This article argues that assumptions surrounding 360° ratings should be examined; most notably, the assumptions that different rating sources have relatively unique perspectives on performance and multiple rating sources provide incremental validity over the individual sources. Studies generally support the first assumption, although reasons for interrater disagreement across different organizational levels are not clear. Two research directions are suggested for learning more about why different organizational levels tend to disagree in their ratings and thus how to improve interpretation of 360° ratings. Regarding the second assumption, it is argued we might resurrect the hypothesis that low-to-moderate across organizational level interrater agreement is actually a positive result, reflecting different levels' raters each making reasonably valid performance judgments but on partially different aspects of job performance. Three approaches to testing this hypothesis are offered.

Recently, there has been a surge of interest in 360° rating programs (e.g., Bracken 1996; Church & Bracken 1997; Hazucha, Hezlett, & Schneider 1993; London & Beatty 1993; Tornow 1993). 360° programs involve generating performance evaluations on target ratees from multiple sources such as supervisors, peers, subordinates, and customers, as well as from the targets themselves. Often 360° ratings are used for developmental feedback. The basic notion here is that to improve performance, one first needs to get insight about his or her weaknesses in job performance. Evaluations from employees who have close working relationships with the ratee are seen as an effective way to provide this insight. Although feedback for development is a frequent application of the 360° method, additional purposes include for promotion, compensation, succession planning, or other administrative decisions.
Regarding how this method might be viewed in the context of industrial and organizational psychology, Tornow (1993) provided a useful discussion. He pointed out that the science and practice-related views of 360° evaluations may be very different. The science "take" on this method involves interest in accuracy of measurement with the ratings and in reducing interrater disagreement to minimize error in ratings. The practice-related view focuses more on the usefulness of the method for improving individual and organizational effectiveness. Practitioners may value the multiple perspectives on a ratee's performance and stress the learning that can take place from the different sources' views of his or her performance. They may also see considerable value in simply giving all relevant parties the opportunity to express their views on target ratees' performance. With this latter view, the means (i.e., rating process) may be more important than the end (i.e., the validity of the ratings).

In a discussion of the Tornow paper, Dunnette (1993) acknowledged the two different perspectives on 360° ratings, but pointed out that having good science supporting practice is also important. I would add that there are important scientific questions to be asked about 360° feedback ratings that are fundamental in supporting the use of this technique. The most important is do additional rating sources provide incremental validity beyond ratings from a single source (e.g., supervisors)? This is especially critical if the ratings are being used for administrative decision making. However, even in the case of 360° for developmental feedback, receiving reasonably valid performance information (i.e., correlating highly with actual ratee performance levels) is important to properly guide behavioral change. That is, if appropriate change is to occur, precise and focused knowledge about the areas needing improvement must first be provided.

Thus, there are two primary assumptions made by practitioners who use 360° ratings. One assumption is that the multiple sources of ratings (e.g., peers, subordinates, etc.) each offers at least somewhat unique data on ratee performance. It follows that we would not expect or even necessarily desire high interrater agreement across sources. However, provided organizational level is the unit of importance in gaining multiple perspectives on ratee performance, interrater agreement within sources is desirable. That is, if we want the peer perspective, the subordinate perspective, etc., regarding the performance of ratees, we would like to obtain relatively good agreement within each of these perspectives. To clarify this assumption, let's consider the alternatives to the interrater agreement results just described. If we obtain very high interrater agreement across sources, there is little need to collect ratings from multiple sources. If interrater agreement within sources is low, we might suspect that the ratings contain error, with at least some of the raters providing invalid evaluations from that source's perspective. The second assumption is that having additional rating sources providing evaluations yields incremental validity over and above a single source (e.g., supervisors).

Thus, the desirable psychometric properties of 360° ratings are not typically discussed. However, the assumptions around this method point toward the expectation that interrater agreement across sources should be low to moderate and that agreement within sources should be higher. Also, additional
sources of rating data should contribute enhanced validity for the performance evaluations.

