Gender Composition, Situational Strength, and Team Decision-Making Accuracy: A Criterion Decomposition Approach

Jeffrey A. LePine
Department of Management, University of Florida

John R. Hollenbeck
Department of Management, Michigan State University

Daniel R. Ilgen
Michigan State University

Jason A. Colquitt
Department of Management, University of Florida

and

Aleksander Ellis
Michigan State University

This study reexamines conclusions regarding the superiority of all male teams on traditionally masculine tasks. By decomposing the criterion of decision-making accuracy, we illustrate how male-dominated teams may, in some contexts, constitute the worst gender composition. Specifically, as the percentage of males on a team increased, there was an exponential increase in the tendency for making decisions that were overaggressive.

An earlier version of this article was presented at the 1999 Annual Meeting of the Society for Industrial and Organizational Psychology. The authors thank Lori Sheppard, Christine Jackson, Amy Pepper, and Henry Moon for their help with collecting data. This work was supported, in part, by grants from the Office of Naval Research (N00014-93-0983) and the Air Force Office of Scientific Research (F49620-95-1-0314). This support does not imply agreement or disagreement by the supporting agencies of the ideas expressed herein.

Address correspondence and reprint requests to Jeffrey A. LePine, Department of Management, 201 Stuzin Hall, P.O. Box 117165, University of Florida, Gainesville, FL 32611-7165. Fax: (352) 392-6020. E-mail: lepineja@notes.cba.ufl.edu.
However, this bias in male-dominated teams’ decision making was neutralized when teams received feedback on past levels of overaggressiveness. We document these effects in a study of 80 four-person teams using a simulation of a task that is masculine in its characteristics and male-dominated in terms of occupational demography. Possible interventions aimed at improving team decision-making accuracy are suggested.

Two recent trends in organizations include an increased feminization of the work force and a reliance on team-based work structures. Howard (1995) addressed the first, noting the increase in work force participation rates for women across the economy as a whole as well as within specific occupations. In some career fields once exclusively male, women are beginning to constitute the majority of those entering professional schools, including veterinary science, pharmacy, law, medicine, and business. Ilgen (1999) addressed the second of these trends, the shift from structuring work around individual jobs to organizing work around teams. The increased reliance on team-based structures has prompted some to conclude that the entire notion of the individual job is a social artifact that is beginning to lose relevance in organizations (Bridges, 1994). Although this loss of relevance is, in our opinion exaggerated, changes in work force composition and the nature of work suggest that the number of mixed-gender teams in organizations is likely to increase.

Although the popular press extols the performance advantages of work group diversity, this sanguine outlook has not been supported by empirical research on gender composition (Argote & McGrath, 1993). In her meta-analysis, Wood (1987) failed to find any advantages for teams with mixed-gender composition. In fact, all-male teams outperformed any other gender configuration. Wood attributed part of this to the fact that the majority of the studies contributing to this meta-analysis used tasks that were traditionally male-dominated. This is hardly a comforting fact to those concerned with integrating women into traditionally male-dominated jobs or occupations.

Historically, Wood’s meta-analysis appears to have had a dampening effect on research involving gender composition and group-level performance. Although some have examined how gender composition affects individual group members (e.g., Ely, 1995; Mehra, Kilduff, & Brass, 1998), little attention has focused on gender composition as an independent variable affecting group performance (Rogelberg & Rumery, 1996). In the absence of new research addressing these effects, past conclusions regarding the superiority of all-male groups on traditionally male tasks persists.

There remain, however, questions regarding Wood’s conclusions about the superiority of all-male groups on tasks that have been traditionally male-dominated. First, roughly half of the research contributing to Wood’s meta-analysis was conducted prior to 1970, thus predating the sexual revolution believed to be largely responsible for the currently increasing integration of the work force (Eagly, 1995). While the notion of gender roles persists, a much broader acceptance of women occupying a variety of work roles now exists. The
degree to which these past findings generalize to current organizations is open to question.

Second, prior research has considered the nature of the task context primarily in terms of male or female dominance. Research shows that the strength of the task situation often has a major role in neutralizing effects of individual differences like gender or personality (Mischel, 1977; Weiss & Adler, 1984). In strong situations, or those characterized by formal rules, standard operating procedures, unambiguous feedback, or meaningful incentives, individuals construe and act on information in similar ways, thus neutralizing the effects of between person variability. Because many work contexts can be described as ‘‘strong situations,’’ the need to test the generalizability of gender composition effects across situations of varied strength is also critical.

Finally, prior research examining group performance has examined overall performance differences using global criteria. Campbell (1990) has argued that performance is not a meaningful unidimensional construct and, instead, means different things to different people in different work contexts. In order to further the science of performance prediction and enhancement, researchers in organizational behavior need to develop more refined conceptualizations of the performance construct. That is, rather than developing theories of ‘‘performance’’ per se, many have argued for ‘‘criterion decomposition,’’ whereby researchers develop and test theories of more specific, unidimensional components of performance (Motowidlo & Van Scotter, 1994).

Within the context of work teams, criterion decomposition means developing more specific theories about what precise aspects of team performance may be enhanced or hindered by various types of gender composition, as opposed to trying to answer questions involving gross, multidimensional measures of performance. Although practitioners may demand answers in terms of broader concepts like performance, applied science often cumulates more readily by relying on more narrow, unidimensional, and homogeneous dependent variables (Nunnally, 1978) and using a divide-and-conquer approach (Simon, 1992). Theories of these more narrow constructs can then be integrated at a later point to address the more general issues of interest to practitioners.

The primary purpose of the present research is to use a criterion decomposition approach to reexamine the relationship between gender composition and group performance in a traditionally male task under varying levels of situational strength. We show that there are good reasons to suggest that, for some specific aspects of group performance, all-male teams may constitute the worst configuration of gender composition, especially in contexts characterized by low situation strength. Previous research suggests that male-dominated teams are superior on tasks that favor males’ interests and abilities, therefore we take the most conservative approach and examine the negative implications of all-male teams using a task that is masculine in its characteristics and male-dominated in terms of occupational demography. The task employed in this study is a simulation of a real-world military task where women comprise less than 10% of the work force (Jowers, 1995). We note that the use of a male-dominated task is critical in that this is precisely the type of task that is likely
to see an increase in female participation in the future. We also illustrate the usefulness of criterion decomposition in terms of identifying sets of nonredundant predictors generated by two separate theories (Social Role Theory and the Multilevel Theory of team decision-making).

Criterion Decomposition and Team Decision-Making Accuracy

The relative importance of specific performance dimensions varies across tasks. Thus, the exact nature of the task needs to be specified when trying to understand and predict performance differences between teams composed in different ways. The current research is primarily interested in performance in information processing tasks. We focus on information processing for two reasons. First, technological changes in the nature of work have increased the role of cognitive factors in performance relative to physical strength and endurance (Howard, 1995). Across a wide variety of contexts, the ability to process information and make good decisions is generally more important than executing physical tasks. Second, when jobs demand physical strength and endurance, gender differences tend to be large. Hence, the argument for gender integration is more difficult to make (Halpern, 1986).

Within the realm of cognitive tasks, one aspect of performance that exists in almost every job is the need to make good judgments and decisions. Virtually all job analytic schemes consider this aspect of the work when characterizing a job [e.g., the “Making Decisions” section of McCormick’s Position Analysis Questionnaire (McCormick, Jeanneret, & Mecham, 1972) or the “Reasoning Development” scale in the Job Element Method (Primoff, 1975)]. Thus, the central criterion in this study, rather than global performance in some broader sense, is decision-making at the team level.

Team decision-making criteria: Accuracy versus consensus. Ilgen, Major, Hollenbeck, and Sego (1995) noted that team decision-making criteria could be grouped into two types. Internally referenced criteria are characterizations of the feelings and actions of members that occur in the process of reaching an outcome. Information flow, availability, and use (e.g., Argote, Seabright, & Dyer, 1986; Stasser & Titus, 1985) are examples of internally referenced decision-making criteria. The most common internal decision-making criterion is consensus. In the jury decision-making literature, for example, scholars describe a consensus rule and then examine the extent to which the rule is used and also the factors that predict its use (e.g., Stasser, Kerr, & Bray, 1982; Stasser, Kerr, & Davis, 1989). The degree of group consensus, however, says nothing about the quality of the decision with respect to meeting some organizationally valued standard (Janis, 1972, 1982). There can be high consensus about very poor choices.

