ARI Research Note 2007-03

Collaboration and Self Assessment: How to Combine 360 Assessments to increase Self-Understanding

Joseph Psotka and Peter J. Legree
U.S. Army Research Institute

Dawn M. Gray
George Mason University
Consortium Research Fellows Program

Research and Advanced Concepts Office
Paul A. Gade, Chief

March 2007

United States Army Research Institute
for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

A Directorate of the Department of the Army
Deputy Chief of Staff, G1

Authorized and approved for distribution:

Barbara A. Black Michelle Sams

BARBARA A. BLACK
Acting Technical Director

MICHELLE SAMS
Director

Technical review by

Trueman R. Tremble, Jr., U.S. Army Research Institute
Michelle Zbylut, U.S. Army Research Institute

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
March 2007	Interim	October 2006 – March 2007

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
Collaboration and Self Assessment: How to Combine 360 Assessments to Increase Self-Understanding	5b. PROGRAM ELEMENT NUMBER 611102
6. AUTHOR(S) Joseph Psotka and Peter J. Legree (U.S. Army Research Institute); Dawn M. Gray (George Mason University)	5c. PROJECT NUMBER B74F
	5d. TASK NUMBER
	5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Institute 2511 Jefferson Davis Highway Arlington, VA 22202	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences ATTN: DAPE-ARI-IT 2511 Jefferson Davis Highway Arlington, VA 22202-3926	10. MONITOR ACRONYM ARI
	11. MONITOR REPORT NUMBER Research Note 2007-03

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Subject Matter POCs: Joseph Psotka and Peter Legree.
Contracting Officer's Representative: Paul A. Gade

14. ABSTRACT *(Maximum 200 words)*:
Traditional performance appraisal measures are generally one sided and lack the ability to deliver accurate objective feedback. 360 assessments provide a collaborative tool for professional development utilizing superior, peer, subordinate, and self assessments to create a more balanced circle of feedback. Traditionally self ratings have been found to be less correlated than peer and superior ratings with performance measures. In general peer and superior ratings are more highly correlated than self and peer, or self and superior. By utilizing an objective measure of tacit leadership knowledge, the TKML, we compared self ratings to peer and superior ratings through a factor score. Given this new measure, correlations with superior and peer ratings were improved. However, subordinate ratings were found to be negligible. Self ratings were found to have a stronger correlation with leadership ability than peer or superior ratings and therefore were found to be the most reliable assessment of leadership ability. Additional regressions combining the 360 ratings yielded the strongest correlations with the TKML. Self appraisals should draw on our knowledge of ourselves and the perspectives that others provide for us.

15. SUBJECT TERMS
Performance appraisal, 360 degree assessment, tacit knowledge, TKML, self ratings, peer ratings, superior ratings, subordinate ratings, correlated, factor score

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited		Ellen Kinzer Technical Publications Specialist 703-602-8047

Research Note 2007-03

Collaboration and Self Assessment: How to Combine 360 Assessments to increase Self-Understanding

Joseph Psotka and Peter J. Legree
U. S. Army Research Institute

Dawn M. Gray
George Mason University
Consortium Research Fellows Program

Research and Advanced Concepts Office
Paul A. Gade, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

March 2007

Army Project Number
611102B74F

Personnel, Performance and Training

COLLABORATION AND SELF ASSESSMENT: HOW TO COMBINE 360 ASSESSMENTS TO INCREASE SELF-UNDERSTANDING

CONTENTS

LIST OF TABLES

LIST OF FIGURES

COLLABORATION AND SELF ASSESSMENT: HOW TO COMBINE 360 ASSESSMENTS TO INCREASE SELF-UNDERSTANDING

Introduction

Self awareness and understanding yourself have been recognized as an essential developmental skill for improving life success since the time of Socrates. A reasonable extension suggests that more accurate insight and self awareness of one's own strengths and weaknesses, skills, knowledge, and values, should be related to successful leadership potential and performance (Yammarino & Atwater, 1993). Self rating data compared to familiar others' ratings have often been suggested and studied in multi-source or 360 degree appraisal systems as indicators of self awareness (van Hooft, Flier, & Minne, 2006). Previous research (and the current research) demonstrated that the inter-rater agreement in 360 feedback ratings is typically low to moderate (Conway & Huffcutt, 1997; Harris & Schaubroeck, 1988), rarely providing the confident assessment of self awareness that robust appraisal systems require. Conway & Huffcutt (1997) found uncorrected Pearson correlations ranging from .14 between self ratings and subordinates' ratings to .34 for peer to superior ratings. Yet, without either robust correlations with others' ratings or independent objective evidence, self rating accuracy cannot be determined or evaluated. This research seeks to find coherence in this complicated multisource environment by examining the interrater agreements with the assistance of a validated objective instrument assessing Army leadership, the Tacit Knowledge of Military Leadership (TKML) instrument (Sternberg, Forsythe, Hedlund, Horvath, Wagner, Williams, Snook, & Grigorenko 2000).