By far the most research has been conducted on the first topic listed above. Regarding this question, several studies have shown that interrater agreement within-organizational level (e.g., peer-peer or supervisor-supervisor agreement) does, in fact, tend to be greater than across-level agreement (e.g., peer-supervisor). For example, Berry, Nelson, and McNally (1966) had raters from two supervisory levels evaluate U.S. Navy personnel separated by a two-year interval. Correlations between performance ratings within organizational level across the two years averaged .55. Across the two levels, the correlations were significantly lower ($r = .34$). Similarly, Gunderson and Nelson (1966) analyzed peer and supervisor performance ratings of personnel in a remote setting (the Antarctic) and found that within-organizational level interrater reliabilities were higher than across-level reliabilities on every dimension (medians .74 vs. .50). Also relevant, Borman (1974) compared a multitrait-multimethod (MTMM) matrix using peer-supervisor correlations to estimate convergent validities with a MTMM matrix employing within-organizational level interrater reliabilities to index convergent validity. For 5 of the 7 dimensions, within-level interrater agreement was higher than the corresponding across level interrater agreement ($ms = .45$ vs. .32).

The efficacy and utility of across level ratings, in general, and 360° rating programs in particular, rely on understanding the nature of the rating differences observed across rater levels. That is, in order to interpret and appropriately use across-level rating information for developmental feedback or operational decisions, it is important to know, for example, if raters from different organizational levels actually observe different behaviors, or see essentially the same behavior but interpret it differently or weight it differently. Understanding the reasons for interrater disagreement across organizational level should help to interpret multi-source ratings. We now turn to a review of models attempting to explain reasons for interrater disagreement across organizational levels. This review is restricted to literature dealing with the peer and supervisor rating sources, as there were sufficient studies that examined these rating sources. This was not the case for research involving subordinate or customer rating sources. As discussed above, a central focus in our review is toward understanding why raters from different organizational levels disagree in their performance ratings and how this holds meaning for the 360° approach. Accordingly, considerable attention is directed toward two basic hypotheses that have been offered to explain why different level raters disagree. These are that raters at different levels (1) use different dimensions or weight dimensions differently as a basis for evaluating performance or (2) actually observe different performance-related behaviors in target ratees.

HYPOTHESES FOR PEER-SUPERVISOR DISAGREEMENT IN RATINGS

Table 1 presents the two central hypotheses regarding why peers and supervisors disagree in their performance ratings to the extent they do. As men-
TABLE 1
Hypotheses About Why Across-Organizational Level Interrater Agreement is Relatively Low

1. Different Levels' Raters Use Different Performance Models
   a. They Use Different Dimensions or Define Dimensions Differently
      • When asked to generate performance dimensions, different levels produce different dimensions (Borman 1974; Zedeck et al. 1974)
      • Raters from different levels use different patterns of cues when making performance judgments (Oppler et al. in preparation)
      • Raters from different levels have different implicit performance theories (Oppler & Sager 1992; Pulakos & Borman 1986)
   b. They Use Similar Dimensions but Different Weights for the Dimensions in Making Overall Performance Judgments
      • Across-level interrater agreement on individual dimensions is high relative to agreement on overall performance dimensions (Harris & Schaubroeck 1988)
      • Regression models, with overall performance ratings regressed on dimension ratings, show significant differences for different organizational level raters (Borman et al. 1995; Pulakos et al. 1996; Tsui 1984; Tsui & Ohlott 1988)

2. Different Levels' Raters Use Different Samplings of Ratee Behavior in Making Performance Judgments
   a. They Have Different Opportunities to Observe Ratees, and Thus See Different Ratee Behavior.
      • Different levels have different opportunities to see ratees in various roles (Tsui 1984)
      • Peer raters have more opportunity to view ratees than do supervisors, and therefore self-peer interrater agreement (same organizational level) should be higher than self-supervisor or peer-supervisor interrater agreement (different levels) (Harris & Schaubroeck 1988)

mentioned, the first hypothesis is that different levels' raters use different performance models when making ratings. Beneath the first hypothesis in the table are two different versions of the hypothesis, along with essentially operational definitions of each version. The second hypothesis is that different levels' raters use different samplings of ratee behavior in making performance judgments. Table 2 summarizes results for studies examining these hypotheses. It should be mentioned that the different versions of the hypotheses and operational definitions are not completely independent from each other. This becomes clear when we see that some of the empirical studies to be reviewed apply to more than one version. Nonetheless, we believe this summary of the various explanations for across-organizational level interrater disagreement provides a useful framework for assessing the empirical research bearing on this issue and for helping to interpret 360° ratings.