Externally referenced criteria focus on the quality of decisions with respect to some external organizationally valued standard for evaluation. A decision need not be correct or true in some absolute sense, but there must be one objectively verifiable correct (true) decision against which an actual decision can be compared. Team decision accuracy, therefore, can be captured by the
comparison of true decisions and teams' decisions. For teams of doctors, true scores might be the actual severity of their patients' pathologies. Teams' decisions would then be their actual diagnoses with respect to the severity of their patients' pathologies. In this example, the difference between the true scores and the teams' decisions is a useful index of decision-making accuracy because as diagnoses become more incorrect, recommended treatments becomes less appropriate. For many obvious reasons, the organizations in which these teams of doctors work value decisions that result in the most appropriate treatment for patients.

Decomposing overall decision-making accuracy. Although focusing on decision-making accuracy rather than global performance represents one level of criterion decomposition, researchers have suggested that the criterion of decision-making accuracy can also be meaningfully decomposed into independent elements that have different psychological implications. Gigone and Hastie (1997) argued that team decision-making accuracy, although traditionally treated as a unidimensional construct, is best conceptualized as a multidimensional construct. Moreover, the three distinct dimensions of accuracy—linear consistency, mean bias, and variability bias, discussed further below—may be only weakly related to each other and are likely to be influenced by different antecedents.

Traditionally, team decision-making accuracy has been defined as Mean Absolute Error (MAE) or the absolute difference between a team's decision or judgment and the "true score" (Gigone & Hastie, 1997). However, this definition ignores the distinction between correlational and absolute accuracy. Gigone and Hastie argued that Mean Squared Error (i.e., MSE; the squared difference between the team's judgment and the "true score") is a superior measure of overall decision accuracy for two reasons. First, MSE gives more weight to extreme errors relative to MAE. Second, and perhaps more importantly, MSE is a multidimensional construct of the additive form where it exists at the same level as its dimensions and is formed as a mathematical function of those dimensions (Law, Wong, & Mobley, 1998). Accordingly, relationships between the composite and the dimensions can be completely independent of the structure of the relationships between the dimensions. This implies that MSE can be decomposed mathematically into dimensions that are uncorrelated. This characteristic of MSE makes it an excellent tool for building and integrating theories. By considering the dimensions of the overall criterion in isolation, one can use separate theories to identify sets of nonredundant predictors and, as a result, increase one's ability to predict and explain variance in the global outcome.

Equation 1 shows how MSE decomposes naturally into its three terms, each of which describes a potentially important and conceptually distinct aspect of decision-making error:

\[
MSE = (M_j - M_c)^2 + (S_j - S_c)^2 + 2(1 - r_{xy})S_jS_c, \quad (1)
\]

In Eq. (1), \(M_j\) is the mean of the judgments, \(M_c\) is the mean of the true scores.
or the criterion, \( S_j \) is the standard deviation of the judgments, and \( S_c \) is the standard deviation of the criterion. The first term in the equation, \((M_j - M_c)^2\), reflects mean bias, or the amount of systematic distance error in the judgments either greater than or less than the mean of the criteria. The second term in the equation, \((S_j - S_c)^2\), reflects variability bias. This term captures the degree to which the variability of judgments made by the team is greater or less than that associated with the true scores. Finally, the third term in this equation captures linear consistency \((r_{xy})\), which is the linear relationship or correlation between the judgments and the true scores. Taken together, the three dimensions of accuracy imply that a perfect set of decisions accomplishes three distinct goals: (a) a perfect linear association between the decisions and the true scores, (b) equal means for the decisions and true scores, and (c) equal variability for the decisions and the true scores.

Research on accuracy decomposition. The importance of decomposing accuracy into its constituent parts was shown recently in a study by Hollenbeck and his colleagues (Hollenbeck, Colquitt, Ilgen, LePine, & Hedlund, 1998). In this study, the predictive effects of the variables specified by one theory of team decision making were almost entirely attributable to the impact that the variables had on just one aspect of accuracy. That is, the core constructs of the Multilevel Theory of team decision making (Hollenbeck, Ilgen, Sego, Hedlund, Major, & Phillips, 1995) were found to explain a significant amount of variance in overall decision-making accuracy operationalized as mean squared error (29%), but this was primarily due to the theory's ability to account for variance in linear consistency (52%). One of the variables specified by this theory explained a moderate amount of variance in variability bias (9%), but virtually no variance in mean bias was accounted for. Given that as much as one-third of the variance in overall team decision-making accuracy is attributable to mean bias, the fact that none of the variables specified by the Multilevel Theory impacted this aspect of the criterion suggests a need to develop predictors of mean bias. In fact, because research has never broken down overall accuracy into its component parts, there is not a single documented predictor of mean bias in the decision-making literature.

Hollenbeck et al. (1998) called for more research on the antecedents of mean bias and noted that "because this type of bias reflects a consistent response bias, factors related to cross-situational consistency (i.e., individual differences) might be predictive of this type of bias" (p. 500). This suggestion is highly relevant to the study of gender composition in decision-making teams because members' gender is one type of individual difference that might be theoretically linked to the type of error reflected in mean bias.

Gender Composition, Social Role Theory, and Mean Bias

Social Role Theory. Social Role Theory stands as one of the major theoretical frameworks for understanding the origins of gender differences (Eagly, 1987). Social Role Theory starts with the notion that men and women have been
traditionally segregated into different roles, both in the family and in the workplace, and that these roles are directly relevant to gender differences in adult social behavior. Social Role Theory is a structural theory in the sense that it emphasizes that members of social groups experience common situational constraints and experiences because they tend to be placed in similar social positions within organizations and families.

The sets of expectations that differ for men relative to women are classified by Social Role Theory into two categories. According to the theory, women are ascribed communal characteristics, whereas men are assigned agentic characteristics. The communal content of women's roles is derived primarily from the domestic role and emphasizes what Bales (1950) referred to as expressive behavior, implying that women, relative to men, are more caring, nurturing, emotionally expressive, selfless, and more likely to seek harmony in interpersonal relationships. The agentic content of the male stereotype is derived from men's typical role as provider and emphasized what Bales referred to as instrumental behaviors. This distinction implies that men, relative to women, are more assertive, controlling, aggressive, independent, adventurous, and competitive.

According to Social Role Theory, these sets of expectations are reinforced by differences in skills and abilities that eventually develop as these expectations are put into practice. In this fashion, initial differences that are expectation-based are solidified by selective exposure to differential skill-building opportunities. These differences in expectations and skills only change when the social arrangements that give rise to them change. Although, as we noted above, there has been a major shift of women into the paid labor force, this has yet to spill over into substantially reduced magnitudes regarding beliefs about gender differences (Deaux & Lewis, 1984; Eagly, 1995; Feingold, 1994; Heilman, Block, Martell, & Simon, 1989).

Mean bias and gender composition. Social Role Theory's assertion that men are more assertive, risk-seeking, competitive, and aggressive has been documented in numerous empirical studies and meta-analyses (Eagley & Steffen, 1986; Feingold, 1994; Bauer & Turner, 1974; Coet & McDermott, 1979; DiBerardinis, Ramage, & Leavitt, 1984; Seeborg, Lafollette, & Belohlav, 1980). Although this research does not directly link gender to decision accuracy in general, or mean bias specifically, it does have implications for these criteria. Specifically, to the extent that an individual's assertiveness, risk seeking, competitiveness, or aggressiveness can be seen as having implications for responses, men are more likely than women to manifest extreme judgments in the direction that reflects these characteristics.

To some extent, this expectation might seem counterintuitive in masculine contexts because in these types of situations, aggressive decisions would intuitively seem to be more appropriate. Indeed, we agree that there may be situations where there is an apparent requirement for aggressive decisions because the likelihood of higher M's is greater (e.g., “purchase risky stocks” recommendations in a bull market). However, the same task context is just as likely to
have a requirement for more passive decisions in another situation because the likelihood of lower M_c's is greater (e.g., “purchase money market certificates” recommendations in a bear market). In both situations, the overall task is the same because making correct decisions requires analysts to apply the same cues (e.g., present market conditions, forecasts, and asset characteristics) onto the same decision rule. In both of these situations the act of making a decision is neutral with respect to aggressiveness because the underlying task simply requires the decision maker to consider cue values and then apply a mathematically based underlying decision rule.