Strengths of Multi-Source Rating Environments

In a 360 degree rating environment, raters assess different dimensions of effectiveness, from three or four different perspectives which usually include superiors, peers, subordinates, and customers (Tornow, 1993; Church & Bracken, 1997). According to this approach, differences in rater perspectives are viewed as potentially informative rather than simply error variance. In other words, multiple ratings can represent significant and meaningful sources of variation about perceptions of performance (Salam, Cox, & Sims, 1997).

Self Rating in Multi-Source Rating Environments

Self rating, when compared in meta-analyses (Harris & Schaubroeck, 1988) against others' ratings, correlate less with peers and superiors ($r = .22$) than superiors and peers ($r = .48$).agree with each other. Given these generally low correlations among self and others, it has generally been held not only that someone's views of oneself are generally different from others' views, but that self ratings are generally inaccurate assessments of the self. While the first view is directly given by the data, the second view merits additional investigation by examining converging evidence from other sources. Not everyone holds this belief. Borman (1997) summarizes several explanations as to why discrepancies exist between different rating sources, even though they may have no differential validity. Scullen (1997) offers another interesting

hypothesis that suggests that peer and subordinate ratings may have an artificially inflated appearance for validity because they are aggregated over several ratings, whereas self and superior ratings are usually made just by one person. Averaging several raters in the peer and subordinate groups will tend to make those ratings more reliable, and so correlate better with each other than with the superior or self ratings that are just based on a single rating.

Self Rating and Objective Measures

One of the most important variables that affects self rating is experience on the job. For example, Yammarino and Waldman (1993) showed that experience on the job has a significant effect on self rating of job-relevant skills and abilities. Among officers, rank is a proxy which stands in for job experience, since these are highly correlated, with Captains' subordinates (Lieutenants) usually having only one or two years of Army experience, whereas their superiors (Lieutenant Colonels) have ten to twenty years or more of job experience.

Personality and ability variables also seem to affect self ratings. Self esteem appears to be a directly related variable. Cognitive complexity and ability not only modify an individuals' self awareness but also one's ability to incorporate criticism and feedback to adapt to the prevailing culture (Yammarino & Atwater, 1993). Since there are many variables that may affect ratings in a 360 degree environment, a comprehensive objective measure must be multidimensional.

Sources of Error in Self and Other's Assessments

Both self assessment and others' assessment of an individual officer are subject to a wide range of error sources that can distort judgment (Ashford, 1989.) Self-presentation pressures, desire to please, social conformity, social desirability, and unfamiliarity with a position can all add to the noise of self rating. Similarly, others' unfamiliarity with a person, differences in job and task characteristics, differences in rank, attribution biases, halo effects, and personality, interests, and cognitive differences between rater and ratee can all create error in other's judgments. All too often, 360 degree research is carried out with ad hoc groups who do not know each other well. In this paper the research was conducted with intact chains of command in an Army setting.

In spite of all these sources of error and bias in self and others' assessments, self assessment has one critical advantage: a superior knowledge of past performance histories and capabilities. Given this decided advantage, it is unclear why self assessment enjoys such decidedly widespread, negative perceptions as biased, self serving, distorted, unreliable, and inaccurate in the literature (Yammarino & Atwater, 1993).

Self Assessment vs. Superior Assessment

Since this report deals with the leadership culture of US Army officers, the leadership, self rating, and assessment research of Bass and his colleagues is worth noting. Bass and Yammarino (1991) in a study of Naval officers determined that self ratings were often inflated, but less so in successful officers, who were more promotable. Similarly, McCauley & Lombardo (1990) showed that the more accurate an officer's self rating, the more likely it was the officer would be promoted. Of course if accuracy of self ratings is defined by the match

between self rating and superiors' rating, then the higher the correlation between self and superior the higher the probability of promotion has to be. If it is defined by match between self and subordinates and peers, it again does not make a lot of sense, since if they both agree that they are poor leaders, then promotion should be counterproductive. Instead these relationships only make sense if they are mediated by self awareness, and the match between superiors and self ratings shows that high self awareness, beyond the match between superiors' and self ratings, is what is being rewarded by promotion.

Self Rating and Superiors' ratings should agree in the Army system, since both officers and their supervisors are mandated to come together several times a year to review the superior's Officer Evaluation Report (OER) of each subordinate. The OER is a mutual assessment tool in which raters and ratees come to an agreement on critical issues that have been spelled out by detailed leadership task analyses refined over many years by the Army. This process suggests that self rating and superior ratings should agree better than self and subordinates or peers in the Army, since officers and their supervisors are obligated to review and discuss the critical issues frequently. Also, the greater experience self raters and their superiors have with this process should create greater awareness of leadership capabilities than among subordinates, who generally have only one or two years of experience with this process and with Army culture in general.

At the same time superiors generally do not observe officers dealing with tasks or interpersonal issues first hand. Superiors rely more on reports from others about these officers and the products these officers provide them in making their judgments. Peers and subordinates, on the other hand, deal with an officer directly and form their judgments from these personal interactions. It may then be the case that superiors are better at making overall judgments of an officer's leadership qualities but may not be as accurate as subordinates and peers in judging interpersonal or task leadership qualities.