Hypothesis 1, Version 1. This hypothesis is that different level raters use different dimensions or that they somehow define dimensions differently when making ratings. The most direct test of this hypothesis has been to have individuals from different organizational levels simply generate the important performance dimensions from their own perspective and compare these dimension sets across level. Borman (1974) took this approach with
secretaries in a university setting and their instructor "bosses." As part of a behaviorally anchored rating scale development process, each group independently generated what they believed were the important performance dimensions, and the results demonstrated that the dimension sets "showed only modest conceptual similarity" (pp. 105). No quantitative comparison of the two dimension sets was offered, however. Similarly, Zedeck, Imparato, Krausz, and Oleno (1974) had two groups, peers and supervisors of hospital nurses, develop performance appraisal dimensions independent of each other. Subjective comparisons suggested considerable overlap in this case, and a consensus meeting between the groups resulted in agreement on a single dimension set (Zedeck et al. 1974).

Another way to address the different-dimensions version of Hypothesis 1 is to evaluate whether or not raters from different organizational levels use different patterns of factors or cues when making performance judgments. The reasoning is that if different levels attend to different cues, they may have at least partially different dimension sets influencing their judgments about ratee performance. Relevant to this version of Hypothesis 1, in a Project A sample (a large scale selection and classification study conducted in the U.S. Army [Campbell 1990]), Oppler, Pulakos, and Borman (in preparation) developed a path model with the cue (or exogenous) variables being ratee ability, technical proficiency (from work sample tests), job knowledge, personality (two variables), and objective measures of commendable and disciplinary problem behavior. The endogenous variable in the model was the performance ratings. Results showed very similar patterns of path coefficients when a peer rating model was compared to a supervisory rating model, with ratee problem behavior, dependability, and achievement orientation demonstrating the strongest effects on both peer and supervisor ratings.

A study by Borman, White, and Dorsey (1995) came to essentially the same conclusion. In another Project A sample, ratee technical proficiency and dependability had the largest path coefficients to the overall job performance ratings made by supervisors and peers. However, Pulakos, Schmitt, and Chan (1996) studied supervisor and peer ratings of professionals in a large government agency and found that significantly different path coefficients were evident for the paths between cue variables and overall performance ratings for the two sources. Especially large differences occurred for the written technical proficiency-rating and role-play proficiency-rating paths.

Still another approach to test the different-dimension version of Hypothesis 1, albeit quite indirect, comes from the Harris and Schaubroeck (1988) meta-analysis. They reasoned that if there is low interrater reliability across-organizational level, both for the individual dimension and overall performance ratings, then the different levels' raters may be using different dimensions. They argued that moderate levels of interrater agreement for dimension and overall performance ratings across the peer and supervisor levels (corrected r = .62) provided some evidence that similar dimensions were being employed by the two levels' raters.

Finally, comparisons of what might be called implicit performance theories held by supervisors versus peers also seem relevant to this hypothesis. First,
## TABLE 2
### Summary of Studies Examining Reasons for Across-Organizational Level Interrater Disagreement in Ratings