This is not to say that decision makers are completely rational, however. One reason why is because decision makers do not typically have a perfect understanding of underlying decision rules (if they did, they would always make perfect decisions). What we are saying is that if there is a choice between two alternative decisions after a decision maker applies cue values to an imperfectly understood decision rule, males will be more likely than females to choose the more aggressive alternative. After a series of decisions that vary with respect to the level of required aggressiveness, this tendency should result in greater overall mean bias in the aggressive direction for males. Moreover, the possibility that males will be accurate and females will be underaggressive is less likely given a masculine task. This is because participants (both male and female) are likely to believe that aggressiveness is consistent with the demands of the stereotypically masculine task, and therefore, neither group is likely to be underaggressive in an absolute sense (even though they could be underaggressive).

In a team setting, it is likely that these extreme (aggressive) individual-level judgments will translate into more extreme (aggressive) team-level decisions, at least in proportion to the number of males in the team who tend to make such judgments. This expectation is supported indirectly by Eagly and Wood (1991), who suggest that tasks that are perceived to emphasize one gender’s skills or characteristics over the other can intensify gender differences in intragroup behavior.

In terms of the three components of decision-making accuracy, this directional, cross-situational consistency would be most likely to impact mean bias. First, the cross-situational nature of the bias should be less likely to manifest itself in terms of linear consistency. Linear consistency is a function of decision-to-decision variability, and there is nothing inherent in Social Role Theory that would suggest that gender differences should influence this type of variability. Second, the directional nature of the bias would not likely manifest itself in terms of variability bias. Variability bias reflects a bidirectional error, and Social Role Theory says nothing about how gender differences might influence this type of error. Finally, and as mentioned above, mean bias is theoretically independent of linear consistency and variability bias. For example, if the true scores on a set of decisions were 1, 3, 5, and 7, a team would have perfect linear consistency and variability bias with a set of decisions that included 2, 4, 6, and 8; however, mean bias would equal 1.

Thus, based upon the predictions of Social Role Theory and an analysis
employing Gigone and Hastie’s (1997) three-dimensional decomposition of decision accuracy, our general expectation is that gender composition will be related to overall decision-making effectiveness. Furthermore, we expect that this effect will be primarily attributable to the effects that gender composition has on the mean bias aspect of decision-making effectiveness. Finally, and as outlined above, we are expecting a directional mean bias effect ($M_j - M_c$), not the “nondirectional” form originally suggested by Gigone and Hastie ($M_j - M_c)^2$. More specifically, we propose the following:

Hypothesis 1: There will be a significant and positive relationship between the proportion of men in a decision-making group and mean bias in the aggressive direction.

Situation strength as a boundary condition. Because Social Role Theory’s predictions are based on the effects of individual differences, the strength of the situational context becomes a highly relevant factor when testing this theory. Specifically, as Mischel (1977) and Weiss and Adler (1984) have argued, strong situations can neutralize the effects of individual differences by standardizing the expectations, assessments, and interpretations of outcomes of different people. Strong situations are described as those that generate a common perception and motivation among widely varying people and are created by such factors as formal rules and procedures, clear instructions, unambiguous feedback, or powerful incentives tied to specific behaviors. Situational strength, therefore, can be operationalized in many different ways. In one study, for example, Meier (1970) found that gender differences in leadership styles emerged during a discussion task only when the group’s task was unstructured (i.e., when a clear solution to a problem was not prescribed). Meier argued that high levels of task structure created a strong situation by reducing uncertainty about what participants had to do to be successful, and this suppressed the emergence of gender-related differences in style.

In a decision-making context where the effects of male-dominated team gender composition are at issue, an appropriate manipulation of situational strength would be feedback expressed in terms specifically related to the nature of the decision-making errors likely to occur in male-dominated decision-making teams. Specifically, because we are concerned that male-dominated teams are likely to be overaggressive and manifest mean bias in their decision-making, clear feedback indicating the presence of this specific problem in past decisions should provide the information necessary to neutralize the effect.

The choice of feedback as a manipulation of situational strength is consistent with several current trends in organizations that use teams. Many of today’s most successful work initiatives require teams to receive constant, specific feedback on various dimensions of job performance. For example, Total Quality Management (e.g., Jablonski, 1991) has feedback as one of its central tenets, as teams are given frequent information on quality and efficiency. Initiatives like Gainsharing (e.g., Hatcher & Ross, 1991) and Open Book Management (e.g., Case, 1997) use frequent feedback to show teams where they stand in relation to productivity or budgetary goals. It is not uncommon to see charts and graphs of daily performance taped to team members’ work areas and on
bulletin boards in cafeterias. This type of feedback provides a strong situation because it provides clear information about where teams need to focus their effort in order to be effective. This type of feedback, therefore, neutralizes tendencies for certain types of teams to systematically focus effort in areas that have less value to the organization.

Establishing the moderating effect for situational strength has important theoretical and applied implications. Although the failure to find gender composition effects in strong situations does not necessarily refute Social Role Theory, boundaries for the predicted effects do become established. As a result, more realistic expectations are created for where and when gender composition is likely to matter in real-world contexts. Establishing the moderating role of situational strength is also important because it suggests how to potentially reduce the negative effects of suboptimal gender configurations. Thus, our second hypothesis is as follows:

Hypothesis 2: Feedback on past levels of mean bias will moderate the relationship between gender composition and mean bias, such that gender composition will have stronger effects in the absence of feedback, and weaker effects in the presence of feedback.

Integrating Social Role Theory and the Multilevel Theory

As we noted at the outset, one of the values of criterion decomposition is that, by isolating different dimensions of an overall criterion, one can use separate theories to predict and explain different aspects of this criterion. Criterion decomposition, therefore, can be a powerful vehicle for meaningfully integrating theories and explaining variance in complex phenomena. In this next section of the article, we illustrate the value of this approach.

We have already suggested that Social Role Theory can be used to derive predictors of the mean bias portion of overall team decision-making accuracy (i.e., gender composition and situational strength expressed in terms of mean bias feedback). Past research, on the other hand, has shown that the Multilevel Theory of team decision-making can be used to derive predictors of another aspect of overall team decision-making accuracy—linear consistency. Because these two theories predict different dimensions of decision-making accuracy, considering them together should result in an increase in the ability to explain variance in overall team decision-making accuracy (relative to the variance either theory could have explained in isolation).

The Multilevel Theory. The Multilevel Theory of team decision-making uses a Brunswick Lens model for describing the team decision-making task (Ilgen et al., 1995). The theory identifies a core set of outcomes at multiple levels of analysis (i.e., the decision level, the individual level and the dyadic level) that need to be achieved in order for decision-making teams to make accurate decisions. According to the MLT, the set of core constructs mediates the relationship between the traditional variables that have been studied in past research on group decision making (i.e., cohesiveness, experience, and interdependence) and decision-making accuracy.

At the decision level, the most critical variable according to the MLT is
decision informity, defined as the degree to which each member of the team had all the information necessary to make a decision from a given role. Decision informity can be aggregated to a team level construct, referred to as team informity (TI), that captures how well informed a team is, on average, across all the decisions it makes.

At the individual level, the most critical variable is individual validity, the degree to which any one subordinate (i.e., staff member) can generate recommendations to the leader that are predictive of the “correct” decision for the team. In teams with multiple staff members, this variable can be aggregated to a team-level construct referred to as staff validity (SV). Staff validity is the degree to which the team has staff members whose judgments are predictive of the “true score” of the decision object. A team with high staff validity generates judgments about the decision object that predicts the true state of that object. A team with low staff validity has staff members whose judgments fail to predict the true state of that object. While staffs do not make final team decisions, one key to team effectiveness is the degree to which the leader (who makes the final decisions) can use staff judgments to arrive at accurate decisions. According to the MLT, as long as staff judgments are predictive of the true score, it does not matter that they are biased because leaders can theoretically correct for the bias. That is, staff members might be making terrible decisions as individuals because they are biased (either aggressively or passively), yet as long as the decisions are predictive of the correct decisions (i.e., high staff validity), they may be useful in terms of promoting effective team decisions.