Self Rating and Objective TKML Scores

The TKML was developed and validated by Hedlund et al (1998) in order to obtain a measure of how Army officers develop as leaders while they are on the job. Because ability that is demonstrated in the classroom does not necessarily transfer to the field (Dörner & Kreuzig, 1983), and intelligence determined by IQ scores does not capture the full range of skills needed for complex problem solving, Hedlund et al (1998) developed a tool for measuring the practical or tacit knowledge of military leaders. Good Army leaders become familiar with the full range of Army culture through many sources: classroom education, broad reading, especially formal documentation and doctrinal manuals, but mainly through direct operational experience and training. Tacit knowledge is not acquired through formal training, but rather gained through personal experience during operational assignments and used as a template for resolving similar problems when they are next encountered. Tacit knowledge covers a huge range of Army culture and aspects of leadership, so a multidimensional instrument is needed to cover it all. The TKML consists of a series of problems or scenarios, with each scenario accompanied by a set of possible responses to the situation that it describes. Respondents are asked to rate the quality or advisability of each response option using a nine-point Likert scale. The scenarios are representative of actual situations encountered by officers. Knowing how to deal with the everyday routine, not just the extraordinary, unique events is critically important for successful leadership.

The low fidelity simulation approach (Motowidlo & Tippins, 1993) offers a way of creating such an inventory by using short topics and vignettes followed by alternative solutions to the issues in the vignette. In order to obtain vignettes and practical alternative responses for this inventory, interviewers "asked [Army] officers to 'tell a story' about a personal experience from which they learned something important about leadership at the company level" (Hedlund et. al, 1998). From these interviews Hedlund et al (1998) formed the TKML items chosen for their ability to discriminate between officers in the dimensions of experience ("experienced" vs. "novice") and leadership effectiveness ("more" vs. "less"). There are three versions of the TKML, one for each level in the chain of command (platoon, company, battalion) explored. In the process of validating the company level TKML, Hedlund et al (1998) not only obtained 20 situational judgment test (SJT) questions about everyday experiences, but they also obtained 360 ratings of leadership, interpersonal, task, and overall performance from bosses, peers, subordinates, and selves. Hedlund et al (1998) found that the company level inventory had an internal reliability of .75 ($p < .05$), and that tacit knowledge was related to effectiveness ratings at all leadership levels. Overall, Hedlund et al (1998) found that the TKML provided a potential developmental tool for Army leaders. Further work (e.g. Antonakis et al, 2002; Cianciolo et al., 2006) has demonstrated the validity of the TKML as a measure of Army leadership, showing, for instance, that it effectively differentiates experts (senior officers) from journey-level experts (junior officers). This increase in validity has come about with changes in the scoring procedure for the TKML, which will be used in this paper. Also, in the intervening decade SJT's have become popular means for assessing many different areas of Army performance (McDaniel & Nguyen, 2001), so now we are returning to use these data to compare the quality and rating accuracy of each level of these 360 assessments.

The TKML is a multifactor instrument that deals with many of the same issues as the OER. It explicitly has questions about interpersonal, task and general components of leadership, and it implicitly deals with the tacit, practical knowledge of such common Army concerns as taking care of Soldiers, dealing with your superior, and interpersonal and task issues. It is constructed to provide an objective assessment of the accuracy of self ratings that can be compared with others' ratings as well. It offers the possibility of providing a new perspective and informing theoretical account of the relationship of the accuracy of self assessment and leadership performance. It offers some possibility of countering the one-sided consensus that self assessment is inaccurate and unreliable.

Hypotheses:

H1: The strongest correlations are between the TKML and superior and self ratings, since these two groups are best informed about the issues and the focal individual.
H2: Overall, self ratings are the most accurate, not as assessed by others' ratings, but as measured by the TKML.

Methods

Sample

Hedlund et al. (1998) administered our battery of assessment instruments (each described in detail below) to a representative sample of Army officers at the three levels under investigation. For this new analysis, we studied 93 male Captains, from the perspective of their own ratings and performance and from the ratings of their commanders, peers, and subordinates from this sample, drawn from a pool of 44 battalions stationed at six posts around the United States. We had supervisor ratings for 63 Captains, at least one peer rating for 89, and at least one subordinate rating for 76. Mean scores were computed for missing data. Table 1 shows the distribution of battalions across these six posts. For the purpose of our investigation, combat service support was removed prior to analysis in order to make the data more homogenous. This loss of data is primarily due to the fact that unit operational requirements often precluded the gathering of complete data.

Table 1

Pool of Battalions Sampled by Post

Post	Battalions Sampled
Campbell	10
Drum	5
Carson	4
Bragg	10
Lewis	5
Hood	10

Instruments

Tacit knowledge for military leadership (TKML) inventories

Tacit knowledge inventories of the type developed in our research are intended to measure the experience-based, practically-oriented knowledge of individuals. An inventory consists of a series of problems or scenarios, briefly described. Each scenario is accompanied by

a set of possible responses (5 to 15 items) to the situation that it describes. Respondents were asked to rate the quality or advisability of each response option using a nine-point Likert scale. In the original report (Hedlund et al., 1998) three versions of the Tacit Knowledge for Military Leaders (TKML) inventory corresponding to the organizational levels of platoon, company, and battalion were used. We report on the data collected from the company level inventory. Figure 1 shows a sample question taken from the company commander TKML.