<table>
<thead>
<tr>
<th>Study and Rating Sources</th>
<th>Hypotheses</th>
<th>Test</th>
<th>Results and Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berry et al. (1966)</td>
<td>Use Different Dimensions</td>
<td>Within-level vs. Across-level IR</td>
<td>Within-level IR Higher Than Across-level IR</td>
</tr>
<tr>
<td>(1st and 2nd line supervisors)</td>
<td></td>
<td></td>
<td>Moderate Convergent and Low Discriminant Validity</td>
</tr>
<tr>
<td>Zedeck and Baker (1972)</td>
<td>See Different Samples of Behavior and Value</td>
<td>Convergent and Discriminant Validity of Ratings</td>
<td></td>
</tr>
<tr>
<td>(1st and 2nd line supervisors)</td>
<td>Behavior Differently</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Borman (1974)</td>
<td>Have Different “orientations” toward Ratees and See Different Samples of Behavior</td>
<td>Independent Development of Peer and Supervisor Dimensions; IR Compared Within and Across-level and on Own and Others Dimensions</td>
<td>Different Dimensions Generated; IR Higher Within-level and on Own Dimensions</td>
</tr>
<tr>
<td>Klimoski and London (1974)</td>
<td>Use Different Dimensions</td>
<td>Hierarchical Factor Analysis of All Ratings</td>
<td>Halo and Three Rating Source Factors Emerge, Along with Two Content Factors</td>
</tr>
<tr>
<td>(self-peer-supervisor)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zedeck et al. (1974)</td>
<td>Value Behavior Differently</td>
<td>Independent Development of Dimensions and Rating Behavioral Examples</td>
<td>Similar Dimensions Generated; Peers Gave Higher Effectiveness Values to Examples</td>
</tr>
<tr>
<td>(peer-supervisor)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(peer-supervisor-subordinate-self)</td>
<td>2. See Different Samples of Behavior</td>
<td>2. Comparisons of Ratings of Observed Behavior on Six Ratee Roles across Rating Sources</td>
<td>2. Differences between Sources Were Small</td>
</tr>
<tr>
<td>Pulakos and Borman (1986)</td>
<td>Use Different Implicit Performance Theories</td>
<td>Compared Factor Structures of Peer and Supervisor Ratings</td>
<td>Factor Structures Very Similar</td>
</tr>
<tr>
<td>(peer-supervisor)</td>
<td>1. Use Different Weights for Dimensions</td>
<td></td>
<td>1. IR Same for Dimension and Overall Performance</td>
</tr>
<tr>
<td>Harris and Schaubroeck</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1988) (meta-analysis:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Methodology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tsui and Ohlott (1988) (peer-supervisor-subordinate)</td>
<td>1. Different Types and Amounts of Rater Errors Across Sources</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. Use Different Dimensions</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. See Different Samples of Behavior</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Borman et al. (1995) (peer-supervisor)</td>
<td>Define Performance Differently</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pulakos et al. (1996) (peer-supervisor)</td>
<td>Define Performance Differently</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oppler et al. (in preparation) (peer-supervisor)</td>
<td>Define Performance Differently</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. Not Tested</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. Not Tested</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Policy Capturing Weights and Subjective Weights Assigned Were Compared Across Original Levels</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Compared Peer and Supervisor Rating Models with Several Characteristics of Ratees as Exogenous Variables and Ratings as the Endogenous Variable</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Peer and Supervisor Models Similar</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. IR Moderate for Dimension and Overall Performance</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. IR Same for Self-peer and Self-supervisor</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Small Policy Capturing Differences But Differences in Subjective Weights Across Levels</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Peer and Supervisor Models Had Significant Differences</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Peer and Supervisor Models Very Similar</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Interrater reliability*
in work on Project A (Campbell 1990), Pulakos and Borman (1986) factor analyzed separately supervisor and peer ratings of more than 800 first tour soldiers on 10 performance dimensions and found remarkably similar 3-factor solutions for each set of ratings, with the patterns of loadings almost identical.

Second, Oppler and Sager (1992) analyzed performance ratings of more than 35,000 army trainees, each by two peers and one drill instructor supervisor. A 4-factor model derived from the Peer 1 rating group (one of the two peers randomly selected for each ratee) was then tested for fit with the Peer 2 ratings and with the supervisor ratings using confirmatory factor analysis. Results showed a significantly better fit with the other peer ratings compared to the supervisor ratings, suggesting that the two organizational levels may use at least somewhat different implicit performance theories.

Overall, however, the available evidence does not strongly support the notion that peer and supervisor raters use significantly different dimensions or employ different implicit performance theories when rating job performance. The Zedeck et al., Oppler et al., Borman et al., and Pulakos and Borman studies, each using a different analysis approach, provide relatively strong contradictory evidence for this hypothesis. On the other hand, the Oppler and Sager, and Pulakos et al. studies suggest some support for a different-models explanation of rater disagreement. It should be noted, however, the Oppler and Sager findings might have resulted from the very high power of their design for detecting differences across sources.

Hypothesis 1, Version 2. The second version of Hypothesis 1 is that raters from different organizational levels use similar dimensions in making performance judgments, but that they weight these dimension differently. Harris and Schaubroeck (1988) tested this version by comparing the levels of peer-supervisor interrater agreement on individual dimensions and on overall performance dimensions. The reasoning was that across-level interrater agreement should be adversely affected only with the overall performance dimension because the two levels essentially form different summary composite performance judgments due to the different patterns of dimension weights employed by peers and supervisors. Their meta-analysis results showed no differences in interrater agreement for the two kinds of dimensions, evidence counter to this explanation. Similarly, results from several other studies using peer and supervisor ratings (e.g., Borman & Pulakos 1986; Borman, Mendel, Lammlein, & Rosse 1981; Borman, Toquam, & Rosse 1978) demonstrate that overall performance dimensions have higher across-level interrater reliabilities than do individual dimensions, again providing contradictory evidence for the different weights version of Hypothesis 1.