Finally, at the dyadic level, the most important variable is dyadic sensitivity. This variable reflects the degree to which the team leader correctly weighs each staff member’s recommendation to arrive at the team’s decision. Weights can be identified using policy-capturing techniques, and in teams with multiple staff members, dyadic sensitivity can be aggregated to a team-level variable called hierarchical sensitivity (HS). Hierarchical sensitivity captures the overall optimality of the leader’s use of his or her staff’s recommendations. Effectiveness with respect to hierarchical sensitivity implies that the team has a leader who uses the best possible weight for each staff member’s judgment when combining these judgments to arrive at the team’s decision.

MLT constructs and linear consistency. Several empirical studies have supported the predictive value of the MLT (Hollenbeck et al., 1995, Hedlund, Ilgen, & Hollenbeck, 1998; Hollenbeck, Ilgen, LePine, Colquitt, & Hedlund, 1998) when the criterion is overall decision-making accuracy. However, research has shown that the effects of the three core variables on overall decision-making accuracy are largely attributable to their impact on linear consistency (Hollenbeck, Colquitt, et al., 1998). Here, we offer hypotheses that replicate these findings. First, because decisions often require consideration of multiple pieces of nonredundant information, obtaining a larger pool of relevant information should translate into team decisions that are more predictive of correct decisions. Thus we propose the following:
Hypothesis 3: There will be a significant positive relationship between team informity and linear consistency.

Second, because leaders rely on the judgments of their staffs when making final decisions for their team, linear consistency should be higher for teams with staffs that tend to make judgments that are more valid.

Hypothesis 4: There will be a significant positive relationship between staff validity and linear consistency.

Finally, by applying an appropriate set of weights to staff member recommendations, leaders effectively leverage knowledge regarding the validity of staff member recommendations in order to make valid final decisions. Given that lower levels of hierarchical sensitivity reflect more optimal weighting strategies:

Hypothesis 5: There will be a significant negative relationship between hierarchical sensitivity and linear consistency.

Although there are reasons to expect the core constructs to be related to linear consistency, there is less reason to expect the core constructs to be related to mean bias and variability bias. This is because teams can be highly effective on the core MLT constructs while still having substantial levels of both forms of mean and variability bias (Hollenbeck et al., 1998). For example, on a 1–7 scale, the correct decisions for a set of decisions might include only 3, 4, and 5, but a team with perfect levels of staff validity and hierarchical sensitivity could accomplish this with a set of decisions that included the values 5, 6, and 7. In this example, there would be perfect linear consistency between the team’s decisions and the correct decisions. Moreover, the team’s decisions and the correct decisions would have the exact same variability (range = 2), yet there would still be mean bias (the mean of the correct decisions is 4.00, while the mean of the team’s decisions is 6.00). As another example using the same 1–7 scale and correct decisions that include only 3, 4, and 5, a team with perfect levels of staff validity and hierarchical sensitivity could accomplish this with a set of decisions that included only the values 1, 4, and 7. In this example, there would be perfect linear consistency between the team’s decisions and the correct decisions, and these scores could yield the same mean (4.00). However, there would be a significant level of variability bias (the range of the correct decisions is 2.00, whereas the range of the team’s decisions is 6.00).

Integrating Social Role Theory with the MLT. Although the use of overall decision-making accuracy may be problematic because it masks effects on individual components of decision accuracy (mean bias, variability bias, and linear consistency), it is a popular criterion in the decision-making literature. With this in mind, and in order to illustrate the value of an approach to research that uses theories targeted specifically at the individual components of a multi-dimensional criterion, we examine relationships with overall decision-making accuracy. Specifically, we examine the extent to which the core MLT constructs (Team Informity, Staff Validity, and Hierarchical Sensitivity) account for variance in overall decision-making accuracy (MSE) over and above the variance
attributed to Social Role Theory variables (gender composition and feedback on mean bias).

METHOD

Participant

In return for their participation in our 3-h experiment, 320 undergraduate management students from a large state university received course credit. These students were in either their junior or senior year at the university, and on average, they were about 21 years old. These individuals were arrayed into 80 4-person teams. We note that the participants were also the subjects in another study that focused on contrasting two forms of contextual performance (LePine & Van Dyne, 2001). This other study used an entirely different set of variables and was focused solely at the individual level of theory and analysis. In order to help ensure that participants would be actively involved in the experiment, and also to give team members a common team-level goal, individuals were eligible to earn cash prizes based on their team’s performance. Individuals in top performing teams in each condition received $20, individuals in second place teams received $15, and those in third place teams received $10.

Task

In this study, we used a computerized decision-making task called Team Interactive Decision Exercise for Teams Incorporating Distributed Expertise (TIDE², see Hollenbeck et al., 1995, pp. 301–303 for more details). TIDE² is a flexible computer program that can be used to simulate any decision-making task where the research participant is given quantitative information on multiple attributes and has to combine this information to come up with a decision about what action to take based on this information. The program can simulate decision-making in many different contexts (e.g., personnel selection where the attributes might be scores on various measures of personality, investment decisions where the attributes are various measures of stock performance, or medical decisions where the attributes are values on certain diagnostic tests). TIDE² presents participants with values on attributes associated with a series of decision objects or problems.

We worked with subject matter experts within the U.S. Air Force to configure the program to simulate an airborne, military command and control team. We chose this specific configuration so that the task would deal with information processing at its core, but on the surface would appear masculine in nature. In real-world contexts, command and control teams perform an operational function (monitoring an assigned airspace and using rules of engagement to make classification decisions about unidentified aircraft that enter the airspace) often directly over or near hostile territory. Judgments made about unidentified aircraft can be graded in how aggressive a stance to take (ignore, monitor, engage, etc.) and are based on quantitative measurements associated
with the aircraft (speed, range, altitude, heading, etc.). Although women can perform this function in noncombat zones, command and control teams have traditionally been male-dominated. Members of these teams often have to make quick judgments and be very confident and assertive in communicating these judgments so that they are not discounted by final decision makers (e.g., those who must weigh staff member recommendations and then carry out air engagements or ground attacks). Communal characteristics such as nurturing, caring, interpersonal sensitivity, and emotional expressiveness are inconsistent with these requirements. Thus, overall, it is reasonable to suggest that the command and control task appears to favor agentic characteristics. That is, the task appears to be more consistent with the traditional male role than with the traditional female role.

In our simulation, each team member had a specific role and sat in front of a computer terminal that was networked to the other members, who were located in the same small room in a nearby cubicle. Team members met and interacted with one another in a waiting room prior to the experiment and were fully aware of which individual played which role in the experiment. Thus, while all communications transpired over a computerized network, team members were fully aware of each other’s presence. Team roles consisted of a leader and three staff members. Staff members were responsible for (a) measuring values of the aircraft in a simulated airspace (e.g., speed, altitude, and range), (b) sharing this information with other team members who were responsible for interpreting it but did not have direct access to it, and finally (c) making recommendations to a team leader based on a standard set of rules of engagement.

The team members were provided some initial training on how to interpret individual values of a unique subset of the nine characteristics of an incoming aircraft with respect to their level of threat (i.e., high speed, low altitude, and close range all indicate a threatening aircraft). Individuals also learned how to combine three of these values into a score for the “rule” for which they were uniquely responsible throughout the study. The scores for members’ rules served as the basis for staff members’ recommendations to their leader. Leaders also had responsibility for rules but also had to learn how to weigh each staff member’s recommendation in order to register an accurate overall team decision. Correct decisions were an additive function of the rule values. Recommendations to leaders and final decisions were made on a 7-point continuum that ranged in aggressiveness from Ignore (a nonaggressive response used to deal with a nonthreatening aircraft) to Defend (the most aggressive response used to deal with a highly threatening target). Teams’ decisions were compared to the correct decisions, and this comparison was used as a basis for the performance feedback each member received following each of the 48 trials. The performance feedback consisted of the outcome of each decision trial (members’ decisions, the actual team’s decision, the correct team decision, and the difference between the actual team decision and correct team decision) as well as a summary of all past outcomes. The nature of this feedback and the sequence of the aircraft encountered by these research participants were very similar.
to that employed by Hollenbeck et al. (1998), and thus the results from this study can be directly compared to that previous study.