Figure 1. Sample Question from the Tacit Knowledge Inventory for Military Leaders.

1	2	3	4	5	6	7	8	9
Extremely Bad		Somewhat Bad		Neither Bad Nor Good		Somewhat Good		Extremely Good

You are a company commander, and your battalion commander is the type of person who seems always to "shoot the messenger"—he does not like to be surprised by bad news, and he tends to take his anger out on the person who brought him the bad news. You want to build a positive, professional relationship with your battalion commander. What should you do?

_____ Speak to your battalion commander about his behavior and share your perception of it.
_____ Attempt to keep the battalion commander "over-informed" by telling him what is occurring in your unit on a regular basis (e.g., daily or every other day).
_____ Speak to the sergeant major and see if she/he is willing to try to influence the battalion commander.
_____ Keep the battalion commander informed only on important issues, but don't bring up issues you don't have to discuss with him.
_____ When you bring a problem to your battalion commander, bring a solution at the same time.
_____ Disregard the battalion commander's behavior: Continue to bring him news as you normally would.
_____ Tell your battalion commander all of the good news you can, but try to shield him from hearing bad news.
_____ Tell the battalion commander as little as possible; deal with problems on your own if at all possible.

TKML scoring procedures

Procedures for scoring tacit knowledge inventories pose unique challenges in establishing a "correct" answer for test items. Unlike questions on traditional achievement or intelligence tests, less certainty can be attached to the correctness of specific responses on tacit knowledge tests (Legree, 1995). As the sample question in Figure 1 illustrates, a respondent's ratings depend on his or her interpretation of the problem, an interpretation that is assumed to rely upon

knowledge gained through experience of the Army's cultural values mixed with their own values from prior experience. Therefore, one appropriate standard for response quality is that provided by a group of highly experienced and successful practitioners.

In the previous investigation (Hedlund et al., 1998) of officer tacit knowledge, expert response profiles were obtained for the Company Inventory from a group of highly select officers who had recently demonstrated outstanding performance (as defined by the Army's performance evaluation, promotion, and selection system). These officers completed the TKML inventory, providing us with the raw data to construct expert profiles. Majors and Lieutenant Colonels attending the Pre-Command Course (PCC) served as an expert group for the company level inventory. This is a very select group of officers who, based primarily on their success as company commanders, have been chosen to command battalions. Selection for battalion command is an extremely competitive process. By virtue of their experience and accomplishments at the level in question, this group of officers was deemed to represent the experienced and knowledgeable practitioner.

For the TKML inventory, an expert profile was constructed which represents the mean of the experts' ratings for each response option within a question. The expert profile was correlated with the mean of the 93 Captains' own ratings in this sample on the company TKML. The correlation of 0.997 ($p < .001$) confirmed a virtually identical standard for the Captains in this research. Accordingly we analyzed the TKML using consensus based assessment (CBA) standards (Legree et al., 2000).

CBA and Factor Analysis

CBA is often computed by using Pearson r correlation of each person's Likert scale judgments across a set of items against the mean of all people's judgments on those same items (e.g., Antonakis et al, 2002). The correlation is then a measure of that person's proximity to the consensus. It is also sometimes computed as a standardized deviation score from the consensus means of the groups. These two procedures are mathematically isomorphic. If culture is considered to be shared knowledge; and the mean of the group's ratings on a focused domain of knowledge are considered a measure of the cultural consensus in that domain; then both procedures assess CBA as a measure of an individual person's cultural understanding and internalization of social norms.

However, it may be that the consensus agreement with items is different for some sub-groups of a population. For instance, conservatives who are libertarians may feel more concerned about invasion of privacy than conservatives who feel strongly about law and order. In fact, standard factor analysis (Nunnally, 1967) brings this issue to the fore with its weighting of factor scores by component scores (where component scores are generally the mean correlations of an item's scores with the mean of all scores).

In either centroid or principle components analysis (PCA) factor analysis the first factor scores are created by multiplying each individual's rating by the correlation of the factor (usually the mean of all standardized ratings for each person) against each item's ratings. This multiplication weights each item by the correlation of the pattern of individual differences on each item (the component scores). If consensus is unevenly distributed over these items, some items may be more focused on the overall issues of the common factor. If an item correlates highly with the pattern of overall individual differences, then it is weighted more strongly in the overall factor scores. Now, this weighting implicitly also weights the CBA score, since it is

those items that share a common CBA pattern of consensus that are weighted more in factor analysis. In this sense, the factor scores are generally highly correlated with the deviation or correlation CBA score. So a factor score of Likert ratings is not simply an average score; it is instead a measure of those scores where some consensus holds, and so it can generally be used as the CBA measure.

