INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Social-collaborative complexity of work: Implications for ability test validities

Gehrlein, Thomas Mark, Ph.D.
Rice University, 1993
SOCIAL-COLLABORATIVE COMPLEXITY OF WORK:
IMPLICATIONS FOR ABILITY TEST VALIDITIES

by

THOMAS M. GEHRLEIN

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
DOCTOR OF PHILOSOPHY

APPROVED, THESIS COMMITTEE

Robert L. Dipboye, Ph.D., Director
Professor of Psychology

Barbara B. Gaugler, Ph.D.
Assistant Professor of Psychology

Richard J. Stoll, Ph.D.
Professor of Political Science

Ronald N. Taylor, Ph.D.
Professor of Psychology

Houston, Texas
May, 1993
ABSTRACT

Social-Collaborative Complexity of Work: Implications for Ability Test Validities

by

Thomas M. Gehrlein

The social, collaborative, interpersonal, and interdependence demands of jobs, referred to as social-collaborative complexity (SCC), are rapidly expanding in the work place. One implication for human resources research is that the criterion-related validities of basic ability tests might be affected. This possibility is provocative, because although much research has shown that basic ability tests generally are valid predictors of performance in a wide range of jobs, increased SCC might modify that conclusion. Thus, a field study and a laboratory experiment were concurrently conducted to explore whether SCC moderates ability test validities. In the field study, measures of SCC for a wide range of jobs were obtained from the Position Analysis Questionnaire and the Dictionary of Occupational Titles. For these same jobs, General Aptitude Test Battery validities were obtained from the United States Employment Service. Results suggested that several job-related measures of SCC significantly moderated GATB validities. Furthermore, the moderating effects of SCC were independent of the more predictable moderating effects of other job demands. The lab experiment was conducted to explore whether motivational aspects of SCC moderated test validities. Specifically, the effects of accountability on task performance and test validity were explored in individual and group contexts. Subjects in the 2 x 2 experiment completed, among other measures, a creative thinking ability test and a brainstorming task. Results suggested that ability, accountability, and social context interacted to affect performance, and accountability and social context both moderated test validity in a "moderated moderator" effect. Implications for the work place and suggestions for future research are discussed.
Acknowledgments

I would like to thank the members of my committee, Drs. Robert Dipboye, Barbara Gaugler, Ronald Taylor, and Richard Stoll, for their input into this dissertation. A special thanks to Bob Dipboye for chairing the committee and guiding me through each step of the process. Finally, thanks also to my parents, Gerald and Antonina Gehrlein, for always being there with love and support.

I dedicate this dissertation to my wife, Chusi, who generously gave love, patience, and kindness to me during the roughest times of graduate school. Chusi, I couldn't have done it without you.

Thomas M. Gehrlein
Houston, Texas
May, 1993
# Table of Contents

Study 1 ...................................................................................................................... 2  

Reasons Why Social-Collaborative Complexity Might Be Negatively Related to Test Validity ................................................................................................................. 6  

Reasons Why Social-Collaborative Complexity Might Be Positively Related to Test Validity ............................................................................................................. 10  

The Potential Moderating Effects of Social-Collaborative Complexity on General Aptitude Test Battery Validity ................................................................. 12  

Method ...................................................................................................................... 14  

Description of the Sample. ..................................................................................... 14  

Job Analysis Ratings .............................................................................................. 15  

Position Analysis Questionnaire (PAQ) Data ......................................................... 17  

Dictionary of Occupational Titles (DOT) Data .................................................... 20  

General Aptitude Test Battery (GATB) Data ......................................................... 24  

Moderator Analyses .............................................................................................. 25  

Results .................................................................................................................... 28  

Cognitive GATB Tests .......................................................................................... 28  

Perceptual GATB Tests .......................................................................................... 30  

Psychomotor GATB Tests ....................................................................................... 30  

Discussion .............................................................................................................. 33  

Job Analysis Measures as Moderators of GATB Validity ..................................... 33  

Cognitive GATB Validities ...................................................................................... 33  

Perceptual GATB Validities .................................................................................... 35  

Psychomotor GATB Validities ............................................................................... 36  

Conclusions and Suggestions for Future Research .............................................. 37  

Study 2 ................................................................................................................... 39
Conclusions .................................................................................................................. 71
References ..................................................................................................................... 76
Appendix A--Position Analysis Questionnaire Dimensions in Study 1 ......................... 86
Appendix B--Statistical Procedures to Detect Moderators of Validity ......................... 91
Appendix C--Comparison of Results of Study 1 to Gutenberg et al. (1983) ................. 95
Appendix D--Supplemental Instructions for the Thumbs Problem .............................. 96
Appendix E--Regression Equations for Study 2 ............................................................ 98
Appendix F--Results of Generalizability Analysis ......................................................... 99
Appendix G--Results of Moderator Analyses for All 105 Jobs .................................... 100
List of Tables

Table 1: Criterion-Related Validity of the Minnesota Clerical Test in Weinstein and Holzbach (1973) Study ................................................................. 9

Table 2: Occupations in Study 1, the Dictionary of Occupational Titles, and the U.S. Work Force .................................................................................. 18

Table 3: Means, Standard Deviations, and Intercorrelations of Job Analysis Measures .............................................................................................. 21

Table 4: Means, Standard Deviations, and Intercorrelations of General Aptitude Test Battery ....................................................................................... 26

Table 5: Correlations Between Job Analysis Measures and Cognitive GATB Validities ................................................................................................. 29

Table 6: Correlations Between Job Analysis Measures and Perceptual GATB Validities ................................................................................................. 31

Table 7: Correlations Between Job Analysis Measures and Psychomotor GATB Validities ................................................................................................. 32

Table 8: Means and Standard Deviations of Task Context Manipulation Checks .......................................................... 56

Table 9: Means and Standard Deviations of Accountability Manipulation Checks .......................................................... 57

Table 10: Mean Score on the Thumbs Problem by Condition .................................................................................. 59

Table 11: Regression of Thumbs Problem Performance on the Alternate Uses Test, Task Context, and Accountability .................................................. 60

Table 12: Pearson Correlation Between Score on the Alternate Uses Test and Performance on the Thumbs Problem by Condition ........................................ 65
List of Figures

Figure 1: Flow Diagram for Method (Study 1)..........................................................16

Figure 2: Mean Performance on the Thumbs Problem for Low and High Ability Subjects by Level of Accountability and Type of Task Context..............................................62

Figure 3: Predictor-Criterion Regression Line for Each Cell.......................................64
Social-Collaborative Complexity of Work: Implications for Ability Test Validities

The small, self-regulating group is touted by some organizational theorists as the basic building block of the successful organization (Shea & Guzzo, 1987). Indeed, the work place is undergoing a revolution of groupwork in which quality circles, employee involvement programs, and autonomous teams are proliferating. As a result, principles of collaboration and teamwork are being infused into organizations at a rapid pace (Walton & Hackman, 1986). This trend impacts psychologists, managers, and (especially) workers who experience a fundamental shift from traditional, individually-oriented jobs to more complex and demanding group-oriented work arrangements.

An implication of this shift is that the social, interpersonal, collaborative, and interdependence demands of many jobs are increasing. The term social-collaborative complexity (SCC) will be used here to describe the diverse constellation of job attributes related to working with and among other people. The importance of this people-oriented constellation is widely recognized by theorists of human work, who frequently describe job attributes as related to either people, data, or things (Fine, 1989).

Increased SCC will affect many areas of organizational research including performance appraisal, group development, job analysis, and job evaluation. Increased SCC also might have implications for the relationships between ability and performance that are reflected in the validities of personnel tests. Although much research has shown that ability tests generally are valid predictors of performance in a wide range of jobs (National Research Council, 1989), increased SCC might modify that conclusion. Psychologists cannot afford to ignore these issues, but the effects of SCC on test validity have rarely been addressed.

Given this gap in the literature, the primary question addressed in the present research was: does SCC moderate ability test validities? Study 1, a field study of archival data, and Study 2, a laboratory experiment, were concurrently conducted to
explore this question from different perspectives. In Study 1, measures of SCC for a wide range of jobs in the workplace were correlated with ability test validities to explore whether SCC moderated test validities. Study 2 was conducted to tease out the implications of a few, critical motivational variables related to SCC. Specifically, the effects of accountability and the presence (or absence) of a group context on task performance and test validity were explored in a controlled setting.

Study 1

The purpose of this field study was to broadly explore whether social-collaborative complexity (SCC) moderates ability test validities in the workplace. The study was conducted using two types of archival data that were obtained for a wide range of jobs: (a) measures of SCC and (b) criterion-related validities of several ability tests. The specific measures of SCC and the test validities are described below, but first the rationale for the hypotheses concerning the moderating effects of SCC on test validities will be developed.

The issue of whether SCC moderates test validities really is part of a broader question: do situational variables moderate test validities? Most psychologists probably reject the notion that validity is strictly specific to a particular predictor-criterion combination for a particular job in one organization. Research on validity generalization has done much to discredit the strict, situation-specific hypothesis. Indeed, the National Research Council (1989) strongly supported the use of validity generalization in personnel selection. However, some psychologists do not agree with the Schmidt et al. (1985) conclusion that situational moderators of validity do not merit consideration.

Among them, James, Demaree, Mulaik, and Ladd (1992) suggested that several issues in the debate on moderators of validity are unresolved. First, a basic premise of validity generalization theory is that the between-situation variance in validities is
partially due to situational moderators, but research has almost exclusively focused on the other source of variance, statistical artifacts. Second, the search for moderators has been largely unsuccessful, because it has proceeded in the absence of a theoretical model of how and why situational variables might be important. Finally, in the search for situational moderators of validity, James et al. (1992) suggested that archival data cannot be relied on to the degree that it has been used in the search for statistical artifacts. Instead, proactive research designs also will be required.

As an example, James et al. (1992) described one situational variable that might moderate validity among otherwise highly similar predictor-criterion combinations. They defined restrictiveness of organizational climate as the degree to which a work environment restricts an individual's freedom to express job-relevant behaviors. They hypothesized that validity suffers if a work environment restricts the expression of job-relevant behaviors, because measurement of those behaviors is necessary to empirically estimate ability-performance relationships. For example, it is reasonable to expect that the validity of an oral communication measure will be low in situations where employees have few opportunities to demonstrate their communication skills. Furthermore, a criterion is likely to be unreliable in a restrictive environment, because fewer job-relevant behaviors are available to measure. Thus, a situational variable like restrictiveness of climate might be related to (and, therefore, not independent of) criterion reliability. Nonindependence of situational variables and statistical artifacts violates a basic tenet of validity generalization theory and has serious implications for the soundness of previous validity generalization research.

Other potential moderators of validity have been suggested. Schneider (1978) lamented that little research had explored type of reward system and task characteristics as potential moderators. Peters, O'Connor, and Rudolf (1980) suggested that at least eight resource-related variables (e.g., budgetary support, time availability, and required
services and help from others) affect the expression of job-relevant behaviors and, hence, test validities. Others have suggested that the identification of situational moderators will not necessarily guarantee increased test validities. For example, even if situational constraints on performance are removed, ability does not always predict performance (Terborg, Richardson, & Pritchard, 1980). In some cases, personality traits (e.g., self-esteem and need for achievement) and motivational factors (e.g., social loafing and felt responsibility) might facilitate or hinder the expression of ability.

Some research has demonstrated that situational variables moderate validity. Brown (1981) reported that the validity of a biodata instrument for predicting the sales success of life insurance agents significantly varied among 12 large companies depending on the quality of their personnel recruitment and selection systems. In a field simulation study, Peters, Fisher, and O'Connor (1982) demonstrated that situational variables can moderate test validity in systematic and unsystematic ways. Schmitt, Schneider, and Cohen (1990) reported that situational variables moderated the validity of a supervisory assessment center administered in different locations. All these studies concluded that situational moderators of validity have potentially important implications for the workplace.

Another potential moderator of validity, SCC, is the focus of the present research. Aspects of SCC such as task interdependence (Thompson, 1967), outcome interdependence (Miller & Hamblin, 1963), communication demands (Bass, 1980), and accountability (Tetlock, 1985) can engender strong motivational effects that might affect validity. Furthermore, group tasks sometimes provide opportunities for social loafing (Latane, Williams, & Harkins, 1979) or a feeling of dispensability (Kerr & Bruun, 1983) that can lead to lower individual-level performance and perhaps lower test validity. Conversely, cooperation and collaboration can motivate everyone in a group to
demonstrate their best possible performance; hence, the observed distribution of
performance would contain more true score variance, and test validity would be high.

These issues need to be investigated, but to determine whether SCC moderates test
validities in the work place, it is first necessary to have measures of SCC that are
relevant to real jobs. As mentioned above, people-oriented job demands, which are
critical to SCC, are firmly established in theories of human work. Fortunately, several
well-researched, broadband job analysis instruments include measures of the people-
oriented demands of jobs. For example, the Job Diagnostic Survey (Hackman &
Oldham, 1975) and the Job Characteristics Inventory (Sims, Szilagyi, & Keller, 1976)
both measure a job dimension labeled "dealing with others". However, neither
instrument suggests what worker behaviors or attributes are required to perform SCC-
related work.

Theory suggests that SCC requires workers, among other things, to give and
receive verbal and written information, and to plan, coordinate, and perform work with
and among other people (Shea & Guzzo, 1987). These and other job-required, worker
behaviors or attributes are measured by the Position Analysis Questionnaire (PAQ), a
structured job analysis instrument that has been widely researched and used in many
different organizations to study diverse occupations (McCormick, Jeanneret, &
Mecham, 1972). Because the PAQ is a widely used, well-researched job analysis
instrument that measures many elements of SCC (which are specified below in the
Method section), it provided an opportune way to investigate whether SCC moderates
test validities. Thus, PAQ data and other job analysis measures were used in Study 1 to
explore whether the SCC of jobs moderates ability test validities in the work place.

Alternative hypotheses were explored in Study 1, because if the SCC of jobs
moderates test validities, the relationships might reasonably be positive or negative.
These possibilities are elaborated below, but first the rationale behind them will be
developed. Specifically, there is some evidence that elements of SCC, including task interdependence and cognitive complexity, moderate ability test validities. Furthermore, motivational components of SCC are known to affect work performance, and might, therefore, affect test validity.

Reasons Why Social-Collaborative Complexity Might Be Negatively Related to Test Validity

One reason why social-collaborative complexity (SCC) might moderate test validity is that task interdependence, which is a critical element of SCC, can reduce test validity. Task interdependence has been discussed for decades in the organizational design and small groups literatures and was a central concept in Thompson's (1967) widely cited book on open systems organizational theory. Task interdependence has been variously defined as the flow of work among group members (Wong & Campion, 1991), the connectedness between jobs such that performance of one depends on the successful performance of the other (Kiggundu, 1983), and the amount of task-driven interaction among group members (Shea & Guzzo, 1987). Thus, task interdependence can be viewed as a job characteristic like skill variety, autonomy, etc.

Bass (1980) suggested that task interdependence is one job characteristic that might moderate ability-performance relationships that are otherwise reasonably well established. However, only one published study has explored this possibility. In a 2 x 2 lab experiment, Weinstein and Holzbach (1973) administered the Minnesota Clerical Test to undergraduates who then were randomly assigned to three-person groups in one of four conditions: high or low task interdependence and differential or equal reward distribution. Subjects performed a clerical task for one hour in which they coded responses to a three-part questionnaire. In the low interdependence conditions, subjects worked on separate questionnaires and simply pooled their coding sheets to determine the group's output. In the high interdependence conditions, subjects worked
sequentially such that each subject was instructed to code a specific part of the questionnaire and then pass it to the next subject. In all conditions, six cents went into a pot for each form coded; therefore, the total reward available to the group was a function of total group output. In the differential reward conditions, 1/2, 1/3, or 1/6 of the pot was awarded to each individual depending on performance relative to the other group members. In the equal reward conditions, 1/3 of the pot was awarded to each group member regardless of performance.

As hypothesized, productivity was higher in the low interdependence conditions than the high interdependence conditions ($p < .05$) and was higher in the differential reward conditions than the equal reward conditions ($p < .05$). (The interdependence x reward structure interaction was not significant.) Overall, the Minnesota Clerical Test was a valid predictor of individual performance ($r = .33, p < .01$); however, as shown in Table 1, the test was valid in the low interdependence-equal reward condition only. The difference between validity in the that condition and validity in the high interdependence-equal reward condition was significant ($p \leq .05$). Combined across reward conditions, validity was marginally significantly greater ($p \leq .10$) under low interdependence than high interdependence. Commenting on these results, Weinstein and Holzbach (1973) suggested that:

Should this differential validity hold up in an employment setting, there would be some interesting implications with respect to the Equal Employment Opportunity Commission (EEOC) guidelines on employee hiring and promotion procedures, which require the tester to determine that a significant difference between conditions is absent when 'protected groups' are involved.... the results of this research indicate that the validity... may vary significantly between reward and task-flow-interdependence conditions and that a tester may be violating EEOC guidelines by not testing the moderating effects of the situation (p. 300).
(Note that Weinstein and Holzbach did not use multiple regression to determine whether experimental factors moderated test validity in their study. The question of whether multiple regression or correlational differences is the proper way to assess moderators of validity is a controversial issue that is further explored in Appendix B of this dissertation.)