Variables

Team gender composition. Teams were composed randomly as individual participants arrived at our laboratory. Manipulating gender composition was an alternative we did consider; however, we felt that any benefits of such a manipulation were outweighed by the potential costs of priming our participants. Team gender composition was operationalized as the proportion of males on the team and thus could range from 0 to 100%. Eleven teams has 0% males, 16 teams had 25% males, 21 teams had 50% males, 22 teams had 75% males, and finally there were 10 100% male teams.

Situational strength: Mean bias feedback. In addition to the outcome feedback presented to each participant after each decision (described above), the participants in approximately half the teams (randomly determined and crossed with gender composition) received feedback on their own and their team members’ mean bias in past decisions. This feedback was presented on each team members’ computer screen directly below icons representing each team member. Information was presented on two horizontal lines separated by .5 inches. These lines were scaled in terms of the seven decisions participants could make about each target. Specifically, a “1” on the far left of the line represented the least aggressive response (i.e., Ignore) and a “7” to the far right of the line represented the most aggressive response (i.e., Defend). The top line of the feedback was labeled with a “Y” (i.e., “your” decision) and an arrow on that line indicated the mean of the team member’s decisions up to that point in the scenario. The bottom line of the feedback was labeled with a “C” (i.e., “correct” decisions) and an arrow on that line indicated the mean of the correct decisions up to that point in the scenario. Participants in the mean bias feedback condition were told that they should work to keep their arrow in line with the correct decision arrow. They were also told that if their arrow is further right (i.e., closer to 7) than the correct arrow, they were being overaggressive, and if their arrow is further left (closer to 1) than the correct arrow, they were being underaggressive.

The core constructs of the MLT. For comparative purposes, the core MLT constructs were measured in a fashion consistent with past research (e.g., Hedlund et al., 1998; Hollenbeck et al., 1995; Hollenbeck, Colquitt et al., 1998; Hollenbeck, Ilgen, et al., 1998).

An individual team member was informed about a decision if he or she obtained the information needed to make a recommendation based on their assigned rule. Obtaining all the cue information for each rule was very important. This is because aircraft characteristic values combined multiplicatively in determining true rule scores, and therefore each individual characteristic value had the potential to have a strong impact on the overall rule
recommendation. Accordingly, team informity was operationalized as the number of team members who had all the information they needed to make a recommendation for their role, averaged across all 48 trials. Team informity could range between 0 and 4.

As stated earlier, staff validity is the degree to which the team has staff members whose judgments are predictive of the “true score” of the decision. With this definition in mind, and because leaders can theoretically adjust their decisions for staff bias, staff validity was operationalized as the absolute value of the correlation between each staff member’s recommendation and the correct decision, averaged across each staff member over the 48 decision trials. Staff validity, therefore, could range from 0.00 to 1.00.

As noted above, effectiveness in hierarchical sensitivity implies that the team has a leader who uses the best possible weight for each staff member’s judgment when combining these judgments to arrive at the team’s decision. The best set of weights that can be applied to staff judgments would be given by the $B_{mc}$ weights associated with the regression of the true score criterion on the three judgments made by the staff members. Any deviation from this set of weights by the leader would result in a performance decrement relative to the optimum. The actual weights that the leader applies to the staff members’ judgments are given by the $b_{ml}$ weights associated with the regression of the leader’s decision on the staff member’s judgments. Therefore, hierarchical sensitivity was operationalized as follows:

$$HS = \sum_{m=1}^{n_j} \frac{|B_{mc} - B_{ml}|}{n_j},$$

where $B_{mc}$ is the $B$ weight for member $m$’s judgment in predicting the correct decision, $B_{ml}$ is the $B$ weight for member $m$’s judgment in predicting the leader’s decision, and $n_j$ is the number of staff members. A score of .00 on hierarchical sensitivity indicates that the leader applied an optimal set of weights to staff member recommendations. Higher scores on this measure indicate that the leader did a relatively poor job of weighing the recommendations of members of his or her staff. Although this operationalization is consistent with policy capturing research and past MLT research, there are a number of assumptions with this measure because it is an absolute value of a difference. Although these assumptions have been highlighted before (e.g., Edwards, 1994), they are worth mentioning again here. First, there is an assumption that hierarchical sensitivity has the same effect regardless of whether a leader relies too much or too little on team member inputs. Second, there is an assumption that the level of magnitudes of the correct $B$ weights and actual $B$ weights do not matter. Although violations of these assumptions may have substantive and analytical consequences, they are consistent with the MLT in that the theory claims that any deviation from the optimal set of weights by the leader would result in a performance decrement. Testing these assumptions is beyond the scope of the present research because the central purpose of this article is not to develop
the MLT or its core constructs. However, such effort might be worthwhile for future research.

Dependent variables. We captured teams’ scores for mean bias by taking the simple difference between the mean of the teams’ decisions and the mean of the criterion or correct decisions. We note that the correct decisions were constant across all teams.

As Campbell (1990) suggests in his chapter on the role of theory, the inability of research to accumulate in a systematic fashion has been a real problem in our field. Accordingly, we also calculated scores for MSE (the sum of the squared differences between the team’s decisions and the correct decisions) so that readers could compare our results with results of past research that used overall measures of decision accuracy.

Finally, to verify our assertion that the social role variables would influence overall MSE primarily because of effects on mean bias, and the MLT variables would influence MSE because of effects on linear consistency, we also calculated scores for linear consistency and variability bias. Consistent with Gigone and Hastie (1997), we indexed linear consistency by the correlation of the teams’ decisions and the correct decisions. In order to capture any directional effects on variability bias, we modified Gigone and Hastie’s (1997) measure. Specifically, we measured variability bias by the difference between the standard deviation of the teams’ judgments and the standard deviation of the correct decisions.

RESULTS

The descriptive statistics associated with the team-level variables (n = 80) measured in this study appear in Table 1.

Primary tests of the hypotheses.

To test Hypotheses 1 and 2 we hierarchically regressed mean bias on team gender composition, mean bias feedback, and the interaction of team gender composition and the feedback. The results of this regression are shown in column 1 of Table 2.

Team gender composition was not a statistically significant predictor of mean bias. Therefore, Hypothesis 1 was not supported. Consistent with what one would expect, however, teams given the feedback had lower levels of mean bias than teams not given the feedback, and this distinction accounted for 9% of the variance in mean bias.

With respect to Hypothesis 2, the team gender composition by mean bias feedback interaction was statistically significant, explaining 11% of the variance in mean bias. The nature of this interaction is illustrated in Fig. 1 where we plotted the mean bias scores for the teams within each cell in our design. The pattern is consistent with Hypothesis 2. Although there was no team composition effect on mean bias in the presence of the feedback, there was a
### TABLE 1
Descriptives and Correlations

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Team Gender Composition&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.51</td>
<td>.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Mean Bias Feedback</td>
<td>0.52</td>
<td>.50</td>
<td>.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Hierarchical Sensitivity</td>
<td>0.08</td>
<td>.04</td>
<td>.15&lt;sup&gt;*&lt;/sup&gt;</td>
<td>-.16&lt;sup&gt;*&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Staff Validity</td>
<td>0.55</td>
<td>.07</td>
<td>-.08</td>
<td>.11</td>
<td>-.27&lt;sup&gt;**&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Team Informity</td>
<td>3.15</td>
<td>.36</td>
<td>-.12</td>
<td>-.00</td>
<td>.07</td>
<td>.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Mean Squared Error</td>
<td>1.42</td>
<td>.57</td>
<td>.21&lt;sup&gt;**&lt;/sup&gt;</td>
<td>-.29&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.45&lt;sup&gt;**&lt;/sup&gt;</td>
<td>-.41&lt;sup&gt;**&lt;/sup&gt;</td>
<td>-.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Mean Bias</td>
<td>0.46</td>
<td>.27</td>
<td>.16&lt;sup&gt;*&lt;/sup&gt;</td>
<td>-.29&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.23&lt;sup&gt;**&lt;/sup&gt;</td>
<td>-.10</td>
<td>.05</td>
<td>.62&lt;sup&gt;**&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Linear</td>
<td>0.70</td>
<td>.10</td>
<td>-.02</td>
<td>.07</td>
<td>-.33&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.61&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.05</td>
<td>-.62&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.01</td>
<td></td>
</tr>
<tr>
<td>9. Variability Bias</td>
<td>0.12</td>
<td>.17</td>
<td>.05</td>
<td>-.19&lt;sup&gt;*&lt;/sup&gt;</td>
<td>.38&lt;sup&gt;**&lt;/sup&gt;</td>
<td>-.19&lt;sup&gt;*&lt;/sup&gt;</td>
<td>-.09</td>
<td>.51&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.43&lt;sup&gt;**&lt;/sup&gt;</td>
<td>-.19&lt;sup&gt;*&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Note. N = 80 teams.
<sup>a</sup> Proportion of males 0 to 100%.
<sup>*</sup> p < .10, one tailed.
<sup>**</sup> p < .05, one tailed.

