin.	Small-to-Medium	Lower ratings in Japanese
85. Can deal well with setbacks; resilient; bounces back from failure, defeat.	Small-to-Medium	Lower ratings in Japanese
86. Willing to admit ignorance.	Small-to-Medium	Lower ratings in Japanese
87. Optimistic; takes the attitude that most problems can be solved.	Small-to-Medium	Higher ratings in Japanese
91. Strikes a reasonable balance between his/her worklife and private life.	Small-to-Medium	Lower ratings in Japanese
97. Makes needed adjustments in own behavior.	Small-to-Medium	Lower ratings in Japanese

Note: Items 55, 75, 82, 83, and 91 also exhibited DIF attributable to rater group. In all instances, the type of DIF was uniform. The cut-values of effect that formed the ranges were .1, .3 and .5.

Table 4

Items exhibiting DIF attributable to rater group

Item text with item number	Impact of DIF
10. Good public speaker; skilled at performing, being on stage.	Lower Direct Report and Peer ratings vs. all others
11. Makes his or her point effectively to a resistant audience.	Lower Direct Report ratings vs. all others
24. A team builder; brings people together successfully around tasks.	Higher Direct Report and Peer ratings vs. all others
37. Makes good use of people; doesn't exploit.	Higher Direct Report ratings vs. all others
38. Has good relationships with subordinates.	Higher Direct Report ratings vs. all others
47. Considers personalities when dealing with people.	Higher Direct Report and Peer ratings vs. all others
54. Possesses extensive network of contacts necessary to do the job.	Lower Direct Report and Peer ratings vs. all others
55. Astute sense of "politics."	Lower Direct Report and Peer ratings vs. all others
62. Takes ideas different from own seriously, and from time to time changes mind.	Higher Direct Report and Peer ratings vs. all others
74. At home with graphs, charts, statistics, budgets.	Lower Direct Report and Peer ratings vs. all others
75. Understands cash flows, financial reports, corporate annual reports.	Lower Direct Report and Peer ratings vs. all others
82. Deals with interruptions appropriately; knows when to admit interruptions and when to screen them out.	Lower Direct Report and Peer ratings vs. all others
83. Avoids spreading self too thin.	Lower Direct Report, Peer and Boss ratings vs. self report
91. Strikes a reasonable balance between his/her work life and private life.	Lower Peer and Boss ratings vs. all others

Item text with item number	Impact of DIF
96. Takes good care of self; uses constructive outlets for tension and frustration.	Lower Peer and Boss ratings vs. all others

Note: Items 37, 38, 55, 75, 82, 83, and 91 also exhibited DIF attributable to rater group. In all instances, the type of DIF was uniform. The cut-values of effect that formed the ranges were .1, .3 and .5. All effects due to rater group were Small-to-Medium.