Leadership Effectiveness Survey

We used the Leadership Effectiveness Survey (LES) developed by Hedlund et al. (1998) to measure the criterion of leadership effectiveness. The LES consisted of three single-item measures that asked respondents to rate effectiveness of other officers on a seven-point scale. An example question from the LES is shown in Figure 2. The survey called for three separate judgments of effectiveness in the interpersonal and task-oriented domains of leadership as well as an overall assessment of leadership effectiveness. The format for the LES questions was modeled after the normative process used by senior level raters on the Officer Evaluation Report (OER).

In order to obtain multiple perspectives of a Captain's leadership effectiveness in our investigation, respondents were asked to rate the effectiveness of themselves, their immediate supervisor, their subordinate officers, and peers in their unit. For example, a Captain would rate his superior (Lieutenant Colonel), his peer (another Captain), and his subordinate (Lieutenant). By administering the LES to intact chains-of-command, we also obtained multiple ratings of effectiveness from each perspective, with the exception of supervisors since each officer only has one immediate supervisor. For those cases in which multiple ratings were obtained (e.g., subordinate, peers), the ratings were correlated to establish significant correlations before aggregating the data. A mean rating was computed for each of the effectiveness dimensions (overall, task, and interpersonal).

Figure 2. Sample Question from the Leadership Effectiveness Survey.

Rate your Company Commander:

Think about your company commander. Compared to all other company commanders you have known, **how effective is your company commander, overall, as a leader?** Please circle the number under the statement that best corresponds to your rating.

1	2	3	4	5	6	7
The Best	One of the Best	Better than Most	As Good as Most	Not Quite as Good as Most but still gets the job done	Well Below Most	The Worst

Data Collection Procedures

Hedlund et al. (1998) obtained access to battalions under the auspices of the U.S. Army Research Institute and visited each during its "umbrella weeks" – periods when the units were not deployed on training exercises and were available to participate in research efforts. Selection of units for participation was made by division, corps, or brigade staff. Scheduling and pre-shipment of surveys was coordinated by a point-of-contact at each post. At the scheduled appointment time, the entire available officer chain-of-command for each battalion (approximately 25-30 officers) met at a central location, usually in their battalion conference room, where they completed the test battery including the TKML and the LES as described above.

Data-collection sessions began with an introductory briefing by the visiting researchers. Subjects were introduced to the investigation as follows:

> "We're here as part of a joint Yale/USMA research project under contract to the Army Research Institute. They've asked us to examine the role of informal or "tacit" knowledge in Army leadership. Tacit knowledge is practical knowledge, grounded in personal experience, which is not explicitly taught and is often difficult to articulate. The goal of this research is to improve the process of leader development through job assignment by understanding the hidden or tacit knowledge that makes leaders effective.
>
> Today we are going to ask you to fill out some questionnaires. Some of these will draw on your knowledge of Army leadership and some will draw on more general knowledge. We are also going to ask for some ratings of the people you work with. Some of this you may find difficult, but we are going to strictly protect your anonymity and confidentiality, as I'll describe in a moment, so we hope that you will answer candidly.
>
> All of the data we collect today will help us to answer the questions that the Army has asked us to answer—basically about the relationship between informal knowledge, experience, effectiveness, and other variables. We need your best effort here today—your most thoughtful and candid judgments—in order to ensure that the Army gets its money's worth out of this research."

Officers were assured of the absolute confidentiality of their responses and their informed consent was obtained. Officers, working at their own pace, then completed the instruments in the test battery. Each session ended when all officers in the battalion had completed the test battery, typically after three to four hours. Completed surveys were inventoried, coded to preserve the subjects' anonymity and to facilitate in later analysis, and shipped to Yale University.

Data Analytic Procedures

Since there were initially three different versions of the TKML, one for each level of investigation (platoon, company, battalion), for the purpose of this investigation only the company level data were utilized. This is the level that has a complete 360 degree set of ratings (self, superior, peer, and subordinate.)

The first step was to examine the intercorrelations among the dimensions of the LES (overall, task, interpersonal) for each type of rater (subordinate, peer, superior). Multitrait -

multimethod (MTMM) analysis (Campbell & Fiske, 1959) is typically used to provide evidence of convergent and discriminate validity. Ratings of the same trait (e.g., leadership dimension) are expected to correlate more highly (converge) using different methods (e.g., raters) than ratings across traits using a single method. In our research, we obtained ratings from multiple perspectives based on the assumption that different raters would have different perceptions of leadership effectiveness. Therefore, we expected the correlations to be lower across raters for the same leadership dimensions than across dimensions for a single rater perspective. Within each rater perspective, we also examined the correlations between task, interpersonal, and overall ratings for evidence that these aspects of leadership effectiveness represented distinct constructs.

Our second step was to compute the Consensus Bases Assessment (CBA) score for the TKML. Originally, Hedlund et al. (1998) computed this by using a deviation score from the mean of the experts' scores for each question. However, in the intervening decade, Situated Judgment Tests have been increasingly scored by consensus based methods, either using the mean of the group itself to create a standardized deviation score, or a correlation with the mean of the group. These are equivalent measures. A more powerful technique uses the first factor of the principal component analysis as the CBA score. Since this factor analysis weights each person's score by the correlation of each person with the mean of all the respondents, the factor score offers the best compromise between the deviation score and the correlation coefficient (Legree, Psotka, Tremble, and Bourne, 2005).