### TABLE 2
Hierarchical Regression Results

<table>
<thead>
<tr>
<th>Variables</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Bias</td>
<td>Mean Squared Error</td>
<td>Linear Consistency</td>
<td>Variability Bias</td>
</tr>
<tr>
<td></td>
<td>R&lt;sup&gt;2&lt;/sup&gt;</td>
<td>ΔR&lt;sup&gt;2&lt;/sup&gt;</td>
<td>R&lt;sup&gt;2&lt;/sup&gt;</td>
<td>ΔR&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Social Role Theory</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Team Gender Composition&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.02</td>
<td>.02</td>
<td>.04&lt;sup&gt;*&lt;/sup&gt;</td>
<td>.04&lt;sup&gt;*&lt;/sup&gt;</td>
</tr>
<tr>
<td>2. Mean Bias Feedback</td>
<td>.12&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.09&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.13&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.09&lt;sup&gt;**&lt;/sup&gt;</td>
</tr>
<tr>
<td>3. Team Gender Composition × Mean Bias Feedback</td>
<td>.23&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.11&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.22&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.09&lt;sup&gt;**&lt;/sup&gt;</td>
</tr>
<tr>
<td>Multilevel Theory</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Hierarchical Sensitivity</td>
<td>.23&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.00</td>
<td>.32&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.10&lt;sup&gt;**&lt;/sup&gt;</td>
</tr>
<tr>
<td>5. Staff Validity</td>
<td>.23&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.00</td>
<td>.40&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.08&lt;sup&gt;**&lt;/sup&gt;</td>
</tr>
<tr>
<td>6. Team Informity</td>
<td>.24&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.01</td>
<td>.40&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.00</td>
</tr>
</tbody>
</table>

Note. N = 80 teams.
<sup>a</sup> Proportion of males 0 to 100%.
<sup>*</sup> p < .10.
<sup>**</sup> p < .05.
significant positive effect of team gender composition on mean bias in the absence of the feedback. However, somewhat unanticipated was the clear presence of a nonlinear trend in the relationship between the proportion of men and mean bias in the no mean bias feedback condition.

As a post hoc test of the significance of this nonlinear trend, we reran the regressions reported in Table 2 adding the square of team gender composition and the interaction between the squared gender composition term and mean bias feedback. Adding these two variables contributed an additional 5% of variance explained in mean bias. Teams composed of all females or a female majority or that were balanced in terms of males and females displayed about the same low levels of mean bias. Teams with a male majority showed higher levels of mean bias than these other types of teams, and all-male teams were even worse than those with a male majority.

To test Hypotheses 3–5, we regressed linear consistency on the core MLT constructs (i.e., team informity, staff validity, and hierarchical sensitivity) after we first entered team gender composition, mean bias feedback, and the interaction of team gender composition and mean bias feedback. The first three rows in column 3 of Table 2 show that these three social role variables had no effect on linear consistency. However, rows 4, 5, and 6 of Table 2 show that the MLT variables explain 40% of the variance in linear consistency. Supporting Hypotheses 3 and 4, hierarchical sensitivity explained 10% of the variance in linear consistency and staff validity explained an additional 30%. These effects were consistent with expectations in that linear consistency was higher for teams with staffs that made valid recommendations and for teams with leaders who effectively weighted staff members’ recommendations. The effect of team informity on linear consistency was not significant. Therefore, Hypothesis 5 was not supported.

We performed an additional six-step hierarchical regression of MSE, first on the social role variables and then on the MLT variables. Team gender composition explained a marginally significant 4% of the variance in mean squared error and mean bias feedback explained an additional 9%. Teams with a higher
proportion of males and who received feedback on mean bias made decisions that were more accurate than other teams. The interaction of team gender composition and the feedback explained an additional 9% of the variance in mean squared error. As with the results for mean bias, the effect on mean squared error was neutralized in the presence of mean bias feedback, and there was evidence of a curvilinear effect in the absence of the mean bias feedback. The core MLT constructs explained an additional 18% of the variance in mean squared error, albeit the effect of team informity was not statistically significant. As indicated by the zero-order correlations reported in Table 1, the directions of the effects for staff validity and hierarchical sensitivity were consistent with expectations. That is, teams made more accurate decisions when their staffs made valid recommendations and when leaders effectively weighed those recommendations.

Predicting Variability Bias

To assess the extent that our results replicate and reinforce past findings with respect to the Multilevel Theory, and to confirm our expectations about relationships with the social role variables, we also repeated the six-step regression, this time with nondirectional variability bias as the dependent variable. These results appear in column 4 of Table 2. Of the social role variables, there was only one marginally significant effect—mean bias feedback on variability bias. However, the overall model at this step was not statistically significant. Relative to the MLT variables, only the effect of hierarchical sensitivity was statistically significant, explaining 12% of the variance in variability bias. This effect was also found in a previous MLT study (Hollenbeck, Colquitt et al., 1998), suggesting that leaders who are competent in applying appropriate weights to staff members’ recommendations may also be more competent in neutralizing staffs that are too variable in their recommendations.

Summary of Results for A Priori Hypotheses

Overall, it appears that integrating Social Role Theory with the Multilevel Theory was useful in that it allowed us to identify nonredundant predictors of overall team decision-making performance. Specifically, whereas the variables identified by Social Role Theory influenced overall decision-making performance due to the effects on mean bias, the variables identified by the Multilevel Theory influenced overall decision-making performance primarily due to the effects on linear consistency. Indeed, by considering both of these theories together in the same model, we were able to explain about twice as much variance in overall decision-making accuracy as either theory could have explained alone, thereby demonstrating the strengths of the criterion decomposition approach. In order to seek out additional support for our hypotheses, and also to highlight other issues related to gender composition and team decision making, we conducted several post hoc analyses.
An Alternative Approach to Team Gender Composition

The first additional issue we addressed concerned how we operationalized team gender composition. Whereas we combined the gender of the staff members and the gender of the leader together in our measure, it is possible that leader gender has effects that differ from those caused by the gender composition of the staff. When teams are differentiated hierarchically, leader and staff members have the potential to make unique contributions to overall team effectiveness (LePine, Hollenbeck, Ilgen, & Hedlund, 1997) and our composite measure of team gender composition may have masked these types of differences. Accordingly, we decomposed team gender composition into leader gender (0 = female, 1 = male) and staff gender composition (proportion of male staff members). We then considered these scores along with mean bias feedback and the relevant interactions (i.e., leader gender \times staff gender composition, leader gender \times mean bias feedback, staff gender composition \times mean bias feedback, and leader gender \times staff gender composition \times mean bias feedback) in regressions with mean bias, mean squared error, linear consistency, and variability bias as the dependent variables. Overall, the results of this analysis demonstrated that effects of team gender composition were carried by the staff’s gender composition and not by the leader’s gender. Specifically, there were main and interactive effects for staff gender, but no main or interactive effects for leader gender.

Dynamic Team Gender Composition Effects

A second issue we addressed relates to whether gender differences in social behavior change over time as members become more familiar with each other’s competencies and the actual demands of the task (e.g., Bradley, 1980; Harrison, Price, & Bell, 1998). If this were true, relationships with team gender composition should decrease over time, and this would limit the generalizability of our findings to situations were team members are relatively unfamiliar with one another and their task. In order to examine this issue, we compared relationships between variables in the early trials and late trials. We calculated new scores for mean bias, mean squared error, linear consistency, and variability bias for trials 1–24 (first half of the scenario) and then for trials 25–48 (second half of the scenario). We then repeated the analysis appearing in Table 2. Overall, the parameters were about the same across early and late trials so it does not appear that the effects of the social role variables change as a function of time in any meaningful way.