Weinstein and Holzbach did not suggest how higher interdependence resulted in lower test validity, yet their experimental design holds important clues. In the low interdependence conditions, group members coded separate questionnaires and simply pooled their results. Thus, group members did not pose any task-related contingencies for each other. Not surprisingly, test validity was as high (.47) as might be expected if an unrestricted sample of workers individually performed a clerical task in the work place. In the high interdependence conditions, however, groups performed the coding task sequentially. Thus, an individual’s potential output was directly restricted by those who preceded her/him, and validity was not significant (.09). Sequential interdependence in a group can constrain individual-level performance, and, as will be demonstrated below, severely restrict test validity.

Assume that three persons can be ordered along the X-axis of a predictor-criterion scatter plot. Let "H" represent the person with highest test score, "M" is the middle score, and "L" is the lowest score. The triad can be arranged in six unique sequences: LMH, LHM, MLH, MHL, HLM, and HML. Because sequential interdependence imposes specific performance contingencies, the maximum possible test validity in the first three arrangements is zero regardless of performance on the criterion. The maximum possible validity in the fourth and fifth arrangements approaches but never equals +1. Only when the group is arranged HML can test validity be as high as +1. Maximum validity is achieved only if performance fits a highly specific pattern. Actual validity is likely to be lower and, indeed, is likely be negative in the first three
Table 1

Criterion-Related Validity of the Minnesota Clerical Test in Weinstein and Holzbach (1973) Study

<table>
<thead>
<tr>
<th>Rewards</th>
<th>Low</th>
<th>n</th>
<th>High</th>
<th>n</th>
<th>Combined</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differential</td>
<td>24</td>
<td>18</td>
<td>16</td>
<td>18</td>
<td>08</td>
<td>36</td>
</tr>
<tr>
<td>Equal</td>
<td>67**</td>
<td>18</td>
<td>-23</td>
<td>18</td>
<td>46**</td>
<td>36</td>
</tr>
<tr>
<td>Combined</td>
<td>47**</td>
<td>36</td>
<td>09</td>
<td>36</td>
<td>33**</td>
<td>72</td>
</tr>
</tbody>
</table>


** p < .01.
arrangements. This phenomenon occurs when a group operates in strict sequence, and accuracy or quality of performance is high (e.g., quantity and quality of output were correlated at .95 in the Weinstein and Holzbach study).

As one element of SCC, task interdependence is likely to become more frequent and important in the workplace. Furthermore, task interdependence is one reason why SCC might be negatively related to test validity. However, there is at least one reason to expect that SCC might be positively related to test validity, and that is the relationship between SCC and cognitive complexity.

**Reasons Why Social-Collaborative Complexity Might Be Positively Related to Test Validity**

As the social-collaborative complexity (SCC) of jobs increases, cognitive demands also increase. For example, when workers participate in autonomous teams, they must process information from many sources and make decisions that previously were out of their control. Furthermore, complex, group-oriented arrangements require additional communication and social skills. Thus, if higher levels of SCC are associated with higher levels of cognitive complexity, then there is reason to expect that SCC is positively related to ability test validities.

Gutenberg, Arvey, Osburn, and Jeanneret (1983) demonstrated that cognitive job demands moderated test validities. The rationale for their study was similar to sentiments described above, namely, that although situational factors have been largely dismissed as potential moderators of test validity, the search is incomplete. They noted that most previous research focused on molecular variables (e.g., task differences), but little attention was paid to the potential moderating effects of worker-oriented job demands. They hypothesized that: (a) cognitive tests would be more valid for jobs high on information-processing and problem-solving, and (b) physical tests would be more valid for jobs high on physical demands. To test these hypotheses, they correlated
scores from three cognitive and two physical PAQ dimensions with GATB validity coefficients for 111 diverse jobs. When correlations were computed between cognitive dimensions and GATB validities, physical dimensions were partialled out, and vice versa.

Consistent with hypothesis (a), cognitive PAQ dimensions significantly moderated cognitive GATB validities. Of nine possible correlations between three cognitive PAQ dimensions (decision making, information processing, and overall general responsibility) and three cognitive GATB validities (General Intelligence, Verbal Ability, and Numerical Ability), eight were statistically significant (r range: +.18 to +.37). Hypothesis (b) was not supported; two physical PAQ dimensions (controlled manual activities and handling activities) did not significantly moderate any of three physical GATB validities (Motor Coordination, Finger Dexterity, and Manual Dexterity). Nonetheless, the cognitive results were impressive, and the study was a rare demonstration of potentially widely relevant moderators of ability test validity (James et al., 1992).

The Gutenberg et al. (1983) study supported the relatively straightforward hypothesis that cognitive GATB validities are higher in jobs that are more cognitively demanding. However, it is less clear whether SCC moderates test validities and whether it does so in a positive or negative direction. Thus, in light of recent trends in the workplace, the need to explore whether SCC moderates test validities is pressing. As mentioned above, this issue was explored in Study 1. The Position Analysis Questionnaire (PAQ), which was the primary measure of SCC, was introduced above, and the General Aptitude Test Battery (GATB), which was the other critical element in this research, is described below. Finally, alternative hypotheses about the moderating effects of SCC on GATB validities are set forth before the procedure for Study 1 is described in detail.
The Potential Moderating Effects of Social-Collaborative Complexity on General Aptitude Test Battery Validity

Much of the previous discussion was framed in terms of personnel tests, in general. However, the form and purpose of these tests widely vary; therefore, it is likely that the effects of social-collaborative complexity (SCC) on test validity likewise will vary. For example, personality tests might become more valid as interpersonal skills become more important in the work place. However, basic ability test validities might decline as SCC increases, because the link between ability and performance might be less clear when groups pose complex contingencies for individual performance. Conversely, ability test validities might increase if the social pressures of group tasks motivate everyone to demonstrate their best performance. These issues potentially could be studied with any personnel measure; however, it might be initially useful to examine the General Aptitude Test Battery (GATB), which is the most widely used and validated civilian employment test battery in the U.S. (Holden, 1989; LaFraniere, 1991; National Research Council, 1989; U.S. Employment Service, 1988).

Under continuous development by the U.S. Employment Service (USES) since 1947, the GATB measures nine aptitudes required for successful performance in many jobs: (a) general intelligence, (b) verbal aptitude, (c) numerical aptitude, (d) spatial aptitude, (e) form perception, (f) clerical perception, (g) motor coordination, (h) finger dexterity, and (i) manual dexterity (U.S. Department of Labor, 1983). These nine aptitudes were identified from factor analyses of 59 tests administered in the early 1940's to over 2,000 adults, most of whom were trainees in vocational courses. Additional analyses conducted last decade on scores from over 23,000 job applicants confirmed that the original factor structure is still evident in the current GATB (U.S. Department of Labor, 1983). Results also suggested that the nine aptitudes break into three general clusters: (a) cognitive (general intelligence, numerical, and verbal), (b) perceptual
(spatial, form, and clerical), and (c) psychomotor (motor, finger, and manual dexterity). The cognitive and psychomotor clusters are relatively independent ($r = .46$), whereas both clusters are highly correlated with the perceptual cluster ($r = .88$ and $r = .75$, respectively). As is common practice, the nine aptitudes measured by the GATB will be referred to here as tests.

Because the GATB measures a wide variety of aptitudes, and the battery has been extensively validated, it provided a unique opportunity to broadly explore whether the SCC of jobs moderates ability test validities in the work place. However, it was difficult to predict the direction of potential moderating effects. As described above, Gutenberg et al. (1983) explored the relatively straightforward hypothesis that the cognitive GATB tests would be more valid in cognitively demanding jobs. However, the potential moderating effects of SCC are not straightforward. One reasonable hypothesis was that, in general, SCC is negatively related to any ability test validity. The rationale is that the higher the SCC of a job the more likely it is that job performance is affected by forces outside of an individual's control; hence, ability-performance relationships are weaker.

The effects of outside forces need not be deleterious, however, to yield lower test validities. For example, workers sometimes collaborate and cooperate to perform complex physical tasks, because several pairs of hands and arms are better than one. As a result, everyone's performance increases, but the cooperative nature of performance obscures individual differences that otherwise would be observable if people performed alone. Thus, if people in jobs high on SCC perform physical tasks relatively well, the validity of psychomotor ability tests might be low in these jobs. A similar rationale could be applied to the cognitive arena. Suppose that everyone performs relatively well when a team works together to tackle intellectually challenging tasks. The cooperative nature of performance obscures individual differences and might lead to lower cognitive ability test validities.
Conversely, SCC might be positively related to test validities. For example, it is reasonable to expect that a verbal aptitude test will be more valid in jobs with higher SCC, because these jobs require more communication and personal interaction. Likewise, a general intelligence test will be more valid in jobs with higher SCC, because these jobs involve more complex contingencies. However, if SCC moderates verbal or intelligence test validities, it might be because measures of SCC and measures of cognitive complexity are correlated. Thus, the moderating effects of SCC need to be explored before and after statistically controlling for cognitive complexity. These alternatives were explored in Study 1.

Method

To broadly explore whether social-collaborative complexity (SCC) moderates ability test validities in the workplace, two types of archival data were obtained for a variety of jobs: (a) job analysis ratings, including measures of SCC and other job demands (primarily from the Position Analysis Questionnaire [PAQ]) and (b) test validities (from the General Aptitude Test Battery [GATB]). The units of analysis were 105 diverse jobs for which both PAQ and GATB data were available. A flowchart of the Method is shown in Figure 1, and each step in the process is described below.

Description of the Sample

The first step was to identify jobs for which both PAQ and GATB data were available. Each job had a unique Dictionary of Occupational Titles (DOT) number (U.S. Department of Labor, 1977; 1991). The DOT has been a standard source of occupational information in the U.S. for almost 60 years and is used for a variety of human resource activities including career counseling, job redesign, and occupational research (Cain & Green, 1983). The first three digits of a DOT number identify the occupational group in which a job is classified. The second three digits are ratings that represent the degree to which a job requires working with data, people, and things. The
last three digits differentiate jobs with the same first six digits but different job titles. DOT numbers were used to identify jobs in the current PAQ and GATB data bases (described below), both of which contain data from the early 1970's to the late 1980's.

As shown in the second step of Figure 1, the PAQ and GATB data bases were merged to obtain the data set for the present study. The 105 jobs in this data set represented a broad spectrum of occupations including, for example, electronic technicians, nurses, public safety officers, plant operators, employment interviewers, auto body repairers, psychiatric aides, etc. The distribution of these jobs was compared to the distribution of jobs in the DOT and the 1980 U.S. work force. (1980 was chosen as a benchmark, because it occurred near the middle of the time period covered by the data base.) The results are shown in Table 2.

Compared to the DOT, clerical and sales jobs were overrepresented and blue-collar occupations were underrepresented in the present study. Compared to the 1980 U.S. work force, blue-collar occupations were overrepresented and professional and managerial occupations were underrepresented in the present study. It was not possible to compare jobs in the present study to the Gutenberg et al. (1983) study. Data for the latter were collected in the early 1970's, whereas data for the present study covered most of the 1970's and 1980's. Furthermore, the quality of research-oriented PAQ data has greatly improved in the past decade, because quality control mechanisms have been instituted to detect incorrectly coded forms (R. C. Mecham, personal communication, February, 1992). The job analysis ratings, including measures of SCC and other job demands, and GATB validities obtained for each job are described below.

Job Analysis Ratings
Figure 1. Flow Diagram of Method (Study 1)

<table>
<thead>
<tr>
<th>PAQ Services, Inc.</th>
<th>United States Employment Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAQ dimension scores</td>
<td>DOT ratings, GATB scores &amp; validities</td>
</tr>
<tr>
<td>(early 1970's - late 1980's) for:</td>
<td>(1972 - 1987) for:</td>
</tr>
<tr>
<td>JOB A</td>
<td>JOB A</td>
</tr>
<tr>
<td>JOB B</td>
<td>JOB B (Validity study 1)</td>
</tr>
<tr>
<td>JOB C</td>
<td>JOB B (Validity study 2)</td>
</tr>
<tr>
<td>JOB E</td>
<td>JOB C</td>
</tr>
<tr>
<td>JOB F</td>
<td>JOB D</td>
</tr>
<tr>
<td>(Total Jobs: Hundreds)</td>
<td>(Total Jobs: 168)</td>
</tr>
</tbody>
</table>

V

Study 1 Data Base:

JOB A: PAQ dimension scores, DOT ratings, GATB scores & validities
JOB B: PAQ dimension scores, DOT ratings, GATB scores & validities (2 studies)
JOB C: PAQ dimension scores, DOT ratings, GATB scores & validities
(Total Jobs: 105)

V

For jobs with multiple validity studies, GATB scores and validities were averaged

V

GATB validities were adjusted for unreliability of the criterion and restriction of range of the predictor

V

MODERATOR ANALYSES: GATB validities were correlated with PAQ dimension scores and DOT ratings
Position Analysis Questionnaire (PAQ) Data. The PAQ consists of 195 items organized into six divisions: (1) information input, (2) mental processes, (3) work output, (4) relationships with other persons, (5) job context, and (6) other job characteristics. Division 4 measures relationships with other persons that are necessary to perform the job (McCormick, Jeanneret, & Mecham, 1989; McPhail, Jeanneret, McCormick, & Mecham, 1991). Five factors underlie the 36 items in division 4: (1) communicating judgments and related information; (2) engaging in general personal contact; (3) performing supervisory, coordination, and related activities; (4) exchanging job-related information; and (5) public and outside personal contacts (Mecham, McCormick, & Jeanneret, 1989). Examples of items in division 4 include: exchanging routine and nonroutine information, negotiating, persuading, general job-required personal contact, contact with specific types of personnel, coordination and collaboration activities, serving/catering, and supervising nonemployees such as campers or patients. Thus, PAQ division 4 is a broadband measure that captures many elements of the SCC of jobs.

The PAQ has been successfully used in personnel research to: (a) estimate the interest and temperament attributes of workers in different professions (Marquardt & McCormick, 1972), (b) estimate validities and score ranges of the General Aptitude Test Battery (GATB) (McCormick, DeNisi, & Shaw, 1979) and commercially available aptitude tests (McCormick, Mecham, & Jeanneret, 1989) (c) assess the structural validity of Holland's theory of vocational types (Hyland & Muchinsky, 1991), (d) demonstrate that the decision-making and information-processing dimensions of jobs moderate GATB validities (Gutenberg, Arvey, Osburn, & Jeanneret (1983), (e) predict the Myers-Briggs Type Indicator profiles of workers in different occupations (Mecham, 1988), and (f) empirically demonstrate that job requirements are related to stress and its effects on worker health (Shaw & Riskind, 1983).
Table 2

Occupations in Study 1, the Dictionary of Occupational Titles, and the U.S. Work Force

<table>
<thead>
<tr>
<th>Occupational category</th>
<th>Study 1</th>
<th>3rd edition</th>
<th>U.S. work force</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional, managerial &amp; technical</td>
<td>13%</td>
<td>12%</td>
<td>27%</td>
</tr>
<tr>
<td>Clerical and sales</td>
<td>22%</td>
<td>8%</td>
<td>25%</td>
</tr>
<tr>
<td>Service occupations</td>
<td>9%</td>
<td>4%</td>
<td>13%</td>
</tr>
<tr>
<td>Craft, repair, operators &amp; laborers</td>
<td>47%</td>
<td>67%</td>
<td>32%</td>
</tr>
<tr>
<td>Agriculture and fishing</td>
<td>0%</td>
<td>2%</td>
<td>3%</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>9%</td>
<td>7%</td>
<td>--</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>


In the present study, PAQ dimension scores were the primary measures of SCC and other job demands. These data were obtained from the composite data base maintained by PAQ Services, Inc., in Logan, Utah. A PAQ composite is a summary of all individual PAQs that were scored from the early 1970's to the late 1980's and that have the same DOT number coded on them. For example, if 10 organizations each submitted to PAQ Services a PAQ with the same DOT number, then one composite in the data base represents the average of these 10 PAQs. In the present study, the median number of PAQs per composite was 20 (range: 3 - 657). Information about the number of PAQs per composite in the PAQ Services data base and the Gutenberg et al. (1983) study was not available.

Scores were obtained for all five dimensions of PAQ Division 4 (Relationships with Other Persons) as measures of the SCC of jobs. Scores were also obtained for the three cognitive and two physical PAQ dimensions studied by Gutenberg et al. (1983). Finally, scores were also obtained for three perceptual PAQ dimensions. (The definition of each PAQ dimension explored in the present study is shown along with representative items in Appendix A.) In the original PAQ research, dimensions were standardized to a mean of zero and a standard deviation of one (McCormick, Jeanneret, & Mecham, 1972). Means and standard deviations of PAQ dimensions in the present study are shown in Table 3. Compared to the original PAQ research, mean dimension scores in the present study were somewhat lower and there was somewhat less variability.

Within each division of the PAQ, dimensions theoretically correlate zero in the population; however, dimensions are not necessarily independent in a given sample of jobs. Intercorrelations of PAQ dimensions in the present study are shown in Table 3. Half the correlations within division 4 (dimensions 17-21, Relationships with Other Persons) were significant; however, none was larger than .33. Dimensions 7 and 8, which constitute division 2 (Mental Processes), were virtually uncorrelated ($r = -.02,$
Dimensions 13 and 15, which are in division 3 (Work Output), were significantly correlated ($r = -.36, p < .01$). Finally, dimensions 3 and 5, which are in division 1 (Information Input), were not significantly correlated ($r = -.13, \text{ns}$).