In another study, Tsui and Ohlott (1988) gathered performance ratings from the supervisors, peers, and subordinates of more than 200 manager ratees. Raters also made evaluations of ratees on six managerial roles (e.g., leader, liaison), on need for power and achievement, and on interpersonal competence.
Multiple regression results for the supervisor, peer, and subordinate rater groups were quite similar. In each of the three models, interpersonal competence and the managerial roles accounted for the most variance in the overall performance ratings. Thus, the patterns of weights were quite similar for the different organizational levels, casting doubt on the validity of the different weights hypothesis.

Similarly, Tsui (1984) gathered peer, supervisor, subordinate, and self ratings on the same six reputational roles used by Tsui and Ohlott, and on overall performance. When the overall ratings were regressed on the role effectiveness rating, no significant differences were found between peer, supervisor, and subordinate regression models, providing more support for rejection of the different-weights explanation of across-organizational level interrater disagreement.

Finally, it might be argued that the Oppler et al. (in preparation) and Borman et al. (1995) studies are relevant to the different-weights version as well as the different-dimension version of Hypothesis 1. If we consider the exogenous non-rating variables (such as ratee personality, job knowledge, and proficiency) as dimensions of performance, the findings that the peer and supervisor regression and path models of performance ratings were very similar could be interpreted as supporting the conclusion that the weights different levels' raters place on dimensions are similar. Of course, the Pulakos et al. (1996) study is not as supportive of this conclusion.

**Hypothesis 2.** This hypothesis is that raters from different organizational levels tend to disagree in their ratings because they use different samplings of ratee behavior when making performance judgments. They use different behavior samples presumably because they have different opportunities to view ratee behavior. That is, only a subset of the behavior relevant to effective performance is observable from a particular organizational level. This hypothesis has not been as well tested as the first one, with only a couple of studies at all relevant.

First, Tsui (1984) asked her supervisor, peer, and subordinate raters to indicate how much ratee behavior they observed relative to each of the six reputational roles. Results of a MANOVA, with the six mean behavior-observed ratings as the dependent variables, were significant but the effect size was quite small. At the same time, across-organizational interrater agreement was low, with the average superior-subordinate correlation about .15, the subordinate-peer correlations less than .20, and the average peer-supervisor correlation about .35. Tsui concluded that what she called the information-differences hypothesis is plausible but that it does not appear to be an important cause of low interrater agreement across organizational levels.

Harris and Schaubroeck’s (1988) meta-analysis provided an indirect test of the different-samples hypothesis. They reasoned that peers and supervisors disagree in their ratings because peers see more ratee behavior than supervisors. They asserted that if this is true, self and peer ratings should correlate more highly than either self and supervisor or supervisor and peer ratings. As
mentioned, the peer and supervisor ratings correlated much higher than did the other two combinations, a result negative to the different-samples hypothesis. Harris and Schaubroeck admit, however, that more direct tests of this hypothesis are needed. It might be pointed out that Hypothesis 2 could be true if either the amount of information each organizational level observes relative to each performance dimension is different or the content of the information related to each dimension is different. It should also be noted that a less substantive, more psychometric explanation for interrater disagreement is that raters from different organizational levels might make certain rater errors to a different degree (e.g., leniency, restriction-in-range).

Three final studies offered hypotheses about why raters from different organizational levels disagree substantially in their performance ratings, but results of the studies appear relevant to either of our main hypotheses. Berry et al. (1966) invoked essentially the different-dimensions explanation for their findings showing greater within-organizational level interrater agreement than across-level agreement. Zedeck and Baker (1972) found only moderate convergent validity in performance ratings across first and second line supervisor raters, concluding that the two levels probably observed different samples of ratee behavior (Hypothesis 2) and valued at least some work behavior differently, the latter close to the different-weights version of Hypothesis 1. And, Klimoski and London (1974) conducted a hierarchical factor analysis of peer, supervisor, and self ratings, identifying three rating source factors, along with a general factor and two content factors cutting across the sources. The emergence of strong rating source factors led Klimoski and London to conclude that each organizational level’s raters may use different dimensions in making performance judgments.