**DISCUSSION**

This study was designed to make both a substantive and methodological contribution to the literature on team decision-making. Substantively, this study reexamined status quo conclusions regarding the superiority of all male
teams on traditionally masculine tasks. Methodologically, this study demonstrated the use of criterion decomposition techniques to show how the integration of two theories (Social Role Theory and the Multilevel Theory) can be used to augment the amount of variance explained in a complex, multidimensional criterion (decision-making accuracy). The nature of both these contributions are discussed below.

Gender Composition and Team Decision-Making Accuracy

Current trends in the nature of work and the nature of the work force imply that mixed-gender teams will increasingly be called upon to perform work. The empirical literature suggests that this may create a problem in the sense that, on traditionally male dominated tasks, teams composed of male members have been found to outperform teams that include female members on broad measures of performance (Wood, 1987). The current study, employing a decomposition approach and a more focused measure of performance, challenges this notion.

Despite the fact that the cognitive task employed in this study was masculine in nature, inaccuracy in decision-making was actually an exponential function of the number of males on the team. Accuracy was generally high and invariant for teams that were (a) composed of all women, (b) composed of a female majority, or (c) balanced in gender composition. However, teams where men constituted the majority performed poorly and teams composed of all men performed worse than any other configuration.

Although the masculine nature of the task might lead one to think that men would perform better, an analysis of the cognitive task requirements using Social Role Theory actually suggested men would do worse because of cross-situational tendencies to be overaggressive. This inaccuracy was detected in terms of a specific type of decision-making error—mean bias. Although differences on this specific aspect of accuracy (mean bias) did spill over into performance differences on the more global measure of accuracy, the effect for the global measure was weaker. This was because gender composition was found to have no appreciable effects on the other two aspects of accuracy—linear consistency or variability bias.

Because Social Role Theory could be used to derive predictions regarding mean bias, but not the other two forms of inaccuracy, this shows the need for specificity when hypothesizing relationships between variables like gender composition, on the one hand, and complex, multidimensional criterion like decision-making accuracy on the other. Also, the nonlinear nature of the gender composition results suggests that effects associated with individual differences may be compounded in team contexts. Although past literature on Social Role Theory has documented heightened levels of aggressiveness in males, this study showed that in a team context, the effects of maleness were not additive, but rather exponential.

We should note, however, that the data did not suggest that the male-dominated teams were consciously trying to be aggressive just for its own sake. If
this were true, then the feedback on mean bias would not have totally neutralized the effect for gender composition. If males were trying to be aggressive, it is unlikely that teams dominated by males would reduce their aggressiveness because they received feedback on their aggressiveness. If males were trying to be aggressive for the sake of being aggressive, they would continue to be aggressive with the feedback, perhaps even more so if they were trying to "outdo" each other. It seems that men in this study were not trying to be overaggressive—they were trying to be accurate—they just simply could not temper their aggression without feedback on the mean bias of past decisions.

We also note that the women in this study were not "cautious" in any sense. Indeed, in an absolute sense, female-dominated teams were a little overaggressive relative to the true score. The overaggressiveness for female-dominated teams was just much less than that found for male-dominated teams. Because the "true scores" for the aircraft were rectangularly distributed, female-dominated teams could have shown mean bias in the cautious direction, but this did not occur. Thus, the findings should not be interpreted as indicating that male-dominated teams make one type of error and female-dominated teams make the opposite kind of error. Rather, male-dominated teams were making one type of error that the female-dominated teams did not make.

We were somewhat surprised that teams of each gender composition tended to make overaggressive decisions. In hindsight, however, this finding may be understandable, given that females are no less adaptive to the perceived demands of the task context than males. To the extent that the task was perceived to be masculine, females may have behaved more aggressively than they would have in a gender-neutral setting in order to cope with the perceived demands of the task. Another possible explanation for this tendency is that participants' may have perceived the costs of a Type II Error to be higher than the costs of a Type I Error. That is, participants may have perceived the costs of failing to engage a life-threatening aircraft to be greater than the costs of engaging a non-life-threatening aircraft.

Although generally supportive of Social Role Theory, this study also helped to establish some boundary conditions. Consistent with what has been argued by Mischel (1977) and Weiss and Adler (1984), the effects posited by Social Role Theory were totally neutralized by a strong situational context (at least as operationalized in terms of feedback on mean bias in past decisions). This has potential applied implications because in many situations, constraints caused by labor market shortages or other factors may preclude sexual desegregation of some occupations. Thus, if any employer is constrained to employ decision-making teams where males will constitute the majority or totality of those on the team, feedback on results of past decisions may be useful in terms of warding off some of the potential problems. In situations where feedback on past levels of aggressiveness is not feasible (e.g., where a team is relatively "new" and has no record of past decision outcomes), it might be useful to warn male-dominated teams that they may tend to make decisions that are over-aggressive.

The importance of decomposing the overall criterion could also be seen in
the results for the feedback intervention. The feedback received by the “control” group in this study was actually excellent feedback when evaluated in an absolute sense (Ilgen, Fisher, & Taylor, 1979)—it was accurate, immediate, credible, and expressed directly in terms related to overall accuracy. However, because this feedback was directed at overall accuracy rather than the specific type of inaccuracy occurring in male-dominated teams, it was not useful in reducing the specific type of inaccuracy that occurred in such teams. Thus, for feedback to be effective and for it to constitute a strong situation, there has to be a tight coupling between the nature of the feedback and the precise nature of the errors that are likely to arise in different types of situations.

Criterion Decomposition and Theoretical Integration

Although we have pointed out several benefits that accrue from criterion decomposition, perhaps one of the biggest liabilities to this approach is that increased fragmentation of performance might lead researchers to “lose the forest for the trees.” Particularly in applied disciplines, where researchers are asked to contribute to the prediction of complex, multidimensional organizational outcomes, this could be seen as detriment in terms of bridging the “scientist-practitioner” gap.

However, the research reported here shows that one way of “recomposing” the global criterion is to employ different theories to predict different unidimensional facets and then integrate the theories in an effort to predict the overall measure of performance. For example, in this study, we used Social Role Theory to derive predictions regarding mean bias, and based on prior research (Hollenbeck, Colquitt, et al., 1998), we used the Multilevel Theory to predict linear consistency. Together, the two theories were able to explain roughly twice as much variance in overall decision-making accuracy than either could have explained in isolation. Their integration was particularly fruitful because they each addressed nonoverlapping facets of accuracy. As a result of this, the type of “diminishing returns” one typically sees when adding predictors was less in evidence in this context because of the lack of covariation among predictors.

As was apparent from Table 2, although the Multilevel Theory was able to predict some of the variance in variability bias (teams that were high in hierarchical sensitivity also seemed to show less variability bias), for the most part, this third aspect of decision-making accuracy was not explained well by either theory. At this point in the theory building process, the criterion decomposition approach suggests one of two avenues. First, one may want to continue theory building directed at further elucidating the two facets of accuracy that are relatively well explained at this point. Alternatively, one might direct theory building effort toward the facet of the overall criterion unaccounted for by existing theories—variability bias. The first approach might encounter the type of diminishing returns one typically sees when expanding the number of predictors, whereas the second may avoid this outcome.

In any event, the criterion decomposition approach seems to inform decisions related to the parsimony—explanation trade-off that is central to the theory
building process. This approach points to areas where there may be a maximum return, in terms of variance explanation, for adding one or two specific predictors. Although the integration of different theories is certainly not novel in the applied social and behavioral sciences, criterion decomposition focuses the process in a way that supports the “divide-and-conquer” nature of the scientific process (Simon, 1991).

Limitations and Recommendations for Future Research

As with any single study, there are limitations inherent in the present research that need to be addressed by future studies. Certainly, the laboratory nature of this study may limit its generalizability to certain real-world contexts. However, it is should be noted that it is extremely difficult to conduct rigorous research on gender composition and team performance in real organizational settings in jobs that are male-dominated. In most traditionally male-dominated contexts where there are women, it is often the case that the women have less experience. There might also be subtle differences in the nature of the work conducted by men or women that are not picked up by a job description, but might nonetheless confound performance comparisons. Finally, few organizational contexts have a clear, objective criterion, and hence biases in supervisory evaluations may stand as an alternative explanation for the type of performance differences favoring male groups documented by Wood (1987). Thus, whereas the generalizability of our results clearly needs to be empirically assessed, several factors inherent in this study enhanced the validity of the conclusions relative to what might have been the case had it been conducted in the field. In this study, we know that experience on the task was equal for men and women, that men and women faced the exact same task demands, and that the performance measure was uncontaminated by stereotypic biases on the part of human judges.