Dimensions in different divisions of the PAQ are not necessarily uncorrelated (McCormick, Jeanneret, & Mecham, 1972). As shown in Table 3, many cross-divisional relationships were significant in the present sample of jobs. The implication was that the PAQ dimensions of primary interest in the present study (17-21) shared some variance with cognitive, perceptual, and psychomotor dimensions. Thus, as described below, the moderating effects of SCC-related dimensions on GATB validities were estimated before and after partialing out the effects of other dimensions, and vice versa.

In addition to PAQ dimension scores, two DOT-based job analysis measures were explored in the present study. The DOT measures were obtained along with GATB data from the U.S. Employment Service (USES). All the data obtained from the USES is described below.

**Dictionary of Occupational Titles (DOT) Data.** Additional job analysis measures and the results of GATB validation studies were obtained from the Michigan Employment Security Commission, a repository for research conducted by the USES. The two job analysis measures were part of the data base of ratings on which the DOT was constructed (U.S. Department of Labor, 1972, 1977, 1991). To construct the DOT, the USES conducted a massive occupational analysis program in which 100+ trained analysts in 10 field offices observed, described, and rated thousands of jobs throughout the U.S. Each job was rated on 46 characteristics, three of which appear as the middle digits of the DOT number. As mentioned above, these three "worker function ratings" represent the degree to which a job requires working with data, people, and things. (The theory and research underlying these ratings originated from Functional Job Analysis, a widely-used job and task analysis methodology [Fine, 1989]).
Table 3

Means, Standard Deviations, and Intercorrelations of Job Analysis Measures

Position Analysis Questionnaire (PAQ) dimensions

<table>
<thead>
<tr>
<th>Social-Collaborative Complexity</th>
<th>Cognitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>7</td>
</tr>
<tr>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>19</td>
<td>33</td>
</tr>
<tr>
<td>20</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>53</td>
</tr>
<tr>
<td>53</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>53</td>
</tr>
<tr>
<td>33</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>7</th>
<th>8</th>
<th>33</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>-.51</td>
<td>-.01</td>
<td>-.27</td>
<td>-.25</td>
<td>-.29</td>
<td>-41</td>
<td>-.40</td>
</tr>
<tr>
<td>S</td>
<td>.45</td>
<td>.44</td>
<td>.31</td>
<td>.61</td>
<td>.51</td>
<td>.67</td>
<td>.56</td>
</tr>
<tr>
<td>D</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>18</td>
<td>06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>-33</td>
<td>-26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>19</td>
<td>-14</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>-20</td>
<td>-07</td>
<td>-23</td>
<td>-18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>48</td>
<td>-18</td>
<td>31</td>
<td>53</td>
<td>-41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>53</td>
<td>06</td>
<td>43</td>
<td>25</td>
<td>36</td>
<td>-02</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>84</td>
<td>11</td>
<td>-03</td>
<td>52</td>
<td>-20</td>
<td>76</td>
<td>44</td>
</tr>
<tr>
<td>13</td>
<td>-01</td>
<td>-16</td>
<td>47</td>
<td>56</td>
<td>-38</td>
<td>60</td>
<td>-19</td>
</tr>
<tr>
<td>15</td>
<td>08</td>
<td>39</td>
<td>-31</td>
<td>-18</td>
<td>03</td>
<td>-37</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>06</td>
<td>-37</td>
<td>27</td>
<td>32</td>
<td>-40</td>
<td>60</td>
<td>-17</td>
</tr>
<tr>
<td>5</td>
<td>02</td>
<td>22</td>
<td>14</td>
<td>31</td>
<td>-08</td>
<td>13</td>
<td>-08</td>
</tr>
<tr>
<td>35</td>
<td>34</td>
<td>-16</td>
<td>-32</td>
<td>51</td>
<td>39</td>
<td>08</td>
<td>84</td>
</tr>
<tr>
<td>D</td>
<td>38</td>
<td>15</td>
<td>-24</td>
<td>13</td>
<td>38</td>
<td>04</td>
<td>44</td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>67</td>
<td>-15</td>
<td>-05</td>
<td>40</td>
<td>-13</td>
<td>75</td>
<td>37</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3 (Cont.)

**Means, Standard Deviations, and Intercorrelations of Job Analysis Measures**

<table>
<thead>
<tr>
<th>PAQ dimensions</th>
<th>Psychomotor</th>
<th></th>
<th>Perceptual</th>
<th></th>
<th>DOT Ratings</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13</td>
<td>15</td>
<td>3</td>
<td>5</td>
<td>35</td>
<td>DP</td>
</tr>
<tr>
<td><strong>M</strong></td>
<td>.02</td>
<td>-.05</td>
<td>.17</td>
<td>-.42</td>
<td>-.36</td>
<td>.82</td>
</tr>
<tr>
<td><strong>S</strong></td>
<td>.79</td>
<td>.54</td>
<td>.82</td>
<td>.64</td>
<td>.63</td>
<td>.97</td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>52</td>
<td>-24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>26</td>
<td>-10</td>
<td>-13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>03</td>
<td>08</td>
<td>-04</td>
<td>03</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>-17</td>
<td>-01</td>
<td>-31</td>
<td>15</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td><strong>P</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>31</td>
<td>-21</td>
<td>44</td>
<td>05</td>
<td>38</td>
<td>31</td>
</tr>
</tbody>
</table>

**Note.**  
N = 105 jobs. PAQ dimensions are defined in Appendix A. DP = DOT-based People worker function rating; DC = DOT-based substantive complexity factor score. Decimals are omitted from correlations. An $r$ of .19 is significant at $p < .05$. An $r$ of .26 is significant at $p < .01$. 

The people rating (DOT-People) represents the complexity of interpersonal behavior and interaction required in jobs. Lower values represent more complex job requirements: mentoring (0), negotiating (1), instructing (2), supervising (3), diverting (4), persuading (5), speaking or signaling (6), serving (7), and taking instructions or helping (8). Cain and Green (1983) reported that the interrater reliability of the DOT-People rating was in the .80's. The DOT-People rating was included in the present study as an additional measure of the SCC of jobs; however, the scale was reflected, so that higher values represented higher complexity. The range of the DOT-People rating in the present sample of jobs was narrow (0-3 on a 0-8 scale), and the mean was low (0.82, SD = .97); however, the mean for all jobs in the DOT also was low (1.17, SD = 1.85) (Miller, Treiman, Cain, & Roos, 1980). Correlations between the DOT-People rating and the other job analysis measures are shown in Table 3. DOT-People was significantly correlated with three of the five SCC-related PAQ dimensions; however, none of the relationships was large.

The other DOT-based job analysis measure included in the present study was a factor score that represented the cognitive or substantive complexity of jobs. Roos and Treiman (1980) reported that many of the 46 items on which jobs were rated by the USES were redundant. Their factor analysis of the items revealed four independent factors: (a) substantive complexity, (b) motor skills, (c) physical demands, and (d) undesirable working conditions. Eight items loaded highly on the substantive complexity factor: (a) the Data worker function rating, (b) general educational development, (c) specific vocational preparation, (d) intelligence, (e) verbal aptitude, (f) numerical aptitude, (g) abstract and creative activities, and (h) repetitive/continuous activities (negative loading). Roos and Treiman summed the eight items (with unit weights) to yield a substantive complexity factor score, which Gerhart (1988) reported had a reliability of about .70 (the type of reliability was not indicated).
A variation of the substantive complexity factor score was used in the present study as a comparison to the PAQ cognitive dimensions. Six of the eight items included in the Roos and Treiman substantive complexity factor were available for the present study. Items (g) and (h), which had the smallest loadings in the factor analysis, were not available. Thus, a modified substantive complexity factor score (DOT-Complexity) was calculated for each job by standardizing the six available items and then summing them with unit weights. Correlations between DOT-Complexity and the other job analysis measures are shown in Table 3. DOT-Complexity was highly correlated with two of the three cognitive PAQ dimensions: making decisions (dimension 7) ($r = .75, p < .01$) and overall responsibility (dimension 33) ($r = .77, p < .01$). These relationships were not surprising given that both types of measures are primarily worker-oriented rather than work-oriented; that is, they focus on personnel rather than task requirements (Gerhart, 1988).

The various job analysis measures obtained for the present study only constituted half of the information necessary to explore whether the SCC of jobs moderated test validities. The other information, GATB validities, are described below.

**General Aptitude Test Battery (GATB) Data**

For each job in the present study, five pieces of GATB data were available from the USES: (a) mean score for each of the nine GATB tests, (b) standard deviation of each test, (c) criterion-related validity of each test, (d) reliability of the criterion measure in each validation study, and (e) sample size in each validation study. All the validation studies were conducted from 1972 to 1987. Most studies utilized a concurrent (86.1%) rather than a predictive (13.9%) design. The criterion in most studies was some measure of job performance (88.6%) rather than a measure of training performance (11.4%). The most common measure of job performance was supervisory ratings (80.0%). In most studies (90.1%), the reliability of the criterion measure represented
interrater reliability or rate-rerate reliability. The median validation sample size was 126.5 (range: 27 - 850). The USES had conducted at least one validation study for each of the 105 jobs but had conducted multiple studies (two to five) for 28 jobs. For each job in which the GATB had been validated more than once, validities were weighted by sample size and averaged across studies (Hays, 1981). This process is indicated as the third step in Figure 1.

In the population of U.S. job seekers, each GATB test is standardized to a mean of 100 and a standard deviation of 20. Means, standard deviations, and intercorrelations of GATB tests in the present sample of jobs are shown in Table 4. Seven of the nine means were within five points of 100. Standard deviations ranged from 6.6 (Manual dexterity) to 10.5 (General intelligence), but all were less than 20 suggesting that the range of GATB scores in the present sample of jobs was somewhat restricted. This was not surprising given that 86% of the validation studies involved a concurrent design. Finally, correlations shown in Table 4 suggest that, as in the general population, GATB scores were highly intercorrelated in the present sample.

As indicated in the fourth step of Figure 1, validity coefficients were corrected for criterion unreliability and range restriction due to explicit selection on the basis of the predictor (Osburn, 1987). Mean observed and corrected validities are shown for each GATB test along with a generalizability analysis (Hunter, Schmidt, & Jackson, 1982) in Appendix F. Results suggested that less than half the variance in the corrected validity of each test was due to sampling error.

Moderator Analyses

As the last step in the study, the job analysis measures were correlated with GATB validities to determine whether the former moderated the latter. This analytical approach was described by Arnold (1982):
Table 4

Means, Standard Deviations, and Intercorrelations of General Aptitude Test Battery

<table>
<thead>
<tr>
<th>GATB tests</th>
<th>G</th>
<th>V</th>
<th>N</th>
<th>S</th>
<th>P</th>
<th>Q</th>
<th>K</th>
<th>F</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>98.1</td>
<td>97.5</td>
<td>96.1</td>
<td>101.9</td>
<td>104.8</td>
<td>108.2</td>
<td>102.8</td>
<td>97.5</td>
<td>105.3</td>
</tr>
<tr>
<td>SD</td>
<td>10.5</td>
<td>8.6</td>
<td>9.9</td>
<td>9.5</td>
<td>9.3</td>
<td>9.3</td>
<td>8.2</td>
<td>7.1</td>
<td>6.6</td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td>90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td>98</td>
<td></td>
<td>89</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td>89</td>
<td>66</td>
<td>83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>71</td>
<td>72</td>
<td>76</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>55</td>
<td>72</td>
<td>62</td>
<td>32</td>
</tr>
<tr>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50</td>
<td>71</td>
<td>60</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>56</td>
<td>58</td>
</tr>
<tr>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>39</td>
</tr>
</tbody>
</table>

Note. N = 105 jobs. Decimals are omitted from correlations, all of which were significant at p < .01. G = general intelligence, V = verbal aptitude, and N = numerical aptitude (cognitive cluster); S = spatial aptitude, P = form perception, and Q = clerical perception (perceptual cluster); K = motor coordination, F = finger dexterity, and M = manual dexterity (psychomotor cluster).
To test the hypothesis that the strength of relationship between two variables (or the 'degree of predictive validity') is a function of some third continuous 'moderator' variable,... is accomplished straightforwardly by correlating predictor-criterion correlation coefficients with the values of the 'moderator' variable. (p. 152)

This approach has been used in other published research (e.g., Gutenberg et al., 1983; Schmitt et al., 1990) but is controversial, because the more common way to detect moderators of validity is via multiple regression. The statistical issues surrounding moderators of validity, and the rationale for using the Arnold procedure in the present study, are fully described in Appendix B.

To estimate the moderating effect of a measure of SCC, Pearson correlations were computed between the measure and each GATB validity before and after partialing out conceptually different job analysis measures. For example, the moderating effect of PAQ dimension 17 (communicating judgments) on GATB General Intelligence test validity was assessed with two correlations:

\[ r_{PAQ 17, GATB-G} \]

which is the zero-order correlation between PAQ dimension 17 and GATB General Intelligence test validity, and:

\[ r_{PAQ 17, GATB-G \circ PAQ 7, PAQ 8, PAQ 33, DOT-COMPLEXITY} \]

which is the correlation between PAQ dimension 17 and GATB General Intelligence test validity partialing out all cognitive job analysis measures.

Likewise, to estimate the moderating effect of cognitive job demands, corresponding correlations were calculated. For example, the moderating effect of PAQ dimension 8 (information processing) on GATB General Intelligence test validity was assessed with two correlations:

\[ r_{PAQ 8, GATB-G} \]
which is the zero-order correlation between PAQ dimension 8 and GATB General Intelligence Test validity, and

\[ r_{PAQ\ 8,\ GATB-G} = \rho_{PAQ\ 17,\ PAQ\ 18,\ PAQ\ 19,\ PAQ\ 20,\ PAQ\ 21,\ DOT-PEOPLE} \]

which is the correlation between PAQ dimension 8 and GATB General Intelligence Test validity partialing out all measures of SCC.

Similar analyses were conducted within the perceptual and psychomotor GATB clusters. That is, the moderating effects of SCC measures on perceptual GATB validities were calculated before and after partialing out perceptual measures (and vice versa), and the moderating effects of SCC measures on psychomotor GATB validities were calculated before and after partialing out psychomotor measures (and vice versa).

**Results**

The moderator analyses were conducted twice, once for the 93 jobs in which the criterion was a measure of job performance and once for all 105 jobs. Results for the 93 jobs are shown in Tables 5, 6, and 7 and described below. Results for the 105 jobs are shown in Appendix G.

**Cognitive GATB Tests**

Results for the cognitive GATB tests are shown in Table 5. One measure, PAQ dimension 17 (communicating judgments), significantly moderated General Intelligence test validity \( (r = .23, p < .05) \). The moderating effect of dimension 17 remained significant when cognitive measures were partialed out (partial \( r = .29, p < .01 \)). One measure, PAQ dimension 21 (public and outside personal contacts), significantly moderated Verbal Aptitude test validity \( (r = .21, p < .05) \). The moderating effect of dimension 21 remained significant when cognitive measures were partialed out (partial \( r = .21, p < .05 \)). Two measures, PAQ dimensions 17 and 8 (processing information) significantly moderated Numerical Aptitude test validity \( (r = .29, p < .01, \text{ and } r = .21, p \)
Table 5
Correlations Between Job Analysis Measures and Cognitive GATB Validities

<table>
<thead>
<tr>
<th>Job Analysis Measures</th>
<th>GATB Validity</th>
<th>General Intelligence</th>
<th>Verbal Aptitude</th>
<th>Numerical Aptitude</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>r</td>
<td>partial r</td>
<td>r</td>
</tr>
<tr>
<td>Social-Collaborative Complexity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAQ 17</td>
<td>23*</td>
<td>29**</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>PAQ 18</td>
<td>-11</td>
<td>-07</td>
<td>-17</td>
<td>-12</td>
</tr>
<tr>
<td>PAQ 19</td>
<td>-16</td>
<td>-16</td>
<td>-17</td>
<td>-17</td>
</tr>
<tr>
<td>PAQ 20</td>
<td>-03</td>
<td>-15</td>
<td>03</td>
<td>-04</td>
</tr>
<tr>
<td>PAQ 21</td>
<td>05</td>
<td>04</td>
<td>21*</td>
<td>21*</td>
</tr>
<tr>
<td>DOT-People</td>
<td>10</td>
<td>07</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>Cognitive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAQ 7</td>
<td>10</td>
<td>10</td>
<td>08</td>
<td>13</td>
</tr>
<tr>
<td>PAQ 8</td>
<td>12</td>
<td>-04</td>
<td>12</td>
<td>-13</td>
</tr>
<tr>
<td>PAQ 33</td>
<td>10</td>
<td>-08</td>
<td>07</td>
<td>-05</td>
</tr>
<tr>
<td>DOT-</td>
<td>15</td>
<td>03</td>
<td>16</td>
<td>06</td>
</tr>
</tbody>
</table>

Note. N = 93 jobs. Decimals are omitted from correlations. PAQ dimensions are: 17 = Communicating judgments, 18 = General personal contact, 19 = Supervising and coordinating, 20 = Exchanging job-related information, 21 = Public and outside personal contacts, 7 = Making decisions, 8 = Processing information, 33 = Overall decision, communication, and general responsibilities. When a partial for an SCC measure was computed, all four cognitive measures were partialled out. When a partial for a cognitive measure was computed, all six SCC measures were partialled out.

*p < .05. **p < .01.
< .05, respectively); however, only the partial correlation for dimension 17 was
significant (partial r = .27, p < .05). The moderating effect of PAQ dimension 20
(exchanging job-related information) was only significant when cognitive measures were
partialled out (partial r = -.24, p < .05).