After examining properties of the TKML and LES, we computed the intercorrelations among the TKML and LES.

Results

Table 2

Means and Standard Deviations of all ratings by self, superior, peer, and subordinate

Mean Ratings	Mean	Standard Deviation	N
Superior	2.96	1.24	63
Self	2.77	.72	90
Peer	3.16	.66	89
Subordinate	3.25	1.31	76

The mean ratings were significantly different ($F = 3.94$; df = 3,314; $p < .01$). The self rating was significantly less than subordinate ratings (t=3.1; df = 74; $p < .01$) but self rating was not significantly different from superior ratings (t=1.6; df = 761; $p < .11$). (cf. Table1).

Intrarater agreements

The mean intercorrelations within each of the self, superior, peer, and subordinate raters on the LES among interpersonal, task, and overall ratings were respectively: self ($r = .554$), superior ($r = .762$), peer ($r = .731$), subordinate ($r = .856$). All Intrarater correlations were significant with $p < .001$. The high correlations provided justification for combining them into an average rating for each rater. The intercorrelations among raters for these mean ratings of the Captians are shown in Table 3 below.

Table 3

Inter-rater correlations among mean ratings by self, superior, peer, and subordinate

Intrarater	Self	Superior	Peer	Subordinate
Self	1.0			
Superior	.169	1.0		
Peer	.278*	.277*	1.0	
Subordinate	.298*	.172	.323*	1.0
N	90	63	89	76

$p < .01$*

Table 3 shows a pattern where Peer ratings had the highest intercorrelations with Self, Superior, and Subordinate raters. Superiors' ratings had the lowest correlations with Self and Subordinate raters.

Inter-rater correlations

The mean correlation of each rating set with all the other raters over the same dimensions for each of interpersonal, task, and overall ratings is shown below in Table 4. The interrater correlations within each dimension appear reasonably consistent within each dimension and across dimensions. No set of raters appear noticeably superior to any of the others. The values and pattern of correlations appears very similar to the inter-rater correlations in Table 3.

Table 4

Mean Inter-rater correlations among ratings by self, superior, peer, or subordinate with all other raters for interpersonal, task, and overall leadership ratings

Inter-rater	Interpersonal Leadership	Task Leadership	Overall Leadership	N
Self	.203	.229	.296*	90
Superior	.247	.187	.198	63
Peer	.269*	.275*	.314*	89
Subordinate	..275*	.172	.335*	76

p < .01*

Intercorrelation of TKML with LES

The first principal components factor score was used as the measure of the CBA score on the TKML. It yielded significant correlations (Table 5) with both the Self Ratings and the Superior Ratings. The difference between these two correlations was not significant. The CBA score was regressed onto the standard deviation of the TKML ratings, in combination with the Self, Superior, Peer, Subordinate Ratings, and all two and three way combinations of these ratings. This regression took advantage of the multisource, 360 rating information provided by all sources to demonstrate how to combine the information from all sources. It also took advantage of the relationship between TKML and standard deviation of the TKML ratings;

namely, that officers who feel confident about the strength or weakness of an item will tend to give it a more extreme rating, while those who can see little difference among items will have a lower standard deviation since they tend to give all alternatives the same or similar rating. This comparison resulted in a significantly higher correlation between mean Self Rating and CBA (r = .511) than Superior Ratings and CBA (r = .242; p < .02) (cf. Table 4); however, the general order of intercorrelations remained identical to the unregressed TKML CBA scores.

Table 5

Intercorrelations among Captains' mean Self Ratings, CBA score, and CBA score and SD regressed on mean Peer Ratings

Mean Ratings	**TKML Deviation Score**	**TKML CBA score**	**CBA regressed on peer ratings**	**N**
Self	-.213	.417*	.511*	90
Superior	-.077	.340*	.242	63
Peer	-.007	.220	.233	89
Subordinate	.084	.065	.075	76
N	92	92	92	

p < .01*

Discussion

The novel finding of this research is that self ratings stand out as significantly more accurate than all other ratings when they are compared against an objective standard of leadership tacit knowledge about a broad range of leadership skills, values, and abilities. Intercorrelations among all the raters of the Captains in this research tend to confirm previous studies that the mean accuracy is about 0.3 of overall self ratings correlations with other ratings sources (cf. Table 4.) The intercorrelation data is neutral about whether peers, subordinates, or superiors are better informed about these Captains' leadership potential, skills, values and abilities. None of the intercorrelations are significantly different from each other, or from the self ratings. The fact that the correlations of all the ratings from the four different rater groups were similar suggests that others' ratings do not deserve special consideration in evaluating the accuracy of self ratings. Since all of the others have a similar perspective on these Captains it is somewhat surprising that they do not correlate more with each other than with the self ratings. Since all of the others have a similar/external perspective on these Captains, it is surprising that they do not correlate more with each other than with the self ratings. Since all self ratings are given from a different and unique, internal subjective perspective it is not surprising that self ratings should correlate less well with the others' ratings. Of course, this subjective view is tempered by continuous feedback form other sources, especially superiors, subordinates and peers. This feedback and social interaction acts to maintain some similarity among these intercorrelations.