In this study, we note that our task and subjects achieved a fair level of mundane realism as described by Carlsmith, Ellsworth, and Aronson (1976) and Berkowitz and Donnerstein (1982). First, most members of military command and control teams are lieutenants, and thus, our participants are about the same age and education level. Second, the command and control task is one where individuals sit in a small area (about the same size as the room in our laboratory), at video display terminals, and collect quantitative information in order to make a series of judgments and decisions over a fairly short period of time. Finally, command and control team members generally share information through computer mediation, just as is done in our study.

Of course, a military command and control environment does not resemble the majority of jobs done by workers in the United States. Additionally, the expectations for females in military environment may be quite different in the United States than in other countries. Accordingly, research needs to be conducted with a focus on other types of tasks and cultures. Perhaps most importantly, research should be conducted in settings were group activity is
not computer-mediated and where time between decision cycles is not so compressed.

We do acknowledge that the consequences for failure in our task were much lower than in real command and control contexts. However, there were consequences that mattered to our participants, and thus, we believe our task also had experimental or psychological realism (Berkowitz & Donnerstein, 1982; Carlsmithe et al., 1982). Specifically, participants seemed very engaged in the task and often made comments to each other that reflected a strong desire to perform well so that they could earn one of the financial bonuses. We must also acknowledge that the presence of such incentives might preclude the applicability of our findings to situations where individuals do not care if they succeed or fail. Moreover, if participants were uniformly motivated, the incentive may have controlled for unmeasured individual differences related to motivation (e.g., conscientiousness). Reducing the variability in motivation is also likely to have reduced the variability in team performance. Accordingly, it is possible that the observed effect sizes are more conservative than they would have been had teams not been so constrained.

While it is often the case that members of command and control teams do not know one another well and have to work together as a team to do missions that could last just minutes, the short-term nature of the task may stand as a boundary condition to our findings. This is because some of the problems documented in male-dominated teams may dissipate with more time on the task or more time spent working together. Although hardly the most rigorous test of this possibility, we did examine the all-male teams that did not receive feedback on mean bias to see if they improved over time, and there was no evidence that they did. That is, despite timely and objective feedback in terms of overall accuracy on nearly 50 separate decision trials, the male-dominated teams showed no signs of “catching on” to the fact that their decisions were overaggressive. Although teams in organizations typically work together for longer periods than those studied here, the teams in this study made decisions over a fairly large number of iterations. In fact, it seems reasonable to expect that the compressed nature of time and the clear feedback should have heightened learning—especially relative to what one might expect in more drawn-out, noisy, real-world environments. If this is true, then the compressed nature of the task may have unintentionally provided a conservative setting for assessing our hypotheses.

An additional limitation is that we did not collect direct data on agentic and communal characteristics. Accordingly, we cannot directly test the process described by Social Role Theory. However, while we cannot rule out the possibility that some other variable other than agentic or communal characteristics might be able to explain the observed effects for gender composition, there are a number of factors that should reduce this concern. First, there is convergent meta-analytic evidence that males and females tend to differ in ways that are consistent with the agentic versus communal distinction (Eagly, 1995; Feingold, 1994). In terms of behavioral tendencies, males tend to be more aggressive and assertive, and females tend to be more tender-minded, empathetic,
helpful. In terms of behavior within groups, males tend to be more instrumental and focused on the task, while females tend to be more emotionally expressive and focused on maintaining social harmony. Perhaps most important, the empirical literature shows that these gender differences are not trivial in size (d’s range from about .33 to 1 SD depending on the trait) and appear to be stable across cultures, age groups (especially when comparing college age respondents to adults), and the time periods.

Second, while we do not measure aggression as an individual difference, we do capture it behaviorally and explicitly include it as an integral component of our model. Specifically, we show that in certain circumstances, male-dominated teams make poor decisions because their decisions are overaggressive and not because they are too variable or because they are not predictive of the correct decision. Furthermore, we showed that this bias increases exponentially as the number of males on a team increases. Thus, while men and women do differ on a number of dimensions, at least some of the effect of team gender composition on decision-making performance can be attributed to the tendency for males to behave aggressively, especially when the group consists of a large proportion of males.

We also did not measure participants’ experience with military games and computers, and therefore it is impossible to rule out potential gender differences in these experiences as alternative explanations for our results. However, there are factors that weigh against this concern. First, assuming that males had more experience with computers and military games, then it would be reasonable to expect that male-dominated teams would have performed better than female-dominated teams. However, in our experiment, female-dominated teams performed better. Second, the computer simulation is a fairly “low-tech” DOS-based program that uses simple pull-down menus. There was nothing moving on the screen during the scenario, thus, any gender differences in spatial ability should not have influenced our results. Participants only needed to use a mouse to point to a window and they selected the windows by clicking the mouse. No special computer skills were needed that would have favored males. Anyone familiar with Windows had more than enough computer skill to be highly effective in the mechanical aspects of task. There were also no joysticks or triggers or graphics that might have been more familiar to males than females. We suspect that regardless of gender, all college juniors should have been able to operate the procedural aspects of the task easily and without confusion.

For the most part, this was a study of young people, and perhaps the negative effects of male-dominated gender configurations might be limited to contexts involving young men. Social Role Theory does not present separate characterizations for younger and older men, and hence, this context does seem to be a legitimate test and application of this theory. We do acknowledge, however, that many of the traits this theory ascribes to men (assertive, controlling, aggressive, independent, adventurous, and competitive) may be exacerbated by youth. Of the potential limitations of this sample and context, we feel this is the one in most serious need of future attention because Social Role Theory
may need to be amended to take age, as well as gender, into consideration. This issue also has applied implications because it suggests that in contexts where teams cannot be diversified by gender, they may need to at least be diversified in terms of age in order to offset the overaggressive tendencies associated with homogeneous groups of young men. Again, this is speculative because to date no one has shown that age is a critical boundary condition for Social Role Theory. We are merely raising this issue because this is an interesting possibility that cannot be tested in the present context.

We also note that gender is only one dimension on which team members could differ. While our sample was relatively homogeneous relative to age, race, ethnicity, experiences, and skills, research focusing on diversity with respect to these other characteristics should be conducted. It might be interesting to examine whether younger teams are more aggressive than older teams, and if so, whether including older members on younger teams can neutralize this tendency.

Finally, there are potential issues with the measurement of some of the MLT variables and it is possible that these issues may have played some role in our findings. Scores for hierarchical sensitivity, as stated above, were absolute values of differences, and there are issues with this type of measure that influence the extent to which one can draw conclusions from one's data. Future research that focuses on the MLT should examine the core MLT variables and, at the very least, verify whether the assumptions inherent in the scores used to measure the constructs are appropriate. Research might also focus on different ways of measuring the core constructs altogether. For example, if one decomposes hierarchical sensitivity, it could also be thought of as the similarity of predicted values from a regression of the leader's decision on the staff members' decisions and the predicted values from the regression of the correct decisions on the staff members' decisions. Future research on the MLT could verify the extent to which measures such as this are improvements over the original formulations.

Conclusion

Although many books and popular press articles have lauded the role of diversity in teams, there is actually very little rigorous empirical research that one can point to that supports the widespread superiority of diverse groups over homogeneous ones when it comes to task performance (Argote & McGrath, 1993). Indeed, one of the most widely cited meta-analyses on this topic suggests that all-male teams show a slight performance advantage over mixed teams or all-female teams, especially on masculine tasks (Wood, 1987). Consistent with what would be expected from Social Role Theory (Eagly, 1987), our study showed that, on a decision-making task, all-male teams actually comprised the worst configuration, despite the fact that the task was masculine in nature. The problems in male-dominated groups can be attributed to a specific type of error—mean bias—and this bias can be neutralized when teams are given feedback on past decisions expressed in these terms. Integrating Social Role
Theory (which predicted mean bias) with the Multilevel Theory (which predicted linear consistency) virtually doubled the amount of variance in overall team decision-making accuracy that could be accounted for in this study because of the nonredundancy of the predictor set generated by each theory.

REFERENCES


Received September 20, 1999; published online April 25, 2002