It would be expected that one or two r's and one or two partial r's in Table 5 would
be significant at traditional levels on the basis of chance alone. Thus, the results were
better than chance.

Perceptual GATB Tests

Results for the perceptual GATB tests are shown in Table 6. Most cells in the
lower half of the table are blank, because each perceptual PAQ dimension was
conceptually related to only one GATB aptitude. One measure, PAQ dimension 21
(public and outside personal contacts), significantly moderated Spatial Aptitude test
validity but only when PAQ dimension 5 was partialed out (partial r = -.22, p < .05).
One measure, PAQ dimension 20 (exchanging job-related information), significantly
moderated Form Perception test validity (r = -.25, p < .05). The moderating effect of
dimension 20 remained significant when PAQ dimension 3 was partialed out (partial r =
-.21, p < .05). One measure, PAQ dimension 20, significantly moderated Clerical
Perception test validity (r = -.22, p < .05); however, the moderating effect was not
significant when PAQ dimension 35 was partialed out.

It would be expected that one r and one partial r in Table 6 would be significant at
traditional levels on the basis of chance alone. Thus, the results were better than chance.

Psychomotor GATB Tests

Results for the psychomotor GATB tests are shown in Table 7. PAQ dimension
20 (exchanging job-related information) significantly moderated all three psychomotor
GATB validities before and after PAQ dimensions 13 and 15 were partialed out. PAQ
dimension 15 significantly moderated Motor Coordination test validity (r = .23, p < .05)
Table 6

Correlations Between Job Analysis Measures and Perceptual GATB Validities

<table>
<thead>
<tr>
<th>Job Analysis Measures</th>
<th>Spatial Aptitude</th>
<th>Form Perception</th>
<th>Clerical Perception</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GATB Validity</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( r )</td>
<td>partial ( r )</td>
<td>( r )</td>
</tr>
<tr>
<td>Social-Collaborative Complexity</td>
<td>PAQ 17</td>
<td>06</td>
<td>07</td>
</tr>
<tr>
<td></td>
<td>PAQ 18</td>
<td>02</td>
<td>-02</td>
</tr>
<tr>
<td></td>
<td>PAQ 19</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>PAQ 20</td>
<td>-01</td>
<td>01</td>
</tr>
<tr>
<td></td>
<td>PAQ 21</td>
<td>-20</td>
<td>-22*</td>
</tr>
<tr>
<td></td>
<td>DOT-People</td>
<td>-09</td>
<td>-11</td>
</tr>
<tr>
<td>Perceptual</td>
<td>PAQ 5</td>
<td>05</td>
<td>07</td>
</tr>
<tr>
<td></td>
<td>PAQ 3</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>PAQ 35</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Note. \( N = 93 \) jobs. Decimals are omitted from correlations. PAQ dimensions are: 17 = Communicating judgments, 18 = General personal contact, 19 = Supervising and coordinating, 20 = Exchanging job-related information, 21 = Public and outside personal contacts, 5 = Being aware of environmental conditions, 3 = Watching devices and materials for information, 35 = Performing clerical and related activities. When a partial for an SCC measure was computed, the relevant perceptual measure was partialled out. When a parital for a perceptual measure was computed, all six SCC measures were partialled out.

\*_{p < .05}.
Table 7

**Correlations Between Job Analysis Measures and Psychomotor GATB Validities**

<table>
<thead>
<tr>
<th>Job Analysis Measures</th>
<th>Motor Coordination</th>
<th>Finger Dexterity</th>
<th>Manual Dexterity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
<td>partial r</td>
<td>r</td>
</tr>
<tr>
<td>Social-Collaborative Complexity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAQ 17</td>
<td>-04</td>
<td>02</td>
<td>-03</td>
</tr>
<tr>
<td>PAQ 18</td>
<td>02</td>
<td>-07</td>
<td>04</td>
</tr>
<tr>
<td>PAQ 19</td>
<td>-19</td>
<td>-16</td>
<td>15</td>
</tr>
<tr>
<td>PAQ 20</td>
<td>-30**</td>
<td>-29**</td>
<td>-37***</td>
</tr>
<tr>
<td>PAQ 21</td>
<td>10</td>
<td>08</td>
<td>-02</td>
</tr>
<tr>
<td>DOT-People</td>
<td>03</td>
<td>07</td>
<td>-06</td>
</tr>
<tr>
<td>Psychomotor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAQ 13</td>
<td>-10</td>
<td>21*</td>
<td>08</td>
</tr>
<tr>
<td>PAQ 15</td>
<td>23*</td>
<td>18</td>
<td>-01</td>
</tr>
</tbody>
</table>

**Note.** N = 93 jobs. Decimals are omitted from correlations. PAQ dimensions are: 17 = Communicating judgments, 18 = General personal contact, 19 = Supervising and coordinating, 20 = Exchanging job-related information, 21 = Public and outside personal contacts, 13 = Performing controlled manual & related activities, 15 = Performing handling & related manual activities. When a partial for an SCC measure was computed, both psychomotor measures were partialled out. When a parital for a psychomotor measure was computed, all six SCC measures were partialled out.

* p < .05. ** p < .01. *** p < .001.
and Manual Dexterity test validity ($r = .35$, $p < .001$); however, neither moderating effect was significant when measures of SCC were partialed out. PAQ dimension 13 significantly moderated Motor Coordination test validity and Finger Dexterity test validity but only when the measures of SCC were partialed out (partial $r = .21$, $p < .05$, and partial $r = .30$, $p < .01$, respectively).

It would be expected that one $r$ and one partial $r$ in Table 7 would be significant at traditional levels on the basis of chance alone. Thus, the results were better than chance.

**Discussion**

**Job Analysis Measures as Moderators of GATB Validity**

**Cognitive GATB Validities** PAQ dimension 17 (communicating judgments and related information) significantly moderated General Intelligence and Numerical Aptitude test validities before and after cognitive measures were partialed out. PAQ dimension 21 (public and outside personal contacts) significantly moderated Verbal Aptitude test validity before and after cognitive measures were partialed out. These results suggest that job-related measures of SCC are more important than cognitive measures as moderators of cognitive GATB validities. Furthermore, one of the alternative hypotheses described above was supported, namely that General Intelligence test validity would be higher for jobs with higher SCC, because these jobs involve more complex contingencies.

To understand the moderating effects of dimension 17, it is helpful to examine PAQ items that are highly correlated with it: (a) giving/receiving nonroutine or unusual information; (b) advising, counseling, or guiding individuals; (c) job-required personal contact with middle and upper level managers and professional personnel; (d) staff functions (i.e., advising, consulting, and/or assisting managers); (e) negotiating; (f) persuading; and (g) coordinating activities of others (Mecham, McCormick, &
Jeanneret, 1989). Thus, the GATB General Intelligence test apparently is more valid in jobs that involve more complex interactions with other people, and this moderating effect of SCC is independent of the moderating effect of cognitive complexity.

One reason why dimension 17 moderated Numerical Aptitude test validity could be that the distribution of performance on numerical tasks contains more true score variance in jobs with higher SCC. For example, suppose workers in a socially complex environment have to obtain the information necessary to perform job-required calculations from other people. Because information must be obtained from others and is not readily available, the distribution of performance across workers is wider and yields higher Numerical Aptitude test validity. For example, it is likely that social and collaboration skills affect an accountant's or auditor's ability to obtain critical data in a sensitive financial investigation. Data to test this hypothesis is not available, but future research on the moderating effects of SCC on test validity will need to address the effects of SCC on distributions of performance as well.

PAQ dimension 21 (public and outside personal contacts) significantly moderated Verbal Aptitude test validity before and after cognitive measures were partialed out. Thus, the more a job involves personal contact with clients, customers, special interest groups, etc. the higher the validity of verbal aptitude in the job. This moderating effect makes sense, because jobs high on personal contact provide many opportunities for incumbents to express verbally-oriented behaviors that affect overall job performance. Thus, the distribution of performance contains substantial true score variance which results in higher test validities.

Overall, it was difficult to compare the present results for the cognitive GATB cluster to results reported by Gutenberg et al. (1983), because they only reported partial correlations and not zero-order correlations. Furthermore, they partialed out physical measures (i.e., PAQ dimensions 13 and 15) to estimate the moderating effects of
cognitive measures (i.e., PAQ dimensions 7, 8, and 33), whereas, in the present study, measures of SCC were partialed out. Finally, for reasons discussed above, data in the two studies were not directly comparable. However, ancillary analyses were conducted, so that results for the cognitive GATB cluster would be somewhat comparable across the two studies. Results are shown in Appendix C. Results suggested that the cognitive PAQ dimensions (especially dimension 7) were stronger moderators of cognitive GATB validities in the Gutenberg study than the present study, at least when physical measures were partialed out. Of course, it is impossible to know what results Gutenberg et al. would have obtained had they partialed out measures of SCC instead of physical measures, but results of the present study suggest that this is a critical issue.

Perceptual GATB Validities  PAQ dimension 21 (public and outside personal contacts) significantly moderated Spatial Aptitude test validity but only after PAQ dimension 5 was partialed out. As the description in Appendix A suggests, dimension 21 is heavily orientated towards interpersonal communications. Because it is likely that the verbal and spatial demands of jobs are negatively correlated, it makes sense that Spatial Aptitude test validity would be lower in jobs that have a larger verbal component. Furthermore, jobs that are highly verbal offer few opportunities to demonstrate spatially-oriented behaviors, thus the ability-performance relationship is difficult to assess. For these same reasons it makes sense that the moderating effects of PAQ dimension 20 (exchanging job-related information) on Form Perception and Clerical Perception test validities also were negative.

It is difficult to understand why the perceptual PAQ dimensions failed to moderate perceptual GATB validities, because, for example, the construct-level link between the GATB Clerical Perception test and PAQ dimension 35 (performing clerical and related activities) was strong. However, these results suggest that the empirical link between ability test validity (as estimated with the GATB) and job requirements (as estimated
with the PAQ) generally is weaker in the perceptual domain than the cognitive and psychomotor domains. As an alternative explanation, even in heavily clerical jobs criterion measures are contaminated by situational and motivational factors that are not related to clerical ability.

Psychomotor GATB Validities. A measure of SCC, PAQ dimension 20 (exchanging job-related information), significantly moderated all three psychomotor GATB validities before and after psychomotor job measures were partialed out. Indeed, the moderating effect of dimension 20 on Finger Dexterity test validity increased from -.37 to -.51 when the psychomotor PAQ dimensions were partialed out. Therefore, dimension 20 was a strong moderator of psychomotor GATB validities.

These results support the alternative hypothesis that higher SCC is negatively related to ability test validity, because the higher SCC the more likely it is that job performance is affected by forces outside of an individual's control; hence, ability-performance relationships are weaker. For example, if workers collaborate to perform complex and demanding physical tasks, everyone's performance increases. However, this situation obscures individual differences and yields lower psychomotor ability test validities.

Alternatively, as workers spend more time exchanging job-related information, they have fewer opportunities to exhibit the job-relevant behaviors necessary to empirically establish ability-performance relationships. This explanation is consistent with a more benign (but probably more realistic) interpretation of the restrictiveness of climate notion advanced by some theorists (e.g., James et al., 1992; Schneider, 1978). Thus, one reason that the moderating effects of dimension 20 on psychomotor GATB validities were negative is that the link between ability and performance is weaker when the social context poses complex contingencies for individuals. Again, additional research is needed to explore differences in distributions of performance as well as test validity.
Another potential reason that the moderating effects of dimension 20 were negative is that individuals with higher levels of psychomotor abilities perform lower in jobs that demand higher-level social and collaboration skills. Thus, the fit between personal attributes and job demands is poor for these individuals (e.g., Caldwell & O'Reilly, 1990). Future research in this area will need to include additional measures of personal attributes (e.g., personality traits) to more thoroughly explore these possibilities.

Conclusions and Suggestions for Future Research

Results of the present study and the Gutenberg et al. (1983) research support the theory that job requirements moderate the validities of a widely used personnel test battery like the GATB. While a return to the notion that test validity is strictly situation-specific hardly seems warranted, evidence was presented here that the social-collaborative complexity (SCC) of jobs moderates ability test validities. Furthermore, the present results suggest that the moderating effects of SCC are stronger than the moderating effects of cognitive complexity. Finally, as James et al. (1992) argued, there are strong theoretical reasons to suggest that situational moderators of validity are not only not independent of statistical artifacts in validation studies, but that situational variables cause statistical artifacts. For example, a restrictive environment could cause a criterion measure to be unreliable. Additional research clearly is needed to explore these issues with other job analysis measures and other types of personnel tests, because the results have practical implications for the generalizability of ability test validities.

More research is also needed, in general, to improve measurement of SCC. Dimensions in PAQ division 4 (Relationships With Other Persons) were conceptually similar to popular notions of what it means to work with other persons and somehow be interdependent with them. However, as Fine (1989) noted, measurement of people-oriented job demands is in a primitive state compared to measurement of cognitive demands. For example, of the three DOT worker function ratings, the assumption of
ordinality is more tenuous for the People scale than the Data and Things scales (Fine, 1989). In this study, DOT-People was highly restricted and was not highly correlated with PAQ-based measures of SCC. However, the DOT scales offer the advantage that thousands of jobs have been rated on them over the years. Clearly, however, more detailed measures of SCC need to be developed and compared to the DOT and PAQ measures.

The purpose of this field study was to broadly explore whether the SCC of jobs moderates GATB validities. A distinct advantage of this approach was that a large amount of data from a wide variety of jobs and organizations was simultaneously analyzed to identify moderators with potentially wide implications. A disadvantage was that it was not possible to closely examine, in a controlled setting, specific motivational components of SCC that affect job performance and moderate test validity. This caveat is consistent with the James et al. (1992) argument that archival studies (which have been heavily used in research on validity generalization) are a limited tool in the exploration of situational moderators of test validity. Conversely, a laboratory study would allow one to directly address questions such as the effects of SCC on distributions of performance and the potential value of personality data, both of which are related to the issue of moderators of ability test validity.

Laboratory research on this topic was planned before Study 1 was begun. However, unanticipated results of Study 1 suggested additional reasons why a lab study would be valuable. For example, there was evidence that SCC was a stronger moderator of cognitive GATB validities than were cognitive measures. However, it was impossible to probe motivational factors in the present study, and, as mentioned above, motivational aspects of SCC can lead to decrements (or increases) in performance and, ultimately, test validity. Given these anticipated and unanticipated benefits of lab research, Study 2, a laboratory experiment, was conducted concurrent with Study 1.
Study 2

Another way to explore whether social-collaborative complexity (SCC) moderates test validities would be to focus on underlying motivational factors while holding the issue of cognitive complexity constant. In that vein, a distinction can be made between situations in which an individual performs work alone and situations in which performance occurs in a group, because SCC usually is higher in the latter. These different task contexts engender unique motivational effects that could differentially affect task performance and test validity.

However, the effects of the task context on performance and test validity could be complex. To clarify these effects, a cross-cutting variable, such as accountability, also needs to be explored. The relevance of accountability to the basic question addressed by the present research is described below, but suffice it to say that accountability is likely to become a more important feature of the work place as SCC increases.

The purpose of the present study was to explore these issues in a controlled setting. A laboratory experiment was conducted in which two between-subjects variables were manipulated, task context (subjects either worked alone [individual condition] or in a mere group [group condition]) and accountability (subjects were either accountable or unaccountable for their performance). Task performance and test validity were examined in each cell of the 2 x 2 design. Specific hypotheses are set forth below, but first the rationale behind them will be developed.

Individual Versus Group Task Context

Research in social and organizational psychology has repeatedly shown that individuals perform differently when they work alone versus in a group. Most results have suggested that performance is lower in the latter context. For example, research has consistently shown that interacting groups typically do not perform as well as noninteracting or nominal groups on brainstorming tasks (Diehl & Stroebe, 1991;
Gallupe & Bastianutti, 1991). This general phenomenon has alternatively been labeled social loafing (Latane, Williams, & Harkins, 1979), diffusion of responsibility (Harkins & Petty, 1982), and free riding (Kerr & Bruun, 1983).

Why is performance generally lower in a group? One explanation is that individuals feel less personal responsibility for action in a group (Whyte, 1991). However, there is evidence that not everyone experiences a decrement in felt responsibility in a group and that felt responsibility is a variable on which individuals differ (Pearce & Gergersen, 1991). Another explanation is that individuals sometimes feel that the group to which they belong does not have the requisite skills, resources, and support necessary to get the job done, a situation which Shea and Guzzo (1987) termed "low group potency". Differential perceptions of group potency could affect motivation and disrupt an otherwise stable distribution of performance across group members.

The type of task assigned to a group also can have motivational consequences for performance. For example, many group tasks require that member efforts somehow be summed to yield the group's output, which Steiner (1972) termed "additive" tasks. In additive tasks, the group's performance typically does not rest so heavily on any one individual. However, if group members perceive that individual contributions are potentially dispensable and not identifiable, then they might perform below expectations (Harkins & Petty, 1982). Indeed, it has been suggested that the dual considerations of perceived identifiability and perceived dispensability are central to understanding low performance in a group context (Kerr & Bruun, 1983). An unanswered question is what effect an additive group task has on test validity. This and other issues were addressed in the present study.