The finding is interesting and worthy of more research and investigation, but it is not definitive proof that self-ratings stand out as significantly more accurate. Without knowing the true correlation between tacit knowledge and performance, the interpretation that stronger

correlations between a rating source and TKML scores reflect more accuracy may or may not be correct. Also, this finding does not necessarily dispel the notion that different rating sources have different, but valid, perspectives on a ratee's performance. Thus, differences in correlations between tacit knowledge and ratings from different sources may not be a reflection of inaccurate ratings, but instead reflect that the different perspectives of performance are differentially related to the tacit knowledge construct.

There is no spurious relationship between Self Ratings and TKML

A superficial review of the data might suggest that there is a confound in the research since self ratings and the TKML share a common method: Likert scaling of knowledge, values and skills. However, the overall means on the TKML that the peers and Captains provide correlates .997 with the means provided by an entirely different standardizing group of selected superior officers (Hedlund et al., 1998), suggesting that if the standards provided by these Captains' superiors (Lieutenant Colonels) were used as the components on the TKML factor analysis, the result would not have been different, and the TKML truly provides evidence that self ratings are more accurate than others' ratings.

Pattern of quality of ratings and the TKML

Although there are no significant differences among the ratings intercorrelations; in general, raters appear to be most in agreement with their subordinates. For instance (in Table 3) self ratings correlate best with their subordinate ratings; and superior ratings correlate best with peers (who are the entire set of those superiors' subordinates.) This may have something to do with the fact that subordinates see their leaders most in leadership actions, especially during their official evaluation interviews, and so use the same standards for assessment. This agreement between subordinates and superiors provides some evidence that self ratings may use the same standards, but have a different perspective. When Captains rate themselves and see themselves different from their superiors, this suggests that they are still using the same standards as their superiors but have a different, internal perspective and understanding of themselves. It is therefore somewhat less surprising that regressing the TKML on peers' ratings strengthens the correlation with self ratings and weakens the correlations with superiors' ratings (cf. Table 5, column two versus three). We suggest that this strengthens our interpretation that regressing the TKML on peers' ratings incorporates peers' true views about the other Captains in this investigation from a perspective that superiors do not share. As a result, somewhat counter intuitively, the use of an objective measure, the TKML, provides a stronger rationale for incorporating 360, multisource rating assessment perspectives into the appraisal process.

Superiors versus Peers and Subordinates

Superiors traditionally possess the right to make high stakes administrative decisions about their subordinates for rewards and promotions. The differences in others' ratings are often seen as legitimate differences in perspective on a focal individual (Toegel & Conger, 2003) but we suggest they are not equally accurate. Our results using the TKML as an objective measure of leadership potential, skill and ability suggests that superiors are better judges than others, thus confirming the wisdom of this practice. Although the TKML is most strongly in agreement with

self ratings, superiors' ratings come in a close second (cf. Table 5.) Superiors in this research have gone through this selection process many times and have, in general, twice the length of experience with Army culture and values than the Captains in this research. Not only do they possess this superior cultural understanding, they have many shared goals and experiences with their subordinates. Although the Army and circumstances may have changed since their tenure as Captains, the leadership knowledge and skills remain as personal, ambiguous and undefined as ever. These enduring issues are captured by the TKML and show little change in the pattern of responses even a decade or more after its creation.

Conclusions

The Value of a Validated Objective Measure

The decades of refined CBA measures of leadership ability and potential have paid off in the strong and significant relations reported here. The earlier report (Hedlund et al, 1998) found only weakly significant correlations between self ratings and TKML, because their early TKML scoring standard used deviations from expert means. Converting to CBA factor analytic scoring of the TKML using a peer-based standard based on a sophisticated principal components factor analysis increased the correlation between self ratings and TKML by several multiples. This allowed us to argue convincingly that self ratings truly were superior to the other rating sources in a multi-layered 360 assessment. For instance, van Hooft, Flier, & Minne (2006) found much weaker correlations between ratings and a standardized in-box instrument. The implications of this singular finding need to be explored in future research. In particular, it might be very useful to have an objective measure like the TKML for both raters and ratees to discuss to pinpoint agreement and deficiencies. Subjective estimates are valuable but may perhaps be improved by incorporating objective measures as well.

Epilogue

Van Velsor, Taylor & Leslie (1993) report that self rating accuracy (as measured against others' ratings) have been shown to be related to effective leadership traits, such as self esteem, intelligence, achievement, status, locus of control, adaptation to feedback, and overall performance. All of these variables and more could be investigated with this new paradigm.

References

Antonakis, J., Hedlund, J., Pretz, J., & Sternberg, R.J. (2002). Exploring the nature and acquisition of tacit knowledge for military leadership (Research Note 2002-04). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Ashford, S.J. (1989). Self-assessment in organizations: A literature review and integrative model. *Research in Organizational Behavior, 11,* 133-174.