In sum, evidence for Hypothesis 1, Version 1, the different-dimensions explanation, is mixed, but overall not highly supportive of this explanation of across-organizational level interrater disagreement. Further, the explanation for interrater disagreement of raters at different levels using different weights for a similar set of dimensions (Version 2) is not well supported. Regarding Hypothesis 2, very little evidence is available to support or refute this different-samples explanation for interrater disagreement. Thus, neither of the primary explanations for across-organizational level interrater agreement has received strong support. However, it is our position that more research is needed to investigate further reasons for interrater disagreement across organizational level. Two possible research directions are discussed next.

RESEARCH ON REASONS FOR INTRERRATER AGREEMENT

Application of Personal Construct Theory Techniques. Personal Construct Theory (Adams-Webber 1979; Kelly 1955) could be used to study directly the dimensions different organizational level raters use naturally in making performance judgments. Eliciting dimensions from peers and supervisors, for example, and then comparing the content of these dimensions, both within- and
across-organizational level, would more directly evaluate this aspect of Hypothesis 1 than has been accomplished to date.

An example of how such a study might be carried out was provided by Borman (1987). He had Army officers focus on the job of NCO (i.e., first line supervisor). Each officer independently listed several effective and ineffective NCOs he or she had worked with, and then used Kelly's Repertory Grid Technique to record the performance constructs he/she believed differentiated between the effective and ineffective NCOs. After identifying and defining 6–10 constructs apiece (e.g., Firmness-Ability to control personnel and situations without falling apart), officers rated the similarity of each of their constructs to each of 49 reference constructs representing ability, personal characteristics, and performance domains (e.g., Good with Numbers, Energy Level, and Training Soldiers). The vector of similarity ratings for each performance construct then reflected a numerical definition of the construct, and these numerical definitions could be compared (correlated) across officers.

The purpose of the Borman (1987) research was to explore individual differences between officers in the constructs generated, but clearly this methodology could be used to investigate similarities and differences in the dimensions used across organizational levels as well. A specific hypothesis relevant to the different-dimensions hypothesis is that correlations between constructs within organizational level should be significantly higher than between construct correlations across organizational levels.

**Expanding Harris and Schaubroeck's Moderator Analysis.** A second suggestion is to expand the Harris and Schaubroeck meta-analysis to include two sets of moderators, each representing one of the hypotheses discussed in this paper. For example, the Hypothesis 1, different-models moderators would index for each study in the meta-analysis something like how different the roles, orientations, and perspectives of the peer and supervisor organizational levels toward the target ratees are in that organization. This could be accomplished by defining each moderator on a rating scale (possibly multiple scales) and obtaining ratings on the scales from persons knowledgeable about each organization. The moderator analysis would then proceed to test the hypothesis that different roles, (however defined to best reflect this aspect of Hypothesis 1) affects the correlations between peer and supervisor ratings.

An analogous approach could be followed with the different-samples hypothesis. A rating scale (or scales) would be developed to obtain estimates for each organization for which a peer-supervisor interrater agreement estimate is available of how closely the sampling of ratee behavior is likely to be across the two rater levels. A moderator analysis could then be conducted to assess the effects of the Hypothesis 2 variable on correlations between peer and supervisor ratings. At this point, a comparison should be made to determine the relative strength of these moderator variables to explain why the peer and supervisor rating sources disagree in the ratings to the extent they do.

It might also be illuminating to see what the correlation is between the different moderator variables. This would be important in interpreting results
of the moderator analyses. For example, there may be substantial correlation between similarity in role, orientation, and perspective and similarity in the samplings of behavior observed. Those organizations with peers and supervisors working closely together (similar samplings of behavior) may tend to have similar roles, orientations, and perspectives toward members of the ratee (peer) level. However, the main issue to be addressed with the additional moderator analyses is which of the two hypotheses better explains differences in levels of peer-supervisor interrater agreement.

The above research directions may be especially helpful for informing on theoretical and practical issues involved in understanding the reasons for and conditions under which we can expect across-level interrater agreement. The personal construct research should provide a more direct test of the different-dimensions hypothesis and the moderator study would help us understand better the factors contributing to across level interrater disagreement. However, we believe it will be most illuminating to focus research on the second assumption of 360° ratings discussed previously, that ratings from more than one organizational level enhance the validity of the performance information.