Other theorists have argued that low performance in a group context is due to complex cognitive operations that are affected by many variables that are yet unidentified (Harcum & Badura, 1990). Some of these variables include the
comparability of performance relative to coworkers (Harkins & Jackson, 1985), personality traits such as self-esteem (Terborg, Richardson, & Pritchard, 1980), and task-relevant ability or even perceived ability (Kerr & Bruun, 1983).

Finally, through their research on brainstorming, Diehl and Stroebe (1987, 1991) have suggested two additional explanations for lower performance in a group: production blocking and evaluation apprehension. Production blocking does not represent malicious sabotage but rather a mundane fact of interacting, discussion-oriented groups: only one individual at a time may speak. Listening to other speakers distracts or interferes with creative thoughts and cause one to forget, suppress, or prematurely judge unexpressed ideas as redundant or irrelevant. The other explanation, evaluation apprehension, represents an enigma, because fear of negative evaluations of one's ideas from other group members runs counter to brainstorming instructions. Yet, as psychologists have long been aware, evaluation apprehension is generally pervasive in lab-based research (Rosenberg, 1969).

Evaluation apprehension has implications for the work place regardless of whether an individual performs alone or in a group. However, particularly in group situations where SCC is high there is the potential for social comparisons which can engender evaluation apprehension. Thus, evaluation apprehension might stifle performance. However, it could also be argued that evaluation apprehension is a natural feature of the work place and that higher SCC simply increases its salience. Thus, increased evaluation apprehension might motivate everyone to perform at their best. Potential relationships between evaluation apprehension and task performance and test validity were explored in the present study.

The research described above generally suggests that performance is lower when a task is performed in a group versus alone; therefore, it was hypothesized that:
H1. Mean task performance will be higher in the *individual* conditions than the *group* conditions.

The potential effect of the task context on test validity is more ambiguous. Except for Weinstein and Holzbach (1973), no published research was located that has explored the effects of group-related phenomena on test validity. However, hypotheses about test validity are conceptually and statistically related to hypotheses about performance on the criterion. It was hypothesized that mean performance would be higher in the *individual* conditions. Also, the distribution of performance in the *group* conditions will contain more error variance, because subjects are less motivated to work up to their potential. Therefore, it was hypothesized that:

**H2A. Task context will moderate test validity**

and

**H2B. Test validity will be higher in the *individual* conditions than the *group* conditions.**

**Accountability**

Although task performance and test validity were hypothesized to be higher in the *individual* conditions, it is important to consider a variable that could moderate the influence of the task context. Accountability seemed like a natural candidate, because psychological contracts in the work place typically include provisions that individuals and groups are accountable to the organization. In exchange for pay and other valued outcomes, organizations monitor performance to ensure that the underlying agreement is being fulfilled. Indeed, the tenuous nature of employment can be a powerful motivator of performance in the work place.

Accountability has been broadly defined as pressure to justify one's views to others (Tetlock & Boettger, 1989); however, it could reasonably include pressure to justify
other types of performance. The guiding metaphor in the theory of accountability is that people are politicians who seek to maintain the positive regard of various constituencies to whom they feel accountable. Research on organizational behavior and related psychological processes frequently does not represent this critical feature of the work place, and laboratory subjects rarely are accountable for their performance (Tetlock, 1985). Recently, however, accountability has received more attention, primarily in research on judgment and choice but also in research on more objective task performance. Results of that research generally have supported the theory that accountability can reduce or even eliminate biases in judgments and decisions (Tetlock & Boettger, 1989; Tetlock & Skitka, 1989; Tetlock, Skitka, & Boettger, 1989). Other research that has supported the theory of accountability has been reported in domains of particular interest to industrial and organizational psychologists, including personnel selection (Gordon, Rozelle, & Baxter, 1988, 1989), performance appraisal (Klimoski & Inks, 1990), and organizational politics (Fandt & Ferris, 1990).

Accountability might affect the amount of task-relevant effort that an individual exerts in a work situation and that is reflected in a measure of performance. However, compared to research on judgments and decisions, little research has explored the effects of accountability on task performance. Indeed, only two relevant studies were located.

Shalley (1992) recently examined the effects of accountability on task performance in a 2 x 2 laboratory experiment. Undergraduates worked on an in-basket exercise for 30 minutes either alone or in the presence of five others (mere presence manipulation). All subjects were exhorted to generate solutions to as many items as possible; however, half the subjects also were told that their responses would be evaluated by experts and compared to other students (accountability manipulation). As hypothesized, unaccountable subjects were more productive. However, the accountability x mere
presence interaction was significant, and unaccountable subjects who worked alone were the most productive of all.

A critical difference between the Shalley study and the present study is that a group condition was included in the latter but not the former. In the Shalley study, subjects who worked in the presence of others were not told or led to believe that the participants constituted a group. However, in the present study, the procedure was constructed so that half the subjects would feel like they belonged to a group. As mentioned above, the distinction between an individual and group context is critical to understanding the motivational consequences of SCC.

In another study on the effects of accountability on task performance, Yarnold, Mueser, and Lyons (1988) instructed extreme Type A and B undergraduates to tear pages out of phone books. Half the subjects were told that the experimenter would announce the number of pages torn by each individual at the end of the 30-minute work session (accountability manipulation). Each experimental session included three to eight subjects with an approximately equal number of Type A and B students. As hypothesized, Type A’s tore more pages than Type B’s, and accountable subjects tore more pages than unaccountable subjects. However, as hypothesized, the personality type x accountability interaction was significant, and Type A’s in the accountable condition outperformed all other subjects.

Note, however, that accountability and mere presence were confounded in the Yarnold study. Subjects worked in the presence of others, but it was never suggested that they constituted a group. Therefore, it is impossible to determine whether accountability or mere presence was responsible for higher performance. The confusion stems from Yarnold’s ambiguous definition of accountability as public, personal evaluation. Accountability can be personal but not public (e.g., an individual performs alone and is accountable) or public but not personal (e.g., a group is accountable for
their collective efforts but individual members are not accountable). These distinctions were addressed in the present study.

An important contribution of the Yarnold study was that they explored the moderating effect of a personality variable on performance under accountability. Tetlock (1985) suggested that individual differences need to be explicitly considered in research on accountability. Other personality variables such as self-confidence, need for affiliation, and need for approval could be related to motivation experienced by accountable subjects. Thus, more research is needed to explore whether personality variables affect performance under accountability. Another objective of the present study was to address this need.

As the SCC of the work place increases, individuals pose greater contingencies for others with whom they must work closely and frequently, and accountability is likely to be a critical factor in these situations. Thus, accountability manipulations need to be included in lab research on the effects of SCC on task performance and test validity. In the present study, it was felt that brainstorming would be an ideal task with which to explore the effects of accountability, because brainstorming is a cognitive activity that can be performed in an individual or group context. Thus, it was hypothesized that:

**H3. Mean task performance will be higher in the accountable conditions than the unaccountable conditions.**

Note that accountability is meaningful whether an individual performs alone or in a group. Thus, a comprehensive analysis of a situational variable like accountability would require that observations be collected in both individual and group contexts, as in the present study. Furthermore, accountability and the task context might interact if, for example, accountability has a greater impact on subjects who work alone versus in a group. In a group, accountable subjects could perceive that responsibility for justifying performance is diffused. Therefore, it was hypothesized that:
H4A. The main effects of accountability and task context will be qualified by a significant interaction.

and

H4B. Task performance will be higher in the individual-accountable condition than any other condition.

No research on the effects of accountability on test validity was located. However, as mentioned above, hypotheses about test validity are necessarily linked to hypotheses about task performance. It was hypothesized that mean performance would be higher in the accountable conditions. It also seemed likely that the distributions of performance in the unaccountable conditions would contain more error variance, because subjects are not motivated to work up to their potential. Therefore, it was hypothesized that:

H5A. Accountability will moderate test validity

and

H5B. Test validity will be higher in the accountable conditions than the unaccountable conditions.

Finally, it is possible that accountability and the task context simultaneously affect test validity, which would be demonstrated by a significant "moderated moderator" effect. This possibility was explored, but a specific hypothesis was not set forth.

Method

Subjects

The study was piloted with a sample of Rice University undergraduates who participated in exchange for psychology course credit \(N = 64\). Results of the pilot study suggested that the manipulations needed to be stronger to more clearly produce the intended psychological effects. Thus, the manipulations were strengthened, and the revised study was conducted with a sample of University of Houston-Downtown
undergraduates who participated in exchange for social science course credit ($N = 64$ in revised study, too). All data refer to the latter sample.

Subjects were randomly assigned to one cell of a 2 (accountable or unaccountable) x 2 (individual or group task context) design. No fewer than two subjects and no more than five subjects participated simultaneously. Subjects who participated simultaneously were initially seated at separate tables in the same room. A group condition was run only if at least three subjects were present. Thus, the mean number of subjects per session was higher in the group conditions (4.25, $SD = 0.7$) than the individual conditions (3.3, $SD = 0.9$). The mean number of subjects per session was similar in the accountable conditions (3.9) and the unaccountable conditions (3.7). Because of these constraints, cell size ranged from 14 to 17, but all analyses were adjusted accordingly. Females constituted 62.5% of the sample ($n = 40$). The percentage of females per cell ranged from 59% to 64%.

Procedure

The five-part procedure required one hour. Subjects first completed the Alternate Uses test, a paper-and-pencil measure of creative thinking ability. Score on the Alternate Uses test was the predictor. Next, subjects completed a warm-up exercise called the Vegetable Problem either alone (individual conditions) or as a group (group conditions). Subjects then provided answers to the Thumbs Problem, a brainstorming task, under different conditions. Performance on the Thumbs Problem was the criterion measure. Subjects then completed a questionnaire to check the manipulations and measure evaluation apprehension. Finally, after the manipulations were exposed, subjects completed the Adjective Check List, a broadband personality measure included to explore the effects of personality traits.

Measures
**Ability Measure.** The primary predictor was the Alternate Uses (AU) test. The AU test measures spontaneous flexibility, a type of creative thinking defined as the ability to spontaneously produce ideas in response to objects or other ideas (Christensen, Guilford, Merrifield, & Wilson, 1960). The AU test was developed at the University of Southern California to study the structure of intellect in the Aptitudes Research Project (Guilford & Hoepfner, 1971). The AU test has been widely used in research on relationships between intelligence and creativity, and many studies have contributed evidence of its construct validity. Consistent with the general thesis of the Aptitudes Research Project, correlations between the AU test and traditional intelligence tests typically are low, around .30 (Guilford, Christensen, Merrifield, & Wilson, 1978).

The AU test is available in two parallel forms, B and C, each of which consists of the names of six common objects, the common use for each object, and blank lines to list up to six other uses for each object or parts of the object. For example, a newspaper typically is used for reading but also can be used to swat flies, start a fire, pack boxes, etc. Score on the AU test is the total number of nonredundant responses listed across objects in eight minutes. Thus, total score can range from 0 to 36 (Guilford et al. 1978). The score is a measure of spontaneous flexibility or, more technically, divergent production of semantic classes. Responses are evaluated on the basis of reasonable feasibility and not creativity, because the test was not specifically designed to measure originality or, more technically, divergent production of semantic transformations (Guilford et al. 1978). Nevertheless, Parnes and Meadow (1959) reported that total responses and high quality responses were correlated around .70.

In the present study, subjects completed Form B of the AU test while facing away from each other. The test was scored according to rules provided in the manual. Subjects did not receive feedback on their scores.
Warm-Up Exercise. Subjects then completed the Vegetable Problem (VP), a warm-up exercise designed for this study to introduce the task context manipulation. Subjects in the individual conditions continued to face away from each other. Subjects in the group conditions were instructed to gather around a large, central table, and the experimenter read aloud these instructions:

The purpose of the rest of this study is to explore how people work together in groups to solve creative problems. All of you have been assigned to the same group; the code letter for your group is "C". I don't know whether you know each other or not, so to introduce everyone to the group, let's go around the table and have each of you tell us your name, year in school, and major or area that you're studying. (Pause). OK, as I said, all of you are members of Group C, and you will work together on several exercises.

Subjects completed the VP individually (individual conditions) or as a group (group conditions). The experimenter read the directions aloud, but the next to last sentence was omitted in the individual conditions:

This next exercise is called the Vegetable Problem. We all know that young children sometimes refuse to eat vegetables at mealtimes. I want you to think of as many different ways as possible to get children to eat vegetables. Please write each idea you have on a separate card. Don't worry about spelling or grammar. Discuss the problem among yourselves, but have one person, it doesn't matter who, do all the writing for your group. You will have only a few minutes for this exercise.

In all conditions, the experimenter left the room and returned after four minutes to collect answers, which were immediately placed in a large envelope that was set aside.

Criterion Measure. The criterion measure was a brainstorming task called the Thumbs Problem (TP) (Bouchard & Hare, 1970; Gallupe et al., 1991). The
experimenter read aloud an introduction, but the last sentence was omitted in the individual conditions:

Now it is time for the next exercise, which is called the Thumbs Problem. Here are the directions and blank cards for your answers. Unlike the Vegetable Problem, the Thumbs Problem does not involve any discussion, so please don't talk about it among yourselves.

Each subject was given a sheet that included general instructions plus supplemental instructions that elaborated the task context manipulation and introduced the accountability manipulation. The general instructions were the same for all subjects:

This exercise is called the "Thumbs Problem". Imagine what would happen if everybody suddenly had one extra thumb on each hand. These new thumbs are built like the old ones but are on the other side of the hand. These new thumbs face inward, so that they can press against fingers just like the old thumbs do. Here is a question for you: WHAT BENEFITS OR DIFFICULTIES WILL OCCUR WHEN PEOPLE HAVE THESE NEW THUMBS? There are no right or wrong answers. Your job is to write as many different answers as possible, even wild answers, as long as they make sense. Write each answer on a separate card. Don't worry about grammar or spelling.

The accountability manipulation was introduced by supplemental instructions appended to the general instructions. (Some aspects of the accountability manipulation were based on ideas put forth by Hayes [1990] in his research on assessment centers.) In the unaccountable conditions, the supplemental instructions stated that answers were anonymous. Subjects were not to write their name (individual and group conditions) or group code letter (group condition) on the instruction sheet or answer cards. (Later, answers were attributed to individuals based on differences in handwriting.) The instructions also stated that a subject's answers would be mixed with answers collected
from 50 other students (individual condition) or 10 other groups of students (group condition). Finally, the experimenter brought a large envelope stuffed with cards into the room at the beginning of the Thumbs Problem. As indicated in the instructions, when time was up the experimenter told subjects to place all their used and unused answer cards in the envelope, which was set aside.

In the accountable conditions, subjects were instructed to write their name (individual condition) or group code (group condition) on each answer card. The instructions also stated that when time was up the experimenter would count the total number of answers provided by the subject (individual condition) or group (group condition). Then, the experimenter would sit down with the individual (individual condition) or group (group condition) to discuss and evaluate score on the Thumbs Problem. If the score was high compared to other University of Houston students, the evaluation would end. Otherwise, the individual (individual condition) or group (group condition) would have to justify the score and explain why not many answers were provided relative to other students in the University. Also, subjects in both the individual and group conditions were asked to provide their signature, so that the individual score (individual condition) or group score (group condition) could be released to other researchers. Finally, the instructions indicated that the experimenter would call subjects within two days to further discuss the individual score (individual condition) or group score (group condition) if it was low. Subjects were asked to provide their phone number and a convenient time for the experimenter to call.

The time limit for the Thumbs Problem was extended from 20 to 22 minutes in the accountable conditions, so that subjects would have time to read and process the supplemental instructions. As indicated in those instructions, when time was up the experimenter collected answers from each subject (individual-accountable condition) or asked subjects to deposit their answers in an empty box marked "Group C" that was in
the middle of the table (group-accountable condition). The purpose of the collection box was to increase the likelihood that subjects in the group-accountable condition believed that the experimenter was interested in the group's total output and not individual performance. The complete text of the supplemental instructions for each condition is shown in Appendix D.

Score on the TP is the number of nonredundant ideas written. Quality of responses typically is not evaluated, because all other scores that can be derived from this type of data correlate highly with quantity (Bouchard & Hare, 1970). This is a common finding in lab research on brainstorming (Bouchard, 1972). For example, Diehl and Stroebe (1987, 1991) concluded that the additional information gained by evaluating quality, originality, or feasibility of ideas was not worthwhile. In the present study, responses to the TP were scored according to rules described by Bouchard and Hare (1970). Subjects did not receive feedback on their scores.

Post-Performance Questionnaire. After the Thumbs Problem, subjects completed a questionnaire that included 12 manipulation checks and four questions designed to measure evaluation apprehension. The instructions assured subjects that their responses would be confidential and that the experimenter would not look at their responses until after they left. Six items were designed to assess the task context manipulation. Five items were rated on a 1 (strongly disagree) to 5 (strongly agree) scale:

1. I feel like we (the other students in this room and I) are a team.
2. I feel like I know more about the other students in this room than I did when I got here.
3. If we (the other students in this room and I) are asked to work together on a creative exercise, I know we will work well together.
4. I felt like the other students in this room were counting on me to do a good job on the Thumbs Problem.
If I sign up for another psychology study, I would like to work with these same students again. The last item was a Yes/No question:

Were you told that you are a member of a group?

Six items were designed to assess the accountability manipulation. Five items were rated on a 1 (strongly disagree) to 5 (strongly agree) scale:

I felt pressure to write a lot of answers to the Thumbs Problem.

I was worried that if I did not write a lot of answers to the Thumbs Problem, the experimenter would expect me to explain why.