Bass, B. & Yammarino, F. (1991). Congruence of self and others' leadership ratings of naval officers for understanding successful performance. *Applied Psychology: An International Review, 40,* 141-163.

Borman, W.C. (1997). 360° ratings: An analysis of assumptions and a research agenda for evaluating their validity. *Human Resource Management Review, 7,* 299-315.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Church, A.H. & Bracken, D.W. (1997). Advancing the state of the art of 360-degree feedback: Guest editors' comments on the research and practice of multirater assessment methods. *Group & Organization Management, 22,* 149-191.

Cianciolo, A.T., Matthew, C., Wagner, R.K., & Sternberg, R.J. (2006). Tacit knowledge, practical intelligence, and expertise. In K.A. Ericsson, P.J. Feltovich, R.R. Hoffman (Ed.), *The Cambridge handbook of expertise and expert performance* (pp. 613-632). New York, New York: Cambridge University Press.

Conway, J.M. & Huffcutt, A.I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self- ratings. *Human Performance, 10,* 331-360.

Dörner, D. & Kreuzig, H. (1983). Problemlosefahigkeit und intelligenz. *Psychologische Rundschaus, 34,* 185-192.

Harris, M.M. & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41,* 43-62.

Hedlund, J., Forsythe, G. B., Horvath, J. A., Williams, W. M., Snook, S., & Sternberg, R. J. (2003). Identifying and assessing tacit knowledge: Understanding the practical intelligence of military leaders. *Leadership Quarterly, 14,* 117-140.

Hedlund, J., Horvath, J.A., Forsythe, G.B., Snook, S., Williams, W.M., Bullis, R.C., Dennis, M., & Sternberg, R.J. (1998). Tacit knowledge in military leadership: Evidence of construct validity (Tech. Rep. No. 1080). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hedlund, J., Sternberg, R. J., & Psotka, J. (2000). Tacit knowledge for military leadership: Seeking insight into the acquisition and use of practical knowledge (Tech. Rep. No. ARI TR 1105). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Legree, P. J. (1995). Evidence for an oblique social intelligence factor. *Intelligence, 21,* 247-266

Legree, P. J., Martin, D. E., & Psotka J. (2000). Measuring cognitive aptitude using unobtrusive knowledge tests: A new survey technology. *Intelligence, 28,* 291- 308.

Legree, P. J., Psotka J., Tremble, T. R. & Bourne, D. (2005). Using Consensus Based Measurement to Assess Emotional Intelligence. In R. Schulze & R. Roberts (Eds.), International Handbook of Emotional Intelligence. (pp 99-123). Berlin, Germany: Hogrefe & Huber.

McCauley, C. & Lombardo, M. (1990). Benchmarks: An instrument for diagnosing managerial strengths and weaknesses. In K.E. Clark & M.B. Clark (Eds.), Measures of leadership (pp. 535-545). West Orange, NJ: Leadership Library of America.

McDaniel, M. A. & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. International Journal of Selection and Assessment, *9,* 103-113.

Motowidlo, S. J., Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. Journal of Occupational and Organizational Psychology, *66,* 337-344.

Nunnally, J.C. (19676). Psychometric theory. New York: McGraw-Hill.

Paulhus, D. L., Lysy, D. & Yik, M. S M. (1998). Self-report measures of intelligence: Are they useful as proxy measures of IQ? *Journal of Personality, 66,* 525-554.

Scullen, S.E. (1997). When ratings from one source have been averaged, but ratings from another source have not: Problems and solutions. *Journal of Applied Psychology, 32,* 880-888.

Salam, S., Cox, J.F., & Sims, H.P. (1997). In the eye of the beholder: How leadership relates to 360-degree performance ratings. *Group & Organization Management, 22,* 185-209.

Sternberg, R.J., Forsythe, G.B., Hedlund, J., Horvath, J. A., Wagner, R.K., Williams, W.M., Snook, S., & Grigorenko, E.L. (2000). Practical intelligence in everyday life. New York: Cambridge University Press.

Toegel, G. & Conger, J.A. (2003). 360-degree assessment: Time for reinvention. *Academy of Management Learning and Education, 2,* 297-311.

Tornow, W.W. (1993). Perceptions or reality: Is multi-perspective measurement a means or an end? *Human Resource Management, 32,* 221-229.

van Hooft, E.A.J., van der Flier, H., & Minne, M.R. (2006). Construct validity of multi-source performance ratings: an examination of the relationship of self-, supervisor-, and peer-ratings with cognitive and personality measures. *International Journal of Selection and Assessment, 14,* 67-81.

van Velsor, E., Taylor, S., & Leslie, J. (1993). An examination of the relationships between self-perception accuracy, self-awareness, gender and leader effectiveness. *Human Resource Management, 32,* 249-264.

Yammarino, F. & Atwater, L. (1993). Understanding self-perception accuracy: Implications for human resource management. *Human Resource Management, 32,* 231-247.

Yammarino, F. & Waldman, D.A. (1993). Performance in relation to job skill importance: A consideration of rater source. *Journal of Applied Psychology, 78,* 242-249.