**RESEARCH TO EVALUATE THE INCREMENTAL VALIDITY OF RATINGS FROM MULTIPLE ORGANIZATIONAL LEVELS**

Plausible conceptual arguments have been made that low interrater agreement across-organizational level may not necessarily signal a lack of validity for ratings from these sources (Borman 1974; Campbell, Dunnette, Lawler, & Weick 1970). In fact, moderate across-level agreement in ratings may indicate that each level's raters are focusing on performance in somewhat different performance domains but providing quite valid ratings for the domains being evaluated. In the context of our earlier discussion regarding the two hypotheses, this outcome might be a function of either using different dimensions or models or reporting on different samplings of ratee behavior.

Accordingly, I recommend more research to follow up on the conception that ratings from more than one organizational level may increase the validity of rating information, when, in fact, there is disagreement across rating levels. Unfortunately, the research strategies to examine this issue will normally require longitudinal designs, placing considerable restrictions on the data appropriate for advancement in this area. Nonetheless, we now describe three possible strategies for testing the central assumption in 360° programs that ratings from more than one organizational level may be more valid than ratings from a single level.

**Correlating Performance Ratings Concurrently with Relevant External Criteria.** First, in the perhaps unusual cases where there are external criteria against which to evaluate the validity of ratings from multiple perspectives, a concurrent validation strategy may be possible. In such cases, peer and supervisor
ratings can each be correlated separately with the criterion and then pooled to assess the incremental validity of the combined ratings over ratings from each source.

Two articles seldom referred to in the performance rating literature provide some empirical evidence and additional theoretical rationale for a model that argues for evaluations from two or more different perspectives yielding more valid judgment than evaluations from only a single source.

Buckner's (1959) study is an example of this approach. He had two different levels of supervisors evaluate the performance of U.S. Navy submariner enlisted men. Buckner found that in cases where interrater agreement across the two organizational levels was relatively low, a composite of the ratings correlated higher with scores on objective criteria (argued to be good indicators of actual performance) than in cases where interrater agreement was higher.

A second study did not employ multiple organizational level raters, but still provides an example of how evaluators from different perspectives might contribute ratings that when taken together are more valid than any single evaluator. Einhorn (1972) had four medical experts with somewhat different specific areas of expertise predict the survival time of persons who had contracted Hodgkin's Disease. Each made ratings on several components related to the severity of the disease, as well as an overall prediction of survival time. Multiple regression results using both sets of ratings showed that the validity against actual survival time was greatest when all four judges' data were used, even allowing for the lower reliability of individuals judges' ratings. These findings seemed to be in part a function of different judges providing valid information on different components. This is analogous to raters at different organizational levels each contributing valid performance information, but on somewhat different aspects of the job. It would be desirable to see more studies like Buckner's and Einhorn's in organizational settings with raters from multiple organizational levels and independently derived indexes of overall performance.

**Predicting Subsequent Job Performance from Ratings in Training.** A second possible setting for the proposed research direction is in industrial or military training environments. In cases where peer and instructor evaluations of potential, likely subsequent job performance, or some similar prediction of future performance are available, correlating these ratings with later overall job performance should address the same hypothesis. In this setting we would predict that the peer and instructor ratings, taken together (i.e., standardized within source and then summed), will provide higher validity against subsequent performance than either of the rating sources itself.

Past research relevant to this hypothesis has focused on the usefulness of peer ratings in training to predict important criteria in military environments. Kane and Lawler (1978) reviewed all of the pre-1978 research in this area. For example, Williams and Leavitt (1947) found that peer ratings of leadership taken in Marine Corps Officer Candidate School (OCS) were better predictors of success in OCS and combat performance than either instructor ratings in
training or objective test scores. The predictive validity of peer and supervisor ratings taken together was not examined.

Similarly, Hollander (1954) gathered leadership peer nominations and instructor ratings of cadets in U.S. Navy pre-flight training and showed that the peer nominations correlated higher \( r = .27 \) with graduation from flight training 14 months later than did the instructor ratings \( r = .18 \). In this case, it was possible to compute a multiple R from data contained in the article, and the results do not support our hypothesis \( R = .27 \). However, it is not clear how much contact relevant to this criterion the instructors had with students in this setting. Also, there is some question for our purposes about the similarity of the leadership rating construct and the construct(s) reflected in the training pass-fail criterion.