I tried hard to write a lot of answers to the Thumbs Problem.

I felt it was important that I write a lot of answers to the Thumbs Problem.

I feel that the experimenter will be satisfied with the number of answers I wrote to the Thumbs Problem.

The last item was a Yes/No question:

Did the instructions for the Thumbs Problem say the experimenter would meet with you (or your group) to discuss scores on the Thumbs Problem?

In addition to the manipulation checks, the post-performance questionnaire included four items to measure evaluation apprehension. Evaluation apprehension was hypothesized to be negatively related to performance (Diehl & Stroebe, 1987; Tetlock, 1985). Each item was rated on a 1 (strongly disagree) to 5 (strongly agree) scale:

I am worried that someone will analyze or criticize the content of my answers to the Thumbs Problem.

I am working very hard in this study, because I want to impress the experimenter.

I am worried that the experimenter is trying to "figure something out" about my personality in this study.
I found it hard to concentrate during the Thumbs Problem, because I was trying to
guess the "real" purpose of this study.

After the post-performance questionnaire, the manipulations were exposed.
Subjects in the **accountable** conditions received a note stating that contrary to the
instructions for the TP, there would not be a meeting with the experimenter. Subjects in
the **group** conditions received a note stating that they would not participate in any more
exercises with other students.

**Personality Measure.** If ability test validity is low when subjects are unaccountable
and/or in a group, personality variables might be better predictors of performance.
Thus, the **Adjective Check List (ACL)** (Gough, 1952), a personality measure, was
included as the last step in the procedure to explore whether personality traits are valid
predictors of performance in any condition where the ability measure was not valid.

The ACL is a self-report measure that takes approximately 20 minutes to complete
(Gough, 1952). Examinees quickly read a list of 300 adjectives and check those
considered to be self-descriptive. Based on analyses of the pilot data, three ACL scales
were identified as potentially useful: Need for Affiliation, Need for Abasement, and
Critical Parent. Two additional scales, Need for Achievement and Need for Exhibition,
were explored on a post-hoc basis to interpret specific results. All five scales were
hand-scored according to instructions provided in the manual. Need for affiliation was
hypothesized to be positively related to performance (Tetlock, 1985), but no specific
hypothesis for the other ACL scales was set forth.

**Analyses**

Planned comparisons (t tests) and an analysis of variance (ANOVA) were
computed to compare performance on the Thumbs Problem among conditions. To
explore the effect of individual differences in ability on the results, an analysis of
covariance (ANCOVA) was computed. Pearson correlations were computed to
compare test validity among conditions. Results of the ANCOVA were used to
determine whether accountability and/or task context moderated test validity.

Results

Manipulation Checks

Task Context Results of the task context manipulation checks are shown in Table
8. The mean for group subjects was significantly greater than the mean for individual
subjects on each item and the average of all five items (coefficient alpha = .78). For
group subjects, four items rounded to "somewhat agree" (4 on the 1-to-5 scale). For
individual subjects, all items rounded to either "somewhat disagree" (2) or "neither agree
nor disagree" (3). Finally, 85% of subjects in the group conditions affirmed that they
had been told they were members of a group. In the individual conditions, 18% of
subjects reported that they had been told they were members of a group.

Accountability Results of the accountability manipulation checks are shown in
Table 9. The mean for accountable subjects was significantly greater than the mean for
unaccountable subjects on two items ("I was worried..." and "I tried hard...") and the
average of all five items (coefficient alpha = .61; if the last item is excluded from the
average, coefficient alpha = .69). On two other items ("I felt pressure..." and "I felt it
was important..."), the difference between the mean for accountable subjects and the
mean for unaccountable subjects approached traditional levels of significance (p < .10
and p < .07, respectively). Finally, 80% of subjects in the accountable conditions
affirmed that the instructions had stated that the experimenter would meet with them to
evaluate their score (whether individual or group) on the Thumbs Problem. In the
unaccountable conditions, 3% of subjects reported that the instructions had stated that
the experimenter would meet with them.

Ability Test Scores
Table 8

Means and Standard Deviations of Task Context Manipulation Checks

<table>
<thead>
<tr>
<th>Subjects in Group Conditions</th>
<th>Subjects in Individual Conditions</th>
<th>t test</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n = 34)</td>
<td>(n = 30)</td>
<td></td>
</tr>
</tbody>
</table>

I feel like we (the other students in this room and I) are a team.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.97</td>
<td>2.60</td>
</tr>
<tr>
<td></td>
<td>(0.97)</td>
<td>(1.25)</td>
</tr>
</tbody>
</table>

I feel like I know more about the other students in this room than I did when I got here.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.85</td>
<td>2.17</td>
</tr>
<tr>
<td></td>
<td>(1.37)</td>
<td>(1.32)</td>
</tr>
</tbody>
</table>

If we (the other students in this room and I) are asked to work together on a creative exercise, I know we will work well together.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4.24</td>
<td>3.27</td>
</tr>
<tr>
<td></td>
<td>(0.74)</td>
<td>(0.78)</td>
</tr>
</tbody>
</table>

I felt like the other students in this room were counting on me to do a good job on the Thumbs Problem.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.00</td>
<td>2.00</td>
</tr>
<tr>
<td></td>
<td>(1.28)</td>
<td>(0.98)</td>
</tr>
</tbody>
</table>

If I sign up for another psychology study, I would like to work with these same students again.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.76</td>
<td>3.07</td>
</tr>
<tr>
<td></td>
<td>(0.74)</td>
<td>(0.74)</td>
</tr>
</tbody>
</table>

AVERAGE OF ITEMS

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.76</td>
<td>2.62</td>
</tr>
<tr>
<td></td>
<td>(0.67)</td>
<td>(0.65)</td>
</tr>
</tbody>
</table>

***p < .001.
Table 9

Means and Standard Deviations of Accountability Manipulation Checks

<table>
<thead>
<tr>
<th>Subjects in Accountability Conditions</th>
<th>Subjects in Unaccountable Conditions</th>
<th>t test</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n = 31)</td>
<td>(n = 33)</td>
<td></td>
</tr>
<tr>
<td>I felt pressure to write a lot of answers to the Thumbs Problem.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.81</td>
<td>3.27</td>
<td>1.70(^a)</td>
</tr>
<tr>
<td>(1.11)</td>
<td>(1.38)</td>
<td></td>
</tr>
<tr>
<td>I was worried that if I did not write a lot of answers to the Thumbs Problem, the experimenter would expect me to explain why.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.39</td>
<td>1.82</td>
<td>5.26***</td>
</tr>
<tr>
<td>(1.28)</td>
<td>(1.10)</td>
<td></td>
</tr>
<tr>
<td>I tried hard to write a lot of answers to the Thumbs Problem.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.71</td>
<td>4.18</td>
<td>2.47*</td>
</tr>
<tr>
<td>(0.59)</td>
<td>(1.04)</td>
<td></td>
</tr>
<tr>
<td>I felt it was important that I write a lot of answers to the Thumbs Problem.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.23</td>
<td>3.73</td>
<td>1.88(^b)</td>
</tr>
<tr>
<td>(0.84)</td>
<td>(1.23)</td>
<td></td>
</tr>
<tr>
<td>I feel that the experimenter will be satisfied with the number of answers I wrote to the Thumbs Problem.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.81</td>
<td>3.79</td>
<td>0.07</td>
</tr>
<tr>
<td>(0.95)</td>
<td>(1.02)</td>
<td></td>
</tr>
<tr>
<td>AVERAGE OF ITEMS (including last item)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.99</td>
<td>3.36</td>
<td>3.90***</td>
</tr>
<tr>
<td>(0.54)</td>
<td>(0.73)</td>
<td></td>
</tr>
<tr>
<td>AVERAGE OF ITEMS (excluding last item)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.03</td>
<td>3.25</td>
<td>4.11***</td>
</tr>
<tr>
<td>(0.60)</td>
<td>(0.89)</td>
<td></td>
</tr>
</tbody>
</table>

\(^{a}p < .10\). \(^{b}p < .07\).

\(^{*}p < .05\). \(^{***}p < .001\).
Maximum possible score on the Alternate Uses (AU) test is 36. In this study, scores ranged from 3 to 29. Compared to norms for university freshmen (Guilford et al., 1978), the grand mean in this study (14.6) was slightly low, but the standard deviation (5.8) was similar. Mean AU score was not expected to significantly differ among conditions, because subjects were randomly assigned to conditions. To test this assumption, an ANOVA was conducted in which AU score was regressed on the accountability factor, the task context factor, and the accountability x task context interaction. The overall F was not significant suggesting that AU scores did not significantly differ among conditions (F [3, 60] = 0.48, ns).

Performance on the Criterion

Scores on the Thumbs Problem (TP) ranged from 3 to 38. Mean score is shown by condition in Table 10. The mean for individual subjects (17.3) was not significantly different from the mean for group subjects (14.4); therefore, hypothesis 1 was not supported (t [62] = 1.54, ns). The mean for accountable subjects (17.1) was not significantly different from the mean for unaccountable subjects (14.5); therefore, hypothesis 3 was not supported (t [62] = 1.38, ns). The task context x accountability interaction, which was tested in an analysis of variance (ANOVA), was not significant (F [1, 60] = 0.18, ns); therefore, hypotheses 4A and 4B were not supported.

To explore whether individual differences in ability affected the results, a factorial analysis of covariance (ANCOVA) was conducted in which performance on the TP was regressed on: (a) AU test score (i.e., the ability measure was the covariate), (b) coded vectors that represented accountability, task context, and the accountability x task context interaction, and (c) vectors that represented the product of the covariate and each coded vector (Pedhazur, 1982). Results are shown in Table 11.

The three-way, AU x accountability x task context interaction, which was the last term entered into the regression, was significant (F [1, 56] = 6.21, p < .05). This
Table 10

**Mean Score on the Thumbs Problem by Condition**

<table>
<thead>
<tr>
<th>Task Context</th>
<th>Unaccountable</th>
<th>Accountable</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>M</td>
<td>n</td>
</tr>
<tr>
<td><strong>Individual</strong></td>
<td>16</td>
<td>16.9</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(8.8)</td>
<td></td>
</tr>
<tr>
<td><strong>Group</strong></td>
<td>17</td>
<td>12.2</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(6.5)</td>
<td></td>
</tr>
<tr>
<td><strong>Combined</strong></td>
<td>33</td>
<td>14.5</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(7.9)</td>
<td></td>
</tr>
</tbody>
</table>

**Note.** The standard deviation is shown in parentheses.
Table 11

Regression of Thumbs Problem Performance on the Alternate Uses Test, Task Context, and Accountability

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU Test</td>
<td>594.5</td>
<td>1</td>
<td>14.49***</td>
</tr>
<tr>
<td>Task Context</td>
<td>126.6</td>
<td>1</td>
<td>3.09</td>
</tr>
<tr>
<td>Accountability</td>
<td>208.5</td>
<td>1</td>
<td>5.08*</td>
</tr>
<tr>
<td>Task Context x Account</td>
<td>31.0</td>
<td>1</td>
<td>0.76</td>
</tr>
<tr>
<td>AU x Task Context</td>
<td>31.8</td>
<td>1</td>
<td>0.78</td>
</tr>
<tr>
<td>AU x Account</td>
<td>129.6</td>
<td>1</td>
<td>3.16</td>
</tr>
<tr>
<td>AU x Task Context x Account</td>
<td>255.0</td>
<td>1</td>
<td>6.21*</td>
</tr>
</tbody>
</table>

Account

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Model</td>
<td>1377.0</td>
<td>7</td>
<td>4.79***</td>
</tr>
<tr>
<td>Error</td>
<td>2298.0</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3675.0</td>
<td>63</td>
<td></td>
</tr>
</tbody>
</table>

*Note. Total $R^2 = .375$. Type I (incremental) SS are shown.

*p < .05. ***p < .001.
result suggested that performance significantly varied as a joint function of ability and experimental factors. To further explore the interaction, subjects in each cell of the design were divided into low and high ability groups based on the median AU score (13.5). Thus, each subject was classified into one of eight subcells, the size of which ranged from six to nine subjects. Mean performance on the TP is shown in Figure 2 for low and high ability subjects by level of accountability and type of task context.

For low ability subjects, the accountability x task context interaction was not significant ($F[1, 56] = 1.04, ns$). (As suggested by Pedhazur [1982], each effect was assessed using the error term from the overall analysis of the factorial design.) The simple main effect of task context was significant ($F[1, 56] = 8.08, p < .01$), but the simple main effect of accountability was not significant ($F[1, 56] = 1.68, ns$). Combined over level of accountability, mean performance on the TP was significantly higher for low ability subjects in the individual conditions (16.5) than the group conditions (10.1) (Tukey's HSD = 3.8, $p < .05$). Thus, hypothesis 1 was supported for low ability subjects.

For high ability subjects, the accountability x task context interaction was significant ($F[1, 56] = 6.45, p < .01$). The simple main effect of accountability was significant for high ability subjects in the group conditions ($F[1, 56] = 6.44, p < .01$) but not high ability subjects in the individual conditions ($F[1, 56] = 1.37, ns$). Mean performance of high ability subjects in the group-accountable condition (22.11) was significantly higher than mean performance of high ability subjects in the group-unaccountable condition (14.44) (Tukey's HSD = 7.48, $p < .05$). Thus, hypothesis 4A was supported for high ability subjects, but hypothesis 4B was not supported, because the pattern of means was not as hypothesized.
Figure 2. Mean Performance on the Thumbs Problem for Low and High Ability Subjects by Level of Accountability and Type of Task Context
Test Validity

The last three terms entered into the ANCOVA (see Table 11) significantly added to the proportion of variance accounted for by the first four terms ($R^2$ first four = .261; $R^2$ last three = .113; $F$ increment [3, 56] = 3.38, $p < .05$). Therefore, the slopes of the predictor-criterion regression lines were not homogenous in the four cells of the design, suggesting that accountability and task context both moderated test validity. Thus, hypotheses 2A and 5A were supported. The regression lines for all four conditions are shown in Figure 3. The regression equation and significance test of the $b$ weight of each regression line is shown in Appendix E.

The criterion-related validity of the AU test (i.e., $r_{AU,TP}$) is shown by condition in Table 12. The pattern of validity in the four cells of the design was indicative of the moderated moderator effect described above. Overall, validity in the group conditions (.48) was not significantly different from validity in the individual conditions (.31); therefore, hypothesis 2B was not supported ($z = 0.78$, ns). Overall, validity in the accountable conditions (.35) was not significantly different from validity in the unaccountable conditions (.55); therefore, hypothesis 5B was not supported ($z = 0.96$, ns). Validity in the accountable-individual cell (.18) was significantly lower than validity in the unaccountable-individual cell (.68) ($z = 2.44$, $p < .05$) and the accountable-group cell (.60) ($z = 2.18$, $p < .05$).

Effects of Evaluation Apprehension

The four items on the post-performance questionnaire that were designed to measure evaluation apprehension were averaged (EA) (coefficient alpha = .81). The grand mean of EA was 2.22 ($SD = 1.03$) on a 1-to-5 scale. Mean EA for low ability subjects ($2.65, SD = 1.10$) was significantly higher than mean EA for high ability subjects ($1.80, SD = 0.76$) ($t_{62} = 3.61, p < .001$). Other analyses were conducted to explore whether EA: (a) varied as a function of accountability and/or task context, (b)
Figure 3. Predictor-Criterion Regression Line for Each Cell
(A = Accountable, U = Unaccountable, I = Individual, G = Group)
Table 12

Pearson Correlation Between Score on the Alternate Uses Test and Performance on the Thumbs Problem by Condition

<table>
<thead>
<tr>
<th>Accountability</th>
<th>Unaccountable</th>
<th>Accountable</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Context</td>
<td>n</td>
<td>r</td>
<td>n</td>
</tr>
<tr>
<td>Individual</td>
<td>16</td>
<td>.68**</td>
<td>14</td>
</tr>
<tr>
<td>Group</td>
<td>17</td>
<td>.39</td>
<td>17</td>
</tr>
<tr>
<td>Combined</td>
<td>33</td>
<td>.55***</td>
<td>31</td>
</tr>
</tbody>
</table>

**p < .01. ***p < .001.
interacted with accountability and/or task context to affect performance, or (c) moderated test validity. None of the results was statistically significant.

Responses to the Adjective Check List

In the population of normal adults, ACL scales are standardized to a mean of 50 and a standard deviation of 10. Results in the present study were close to those values (Abasement $M = 47.1$, $SD = 9.9$; Critical Parent $M = 50.1$, $SD = 10.3$; Affiliation $M = 50.1$, $SD = 9.1$; Exhibition $M = 52.6$, $SD = 8.6$; Achievement $M = 49.2$, $SD = 9.8$). No ACL scale was a valid predictor of performance on the TP overall (Abasement $r = -.01$, Parent $r = .04$, Affiliation $r = -.10$, Exhibition $r = -.04$, Achievement $r = -.01$, each ns) or in any condition. Other analyses were conducted to explore whether any ACL scale score: (a) varied as a function of accountability and/or task context, (b) interacted with accountability and/or task context to affect performance, or (c) moderated test validity. None of the results was statistically significant. Finally, low and high ability subjects did not differ on any ACL scale.

Discussion

Effectiveness of Experimental Manipulations

Responses to the post-performance questionnaire suggested that the task context manipulation was highly effective. Additionally, the experimenter observed that the personal introductions and warm-up exercise served as effective "ice breakers" in the group conditions. Groups laughed and joked during the warm-up exercise, whereas subjects in the individual conditions did not appear to enjoy it as much.

Responses to the post-performance questionnaire also suggested that the accountability manipulation generally produced the intended psychological effects. Two accountability manipulation checks (and the average of all checks) were successful, and two were marginally successful; only one check was unsuccessful. These results were
encouraging given that accountability manipulation checks are notoriously difficult to formulate.

**Performance on the Criterion**

Regardless of the success of the experimental manipulations, none of the hypotheses for performance was supported. However, the hypotheses were not supported, because differences in ability obscured the effects of the experimental factors. Indeed, when ability was treated as a covariate, results suggested a complex, three-way interaction of ability, accountability, and task context. To interpret the interaction, subjects were divided into low and high ability groups, and the effects of accountability and task context were reassessed.

Results for low ability subjects were unambiguous. The simple main effect of task context was significant, and mean performance was significantly lower in the group conditions than the individual conditions. Thus, low ability subjects demonstrated a classic social loafing effect whereby they exerted less effort when working in a group than when working alone, because they were relatively dispensable in the group context (viz., Harkins & Jackson, 1985; Kerr & Bruun, 1983; Latane, Williams, & Harkins, 1979). This effect also can be explained by a combination of efficacy and expectancy theories (e.g., Sanna, 1992). Low ability is associated with an expectation of low self-efficacy which produces an expectation of negative evaluation and lower social performance.

Sanna's (1992) research indicated that social facilitation and social loafing are complementary processes that can be conceptualized in terms of self-efficacy. His research showed that, consistent with social facilitation theory, laboratory subjects who had a high efficacy expectation performed better in a group setting (where there also was the possibility of evaluation) than when alone. However, the research also showed that, consistent with social loafing theory, subjects with a high efficacy expectation
performed worse when the possibility of evaluation was absent than when it was present. Sanna (1992) concluded that the key theoretical construct linking these results is perceived efficacy.

Of course, subjects in the present study did not receive feedback on their ability test scores and had no objective way of determining their relative ability. Nonetheless, the experimenter observed that subjects were sensitive to the issue of ability. As subjects left the lab, the most frequently expressed comments concerned ability, for example, "Sorry, I didn't do well—I'm not very creative" or "I always do well on stuff like this". Also, recall that mean evaluation apprehension was significantly higher among low ability subjects than high ability subjects suggesting that subjects made judgments of their relative ability.

For high ability subjects, the accountability x task context interaction was significant, and the simple main effect of accountability was significant for high ability subjects in the group conditions but not the individual conditions. Consistent with the simple main effect, mean performance of high ability subjects in the group conditions was significantly greater in the accountable condition than the unaccountable condition. Again, Sanna's (1992) theory helps explain the results. High ability is associated with an expectation of high self-efficacy which produces an expectation of positive evaluation and, when paired with a high outcome expectancy (i.e., accountability), leads to even higher social performance. (Note that because the simple main effect of accountability was not significant for high ability subjects in the individual conditions, the apparent downward tilt of the corresponding line in Figure 2 does not represent a significant effect.)

These results suggest that efficacy theory (e.g., Sanna, 1992) has an interesting implication for accountability theory (e.g., Tetlock, 1985). With a relatively simple task like the Thumbs Problem, accountability theory predicts that subjects will take the path
of least resistance and perform as well as possible. However, results of the present study suggest that the effect of accountability is complicated by expectations produced by self-efficacy and a social context. Future research will need to explore these relationships more closely.

Finally, the importance of ability as a nonmanipulated variable in this study suggests that the effects of ability on brainstorming performance be more closely scrutinized in future research. Graham and Dillon (1974) reported that "supergroups" composed of individuals who scored high on individual brainstorming tasks performed much better than typical brainstorming groups in which ability is heterogeneous. One way to more closely explore the effects of ability on brainstorming performance would be to have all subjects brainstorm individually and in a group (in a counterbalanced order) and explore the relationships between ability, motivational factors, and performance in the two contexts. The ability composition of groups could be directly manipulated (e.g., homogeneous versus heterogeneous ability), and then subjects could be informed of this situation. One hypothesis is that loafing would be less prevalent among low ability subjects if they knew they were in a homogeneous group. Of course, related hypotheses concerning test validity would be more difficult to test in homogeneous groups due to an inherent restriction in the range of ability.

Test Validity

Results of the ANCOVA suggested that task context and accountability both significantly moderated test validity. As shown in Table 12, validity was significant in the individual condition but only in the accountable condition, and validity was significant in the group condition but only in the accountable condition.

How can these results be explained? The variance of the criterion in the two cells in which validity was nonsignificant was approximately half that of the other two cells. Experimental factors could have caused this difference in the variability of performance.
That is, if accountability or the task context reduces the range of performance in a work situation, test validity is likely to decline.

The difference between variability in the two group conditions can be explained by accountability theory. The variability of performance in the accountable-group condition was greater, because the motivational consequences of accountability in a group context produced a truer distribution of performance. Accountability theory would suggest that motivation was relatively lower in the unaccountable-group condition; thus, it was not surprising that the variability of performance was lower.

It is not immediately clear why the effect of accountability on the variability of performance would be opposite in individual conditions compared to group conditions. However, Shalley's (1992) research (described above) provides a potential explanation. One result of that research was that accountability can have a detrimental effect on the distribution of performance in an individually-oriented context. Therefore, one conclusion of the present research is that accountability obscures some true variability in the distribution of performance when work is performed alone, which lowers test validity, whereas the opposite effect occurs when work is performed in a group.

**State and Trait Variables**

Evaluation apprehension was higher among low ability subjects than high ability subjects. As the SCC of jobs increases, it is likely that the effects of evaluation apprehension in the workplace will become more salient. As roles become more complex and interdependent and workers pose greater contingencies for each other, individual differences in evaluation apprehension could affect performance. Hence, the specter of increased evaluation apprehension is one reason why interpretations of task performance and test validity are likely to become more complex in the work place.

The criterion-related validity of personality measures, in general, is likely to become a more important issue in the work place. Even though Adjective Check List (ACL)
scales were not valid predictors of performance on the Thumbs Problem in the present study, additional research is needed to explore personality measures in this type of research. One potential reason why personality measures were not useful in the present study is that there were problems with having the ACL as the last step in this procedure (e.g., fatigue, cumulative hypothesis guessing). However, it was felt that having the ACL earlier in the procedure would have induced massive evaluation apprehension and biased other measures, because subjects would have interpreted the experiment as a study of personality. Future research could address the question of personality variables as in the Yarnold study. That is, personality measures could be obtained beforehand and then only extreme types included in the research design.

Another potential reason why the personality measures were not valid in the present study is that performance on the Thumbs Problem was unidimensional. Criteria in the workplace are more likely to be multidimensional and include measures of interpersonal effectiveness. Thus, more research is needed to explore whether SCC moderates validities of personality tests as well as ability tests. This need will become especially acute if organizations choose to more often depend on personality tests in human resource management for a wider range of jobs.

Conclusions

Study 1 (an analysis of archival field data) and Study 2 (a laboratory experiment) were concurrently conducted to explore the moderating effects of social-collaborative complexity (SCC) on test validity from different orientations. Results of Study 1 suggested that higher SCC was associated with higher GATB validity, independent of the more predictable moderating effects of other job demands. An implication of Study 1 was that future research needs to address the effects of SCC on task performance as well as test validity, because the two issues are closely related. An advantage of the experimental paradigm of Study 2 was that it allowed simultaneous tests of the effects of
accountability and the task context on performance and validity in a controlled setting. Results of Study 2 suggested that accountability and the task context both moderated test validity.

Taken together these results suggest that SCC moderates cognitive ability test validities and that higher SCC is associated with higher validity. In the field study, measures of SCC (PAQ dimensions 17 and 21) was positively and significantly correlated with cognitive GATB validities even when the more predictable moderating effects of cognitive job demands were partialled out. In the lab study, test validity was significant in the group conditions when groups were accountable but not unaccountable, and the former represents a socially more complex situation. Thus, as the SCC of jobs increases, there is reason to be optimistic that cognitive ability test validities actually will increase.

Results of the present research also suggest that results of the Weinstein and Holzbach (1973) study are too pessimistic. Recall that higher SCC, operationalized as task flow interdependence, was associated with lower clerical ability test validity in their study. However, as discussed above, their results were highly predictable based on the particular types of interdependence (i.e., pooled and sequential) they studied. At a minimum, results of their research and the present research suggest that organizations need to attend to the potential moderating effects of SCC on ability test validities, especially when performance occurs in a group context. Furthermore, given that the SCC of jobs is rapidly increasing, psychologists cannot afford to wait another two decades before this issue is addressed again.

Results of Study 1 also suggest that a measure of SCC (PAQ dimension 20, exchanging job-related information) was negatively related to all three GATB psychomotor validities even when the moderating effects of physical job demands were partialled out. These results suggest that psychomotor ability test validities will decline
as SCC increases in the work place. This issue will be especially critical in situations where physically-oriented jobs are redesigned to be more collaborative and team-oriented (e.g., the auto industry). The increased SCC of these situations could reduce psychomotor test validities either because workers have fewer opportunities to express relevant behaviors or because these behaviors are obscured in cooperative work arrangements. Indeed, psychologists will continue to be challenged by the need to develop relevant, reliable criterion measures in situations where performance is increasingly group-oriented.

Results support the James et al. (1992) argument that moderators of validity are more likely to be identified if research proceeds on the basis of theory and utilizes a combination of archival data and proactive research designs. Other research reviewed above provided additional evidence that situational moderators of validity can have important practical implications in real employment settings (e.g., Brown, 1981; Schmitt et al., 1990). Furthermore, ignoring situational moderators of validity ultimately might represent a violation of equal employment opportunity laws (Weinstein & Holzbach, 1973).

A limitation of the lab study is that the criterion measure (i.e., the Thumbs Problem) was unidimensional. Criteria in the work place typically are multidimensional; therefore, future research on the effects of SCC on task performance and test validity will need to examine more complex criteria. It is possible that different aspects of a complex task (e.g., an in-basket exercise or a group discussion) are differentially useful as criteria (from a test validation perspective) in a socially complex work situation. Thus, additional research is needed to expand the experimental paradigm explored here to more complex tasks and more elaborate criteria.

Another limitation of the lab study is that only one type of group task, an additive task (Steiner, 1972), was explored. Additional research is needed to explore other types
of group tasks, such as conjunctive and disjunctive tasks, because the issue of individual differences in ability might be even more important in situations where the success of the group rests more heavily on less or more capable members.

A limitation of the field study is that, except for a measure of people-oriented job demands obtained from the Dictionary of Occupational Titles, measures of SCC were obtained from one instrument, the Position Analysis Questionnaire. As mentioned above, at least two other widely used and well researched job analysis instruments, the Job Diagnostic Survey and the Job Characteristics Inventory, contain scales that measure the degree to which a job requires working with other people. However, the measures of SCC available from the PAQ are more elaborate, even though they are somewhat confounded with supervisory responsibilities and level in an organization. Clearly, more research is needed to develop measures that are specifically targeted at elements of SCC such as task interdependence and job-required, interpersonal relationships.

Another limitation of the field study is that only ability test validities were explored. The several advantages of exploring the GATB were discussed above, but the GATB only measures relatively basic abilities and does not measure higher-level abilities such as oral communication and interpersonal effectiveness. Therefore, this type of research needs to be extended to more complex predictors of performance, such as simulation exercises used in managerial assessment centers (see Thornton, 1992), that measure higher-level skills and abilities.

The increasing social-collaborative complexity of the work place will no doubt continue to provide psychologists with many opportunities and challenges for research on diverse human resource issues, only one of which is moderators of ability test validity. Nonetheless, given the critical relationship between ability tests and organizational success, the issues explored here deserve a place of prominence on the
psychological research agenda for the work place. The present research undoubtedly raised more questions than it addressed, but the results hopefully will help move collective knowledge an incremental step ahead.
References


Appendix A—Position Analysis Questionnaire Dimensions in Study 1

PAQ Division 4: Relationships with Other Persons
Dimension 17—Communicating judgments and related information: This dimension is related to various types of communicating activities including particularly the communication of judgments, opinions, decisions, and information of a non-routine nature. The communication activities include writing, advising, negotiating, and persuading, and the interpersonal relationships involved in generally responsible, often higher level job functions.

Example items:
Nonroutine information exchange, advising, contacts with middle management/staff personnel, staff functions, negotiating, contacts with professional personnel, persuading, coordinates activities

Dimension 18—Engaging in general personal contact: This dimension is characterized by various types of personal communications that are quite varied in terms of content. Hence, the dimension represents something of a general communication type of activity.

Example items:
Supervises nonemployees, serving/catering, public speaking, signaling, contact with special interest groups, contact with the public, coordinates activities

Dimension 19—Performing supervisory and coordination activities: This dimension represents a variety of communication activities such as those involved in supervisory, coordination, and related functions. In some instances it may involve instructing or advising others.

Example items:
Total number of personnel for whom responsible, supervision of nonsupervisory personnel, direction of supervisory personnel, contact with trainees and apprentices

Dimension 20-Exchanging job-related information: This dimension involves activities or contacts with personnel, both within and outside the organization, in which the exchange of job-related information tends to be dominant.

Example items:
Contact with supervisors, contact with manual and service workers, signaling, routine information exchange, overall job-required personal contact

Dimension 21-Public and outside personal contacts: This dimension involves personal contacts with the public or other persons primarily outside the organization, such as in selling, dealing with special interest groups, clients, customers, patients, counselees, etc.

Example items:
Contact with public customers, sales personnel, buyers, special interest groups, and clerical personnel

Cognitively-Oriented PAO Dimensions

Dimension 7-Decision making: This dimension is characterized by the extent to which various mental processes are required in the performance of the job, typically reflected by some type of decision making or problem solving.

Example items:
Reasoning in problem solving, decision making, amount of planning and scheduling, analyzing information or data, using mathematics
Dimension 8-Information processing: This dimension is characterized by job activities involving any of various forms of "processing" or "using" information, perhaps most typically applying relatively standardized procedures or guidelines, although some aspects of decision making usually are also involved.

Example items:
- Transcribing, compiling, coding/decoding, use of short-term memory, combining information, analyzing information or data, using mathematics, reasoning in problem solving

Dimension 33-Decision, communicating, and general responsibilities: This dimension is the most inclusive of all and reflects activities involving considerable amounts of decision making, communicating, and general responsibility.

Example items:
- Advising, reasoning in problem solving, writing, persuading, interviewing, planning and scheduling, decision making, combining information

Perceptually-Oriented PAO Dimensions

Dimension 3-Watching devices and materials for information: The emphasis in this dimension is on observing, or being alert to, devices, materials, processes, and other features of the environment as part of the individual's work activities.

Example items: Use of pictorial materials, use of patterns and related devices, use of measuring devices, observing materials in process

Dimension 5-Being aware of environmental conditions: The dominant aspect of this dimension is that of continually being aware of various aspects of the individual's work
environment, especially various types of events or circumstances. Such awareness might be related to the man made or natural features of the environment.

**Example items:** Man-made features of environment, features of nature, far visual differentiation, depth perception, estimating size, estimating speed of moving objects

**Dimension 35-Performing clerical and related activities:** The dominant aspect of this dimension is involvement in the performance of typical clerical, office, and related types of activities.

**Example items:** transcribing, use of keyboard devices, attention to detail, coding and decoding, routine information exchange, use of written materials, time pressures

**Physically-Oriented PAQ Dimensions**

**Dimension 13-Performing controlled manual and related activities:** This dimension is dominated by the execution of controlled manual activities of various types that may involve the use of tools, equipment, or other devices, or direct use of the hands as in assembling or adjusting tasks.

**Example items:**
use of precision and nonprecision tools and instruments, assembling and disassembling, use of handling devices and tools, use of applicators, use of long-handled tools, hand-arm steadiness

**Dimension 15-Performing handling and related manual activities:** This dimension is characterized primarily by job activities that involve the handling or movement of
materials with the hands and arms or activities that involve manipulating or positioning things with the hands

**Example items:**

- Arranging and positioning, physical handling, feeding and off-bearing, manually modifying, material controlling
Appendix B--Statistical Procedures to Detect Moderators of Validity

The statistical procedure used in Study 1 to detect moderators of GATB validity was first described by Arnold (1982) and has appeared in other published research (e.g., Gutenberg et al., 1983; Schmitt et al., 1990). However, Arnold's procedures for detecting moderators of validity differ from more traditional multiple regression methods and have been severely criticized by Stone and Hollenbeck (1984, 1989). The evidence surrounding this controversy is summarized below, and conclusions about the use of the Arnold approach in Study 1 are drawn.

To adequately address the moderator controversy, as it relates to Study 1, three issues must be considered: (a) the theoretical and statistical issues surrounding moderators of test validity have not been fully resolved, (b) the Stone and Hollenbeck (1984, 1989) criticisms of Arnold's procedures were flawed (Bobko & Russell, 1990), and (c) Arnold (1982) actually discussed several different procedures, which probably have different merit, for detecting moderators in different situations.