**Predicting Subsequent Performance Levels with Ratings of Potential Taken on the Job.** Finally, perhaps the most relevant set of data for testing, within a predictive validity framework, the hypothesis that pooled peer-supervisor ratings will be more valid than ratings from either source alone is one in which peer and supervisor "predictor" ratings are based on job performance and these rating dimensions represent constructs conceptually matched with the subsequently gathered criteria.

One type of setting where such data may be available is when ratings of potential or promotability are made by peer and supervisor raters. Of course, such evaluations are usually generated by supervisors only, but if an organization did collect potential or promotability ratings from peers as well, it would be possible to correlate the ratings against later progress in the organization or performance at higher organizational levels. The hypothesis would be the same, the pooled ratings across organizational levels will be more valid than ratings from either source individually. Similar to assessment center validation research to predict future performance, it will be important in such a study that criterion contamination can be ruled out as an explanation for the results.

**CONCLUSIONS**

I have argued here that the assumptions underlying the use of multiple source ratings in 360° programs should be examined and evaluated empirically. One of the assumptions is that ratings from different organizational levels provide different, relatively unique perspectives. Research supporting this assumption is that intrarater agreement within organizational level has generally been found to be higher than across level. We reviewed research on two hypotheses to explain this result, the different performance models and different samples of behavior hypotheses. Results do not strongly support either hypothesis. This is unfortunate because learning why members of different organizational levels tend to disagree in their ratings would help us to interpret 360° ratings. Two research approaches to making progress in this area were offered.

On balance, practitioners of 360° programs should not be alarmed by low to
moderate agreement across rating sources. In my judgment, when possible, within-level interrater agreement should be evaluated (e.g., for peers, subordinates, or customers). High within-source agreement suggests that the rating source is providing a coherent, valid (from their perspective) view of performance. Practitioners might also keep in mind that within-source agreement can be increased by obtaining larger numbers of raters.

A second assumption in 360° programs is that each organizational levels' ratings provide incremental validity to performance assessment. Regarding this assumption, I argued that it is time to resurrect the hypothesis that moderate or even low interrater agreement across-organization level may reflect different levels' raters making reasonably valid performance judgments but on partially different aspects of performance (Borman 1974; Campbell et al. 1970). It was noted that this hypothesis implies nothing about the hypotheses related to reasons for across organization level interrater disagreement. Rendering potentially valid performance judgments on partially different elements of the job could be a function of either different level raters using different performance models, their seeing different samples of behavior, or some combination thereof. Nonetheless, three strategies were discussed for testing the hypothesis that multiple source ratings taken together may be more valid than ratings from a single source alone.

If disagreement across rater levels exists and the incremental validity hypothesis is confirmed, support is gained for employing 360° assessment programs. Certainly, results demonstrating incremental validity for adding ratings from additional rating levels and sources (e.g., subordinates, customers) would provide much needed empirical support for the administrative and financial costs associated with these types of programs. As a final point, it should be noted that although the research reviewed and evaluated in this article focused on peer and supervisor rating sources, the same conceptual arguments and research directions are appropriate for studying other potential ratings sources as well (e.g., self, subordinates, customers).

ACKNOWLEDGMENT

The author wishes to thank Anne Tsui, Michael Harris, Chris Sager, Mary Ann Hanson, and Elaine Pulakos for their comments on an earlier version of this article.

REFERENCES


sented at the 100th Annual Convention of the American Psychological Associa-
tion, Washington, DC.

ences of Ratee Characteristics on Peer and Supervisor Ratings.”

716). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social
Sciences.

Examination of Ratee Race, Ratee Gender, and Rater Level Effects.” Human
Performance 9: 103–119.


Tsui, A. S. 1984. “A Role Set Analysis of Managerial Reputation.” Organizational Be-
havior and Human Performance 34: 64–96.

Tsui, A. S. and P. Ohlott. 1988. “Multiple Assessment of Managerial Effectiveness:
Interrater Agreement and Consensus in Effectiveness Models.” Personnel Psy-


Expectation Scales: A Multitrait-Multirater Analysis.” Organizational Behavior
and Human Performance 7: 457–466.

Anchored Rating Scales as a Function of Organizational Level.” Journal of Ap-
plied Psychology 59: 249–252.