Two different but related approaches have been used to search for moderators of test validity: (a) correlational analyses and (b) multiple regression. Correlational analyses have traditionally been used to compare validity coefficients obtained for subgroups of a population. This approach has alternatively been called subgroup or differential validity. For example, the validity of a particular test can be calculated separately for females and males to determine whether the test is significantly more valid for one sex. However, as Arnold (1982) noted, correlational analyses can also be used to examine the moderating effect of a continuous variable (e.g., any job analysis measure in Study 1) and not just a categorical or grouping variable. Rather than calculate validity separately for subgroups, values of a continuous variable can be correlated with validity coefficients to determine whether the former moderates the latter. The results of this type of analysis are the correlations, hereafter referred to as $r^*$, that are shown in
Tables 5 and 6 of Study 1. Note that an investigation of $r^*$ only makes sense when the number of units of analysis is relatively large. The units of analysis in this study were 105 jobs; the units of analysis in the Schmitt et al. (1990) study were 16 assessment center locations.

The other way to assess moderators of test validity is to use hierarchical multiple regression. A criterion measure is regressed on a predictor, the potential moderator variable, and the predictor x moderator interaction. If the unique variance accounted for by the interaction term is statistically significant, then: (a) the predictor-criterion relationship is said to be moderated by the third variable, and (b) the slopes of the subgroup regression lines are significantly different.

Hierarchical multiple regression is the most popular way to determine whether a variable moderates test validity. However, this study and the Gutenberg et al. (1983) study are examples of situations in which regression was not feasible, and the use of correlational analyses was the only way to assess potential moderators. The reason that multiple regression was not feasible was that criterion scores were not available for the units of analysis (i.e., the different jobs). Because jobs were studied in a wide variety of organizations, the criterion measure used in each validity study varied somewhat. As mentioned above, most of the validity studies involved a concurrent design with supervisory ratings of employee performance as the criterion. However, the specific measure used in each organization varied somewhat, and criterion scores were not directly comparable. Therefore, the USES did not include criterion scores in the data base of GATB validity studies. Of course, the predictor in each validity study was always the GATB, and the USES included GATB scores on the data base along with GATB validity coefficients. Therefore, it was not possible to use hierarchical multiple regression to assess the moderating effects of job analysis variables in this study, and Arnold's $r^*$ was used instead.
correlational analyses and hierarchical multiple regression can lead to different
conclusions about a moderator variable. In the same article, they also severely criticized
a completely different approach to testing moderators described by Blood and Mullet
(1977). They presented data from a Monte Carlo simulation to demonstrate that the
Blood and Mullet approach was invalid; however, they presented no data to refute
Arnold's (1982) approach, nor did they take issue with Arnold's three empirical
examples which demonstrated that correlational analyses and multiple regression can
lead to different conclusions about a moderator. Therefore, Stone and Hollenbeck
(1984)'s conclusion that "both statistical arguments and the results of a simulation study
indicated that the recently advanced contentions of Blood and Mullet (1977) and Arnold
(1982) are without foundation (p. 212)" was without foundation. Finally, it is important
to note that Stone and Hollenbeck (1984) only criticized Arnold's procedure for
categorical moderators, but they took no issue with his approach when a moderator
variable is continuous (i.e., r* described above).

that:

Hierarchical multiple regression using product terms...is the appropriate analytic
strategy to employ in order to determine whether or not values of Byx_i (slopes of
the y on x regression) are significantly different for different values of the
moderator. However, it is essential to note that the results of such an analysis will
tell us nothing whatsoever regarding whether or not correlation coefficients r_xy_i are
significantly different for different values of the moderator (p. 219).

Arnold (1984) also challenged Stone and Hollenbeck to provide empirical evidence to
substantiate their criticisms of his approach.
Stone and Hollenbeck (1989) subsequently sought to demonstrate that the empirical examples provided by Arnold (1982) were flawed, because the original data violated several of the standard assumptions of multiple regression. They transformed Arnold's original data to reduce problems in the distributions and then reanalyzed the data. The results of their reanalysis generally showed that, contrary to Arnold's (1982) conclusions, correlational analyses and multiple regression yielded similar conclusions about moderators. They concluded that if correlational analyses have any statistical advantages over multiple regression to detect moderating effects, future research must demonstrate this.

Two points about their conclusions must be noted. First, Bobko and Russell (1990) criticized Stone and Hollenbeck's (1989) transformation of Arnold's data as "a critical error....(and) simply inappropriate and should not be used by future researchers in this domain" (p. 96). Thus, while Stone and Hollenbeck (1989) provided some thoughtful comments about the moderator debate, the basis on which they attempted to refute Arnold's (1982) findings was statistically flawed (Bobko and Russell, 1990).

Second, as in their earlier article, Stone and Hollenbeck (1989) primarily criticized the use of correlational analyses for exploring categorical moderator variables. They provided no evidence that $r^*$ is an inappropriate measure of the moderating effect of a continuous variable and instead suggested that $r^*$ would be of only limited value. However, they also suggested that $r^*$ might underestimate the existence of a continuous moderator variable in some situations. As described above, Study 1 of this research and the Gutenberg et al. (1985) study were examples of situations in which $r^*$ was quite useful for exploring the moderating effects of continuous variables, because criterion measures were not available and not directly comparable, and multiple regression was not feasible.
Appendix C--Comparison of Results of Study 1 to Gutenberg et al. (1983)

Pearson Correlations Between PAQ Dimensions 7, 8, and 33 and Cognitive GATB Validities Partialing out PAQ Dimensions 13 and 15 in Gutenberg et al. (1983) and the Present Study 1

<table>
<thead>
<tr>
<th>PAQ Dimensions</th>
<th>GATB Validity</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>General Intelligence</td>
<td>Verbal Aptitude</td>
<td>Numerical Aptitude</td>
</tr>
<tr>
<td>Decision making (7)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gutenberg et al. (1983)</td>
<td>.37***</td>
<td>.35***</td>
<td>.30***</td>
</tr>
<tr>
<td>Study 1</td>
<td>.20*</td>
<td>.11</td>
<td>.20*</td>
</tr>
<tr>
<td>Information processing (8)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gutenberg et al. (1983)</td>
<td>.24***</td>
<td>.18*</td>
<td>.28**</td>
</tr>
<tr>
<td>Study 1</td>
<td>.19*</td>
<td>.13</td>
<td>.25*</td>
</tr>
<tr>
<td>General responsibilities (33)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gutenberg et al. (1983)</td>
<td>.20*</td>
<td>.27**</td>
<td>.14</td>
</tr>
<tr>
<td>Study 1</td>
<td>.22*</td>
<td>.09</td>
<td>.25*</td>
</tr>
</tbody>
</table>

*p < .05. **p < .01. ***p < .001.
Appendix D--Supplemental Instructions for the Thumbs Problem

Individual-Unaccountable condition:

Your answers will be anonymous, so do NOT write your name on this sheet or the cards. Also, your answers will be mixed with answers from 50 other students.

When the experimenter says STOP, put all your cards (used and unused) in the envelope on the table. Then we will start the next exercise. START WRITING NOW.

Individual-Accountable condition:

When the experimenter says STOP, he will count your answers. The number of answers you wrote will be your score. Then, he will sit down alone with you to discuss and evaluate your score. For example, your score will be compared to other students at UH-Downtown and UH-Central Campus. If your score is high, the evaluation will end. Otherwise, you will have to justify your performance, and the experimenter will expect you to explain why you didn't write as many answers as other students did.

Also, sign below, so that we can give your score to other researchers at UH and Rice. Your score is important to our research on creative thinking. Finally, write your phone number and a good time for us to call you. If your score is low, the experimenter will call you in the next two days to further discuss your score. START WRITING NOW, AND WRITE AS MANY ANSWERS AS POSSIBLE.

Print your name: ________________________________

Sign your name: ________________________________

Phone number: _______ Good time for us to call you: ________

Group-Unaccountable condition:
Your answers will be anonymous and will be mixed with answers written by the other people in your group. Do NOT write your name or group code on this sheet or the cards. Also, your group's answers will be mixed with answers from 10 other groups.

When the experimenter says STOP, everyone will put all their cards (used and unused) in the envelope on the table. Then we will start the next exercise. START WRITING NOW.

**Group-Accountable condition:**

When the experimenter says STOP, everyone will put their cards in the box on the table. Then, the experimenter will count the total number of answers in the box. This will be your group's score. Then, he will sit down with your group to discuss and evaluate its score. For example, your group's score will be compared to other groups of students at UH-Downtown and UH-Central Campus. If your group's score is high, the evaluation will end. Otherwise, your group will have to justify its performance, and the experimenter will expect your group to explain why you all didn't write as many answers as other groups did.

Also, sign below, so that we can give your group's score to other researchers at UH and Rice. Your group's score is important to our research on creative thinking. Finally, write your phone number and a good time for us to call you. If your group's score is low, the experimenter will call you in the next two days to further discuss your group's score. **START WRITING NOW, AND WRITE AS MANY ANSWERS AS POSSIBLE.**

Print your name: ___________________________________________

Sign your name: ___________________________________________

Phone number: ________ Good time for us to call you: ________
Appendix E--Regression Equations for Study 2

**Individual-Unaccountable condition:**
y = -0.843 + 1.129 x

\[ F_b (1, 14) = 11.71, p < .01 \]

**Individual-Accountable condition:**
y = 19.792 - 0.149 x

\[ F_b (1, 12) = 0.40, ns \]

**Group-Unaccountable condition:**
y = 4.100 + 0.542 x

\[ F_b (1, 15) = 2.77, ns \]

**Group-Accountable condition:**
y = 6.429 + 0.729 x

\[ F_b (1, 15) = 8.54, p < .05 \]

**Note.** y is performance on the Thumbs Problem, x is raw score on the Alternate Uses test
Appendix F—Results of Generalizability Analysis

<table>
<thead>
<tr>
<th>GATB Test</th>
<th>Average $r_{xy}$</th>
<th>Average $r_{xyc}$</th>
<th>$s^2$</th>
<th>$s^2_e$</th>
<th>% $s^2$ Due to Sampling Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>.22</td>
<td>.31</td>
<td>.0200</td>
<td>.0044</td>
<td>22%</td>
</tr>
<tr>
<td>V</td>
<td>.16</td>
<td>.24</td>
<td>.0167</td>
<td>.0048</td>
<td>29%</td>
</tr>
<tr>
<td>N</td>
<td>.21</td>
<td>.29</td>
<td>.0193</td>
<td>.0045</td>
<td>23%</td>
</tr>
<tr>
<td>S</td>
<td>.13</td>
<td>.18</td>
<td>.0126</td>
<td>.0051</td>
<td>40%</td>
</tr>
<tr>
<td>P</td>
<td>.14</td>
<td>.18</td>
<td>.0128</td>
<td>.0051</td>
<td>40%</td>
</tr>
<tr>
<td>Q</td>
<td>.16</td>
<td>.23</td>
<td>.0176</td>
<td>.0048</td>
<td>27%</td>
</tr>
<tr>
<td>K</td>
<td>.09</td>
<td>.13</td>
<td>.0106</td>
<td>.0052</td>
<td>49%</td>
</tr>
<tr>
<td>F</td>
<td>.10</td>
<td>.14</td>
<td>.0113</td>
<td>.0052</td>
<td>46%</td>
</tr>
<tr>
<td>M</td>
<td>.10</td>
<td>.13</td>
<td>.0127</td>
<td>.0052</td>
<td>41%</td>
</tr>
</tbody>
</table>

Note. Analyses were weighted by sample size. $r_{xy}$ is observed validity, $r_{xyc}$ is validity corrected for criterion unreliability and predictor range restriction, $s^2$ is observed variance of corrected validities, $s^2_e$ is variance expected from sampling error. GATB tests are: G = general intelligence, V = verbal aptitude, and N = numerical aptitude (cognitive cluster); S = spatial aptitude, P = form perception, and Q = clerical perception (perceptual cluster); K = motor coordination, F = finger dexterity, and M = manual dexterity (psychomotor cluster).
Appendix G--Results of Moderator Analyses for All 105 Jobs

Correlations Between Job Analysis Measures and Cognitive GATB Validities

<table>
<thead>
<tr>
<th>Job Analysis Measures</th>
<th>GATB Validity</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>General Intelligence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Verbal Aptitude</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Numerical Aptitude</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social-Collaborative Complexity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAQ 17</td>
<td>.37***</td>
<td>.36***</td>
<td>.18</td>
<td>.18</td>
<td>.36***</td>
<td>.30**</td>
<td></td>
</tr>
<tr>
<td>PAQ 18</td>
<td>-.14</td>
<td>-.09</td>
<td>-.09</td>
<td>.01</td>
<td>-.10</td>
<td>-.07</td>
<td></td>
</tr>
<tr>
<td>PAQ 19</td>
<td>-.06</td>
<td>-.04</td>
<td>-.08</td>
<td>-.08</td>
<td>-.08</td>
<td>-.01</td>
<td></td>
</tr>
<tr>
<td>PAQ 20</td>
<td>.01</td>
<td>-.17</td>
<td>.03</td>
<td>-.06</td>
<td>-.02</td>
<td>-.20*</td>
<td></td>
</tr>
<tr>
<td>PAQ 21</td>
<td>-.02</td>
<td>.01</td>
<td>.06</td>
<td>.07</td>
<td>.10</td>
<td>.10</td>
<td></td>
</tr>
<tr>
<td>DOT-People</td>
<td>.13</td>
<td>.03</td>
<td>.13</td>
<td>.11</td>
<td>.22*</td>
<td>.10</td>
<td></td>
</tr>
<tr>
<td>Cognitive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAQ 7</td>
<td>.24*</td>
<td>.11</td>
<td>.14</td>
<td>.13</td>
<td>.19*</td>
<td>.09</td>
<td></td>
</tr>
<tr>
<td>PAQ 8</td>
<td>.16</td>
<td>-.05</td>
<td>.11</td>
<td>-.05</td>
<td>.23*</td>
<td>-.02</td>
<td></td>
</tr>
<tr>
<td>PAQ 33</td>
<td>.24*</td>
<td>-.14</td>
<td>.11</td>
<td>-.10</td>
<td>.24*</td>
<td>-.16</td>
<td></td>
</tr>
<tr>
<td>DOT-</td>
<td>.34***</td>
<td>.13</td>
<td>.23*</td>
<td>.15</td>
<td>.32***</td>
<td>.11</td>
<td></td>
</tr>
</tbody>
</table>

Note. N = 105 jobs. PAQ dimensions are: 17 = Communicating judgments, 18 = General personal contact, 19 = Supervising and coordinating, 20 = Exchanging job-related information, 21 = Public and outside personal contacts, 7 = Making decisions, 8 = Processing information, 33 = Overall decision, communication, and general responsibilities.

*p < .05. **p < .01. ***p < .001.
### Correlations Between Job Analysis Measures and Perceptual GATB Validities

<table>
<thead>
<tr>
<th>Job Analysis Measures</th>
<th>Spatial Aptitude</th>
<th>Form Perception</th>
<th>Clerical Perception</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
<td>partial r</td>
<td>r</td>
</tr>
<tr>
<td>Social-Collaborative Complexity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAQ 17</td>
<td>.14</td>
<td>.14</td>
<td>.09</td>
</tr>
<tr>
<td>PAQ 18</td>
<td>-.05</td>
<td>-.10</td>
<td>-.01</td>
</tr>
<tr>
<td>PAQ 19</td>
<td>.13</td>
<td>.11</td>
<td>-.10</td>
</tr>
<tr>
<td>PAQ 20</td>
<td>.01</td>
<td>.01</td>
<td>-.19</td>
</tr>
<tr>
<td>PAQ 21</td>
<td>-.22*</td>
<td>-.23*</td>
<td>.01</td>
</tr>
<tr>
<td>DOT-People</td>
<td>-.10</td>
<td>-.13</td>
<td>.04</td>
</tr>
<tr>
<td>Perceptual</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAQ 5</td>
<td>.11</td>
<td>.13</td>
<td>--</td>
</tr>
<tr>
<td>PAQ 3</td>
<td>--</td>
<td>--</td>
<td>-.05</td>
</tr>
<tr>
<td>PAQ 35</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

**Note.** N = 105 jobs. PAQ dimensions are: 17 = Communicating judgments, 18 = General personal contact, 19 = Supervising and coordinating, 20 = Exchanging job-related information, 21 = Public and outside personal contacts, 5 = Being aware of environmental conditions, 3 = Watching devices and materials for information, 35 = Performing clerical and related activities.

*p < .05.*
### Correlations Between Job Analysis Measures and Psychomotor GATB Validities

- **GATB Validity**

<table>
<thead>
<tr>
<th>Job Analysis Measures</th>
<th>Motor Coordination</th>
<th>Finger Dexterity</th>
<th>Manual Dexterity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( r )</td>
<td>partial ( r )</td>
<td>( r )</td>
</tr>
<tr>
<td><strong>Social-Collaborative Complexity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAQ 17</td>
<td>.01</td>
<td>.05</td>
<td>-.03</td>
</tr>
<tr>
<td>PAQ 18</td>
<td>-.01</td>
<td>-.11</td>
<td>-.03</td>
</tr>
<tr>
<td>PAQ 19</td>
<td>-.14</td>
<td>-.09</td>
<td>.14</td>
</tr>
<tr>
<td>PAQ 20</td>
<td>-.26**</td>
<td>-.27**</td>
<td>-.29**</td>
</tr>
<tr>
<td>PAQ 21</td>
<td>.10</td>
<td>.10</td>
<td>.02</td>
</tr>
<tr>
<td>DOT-People</td>
<td>.05</td>
<td>.08</td>
<td>-.06</td>
</tr>
<tr>
<td><strong>Psychomotor</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAQ 13</td>
<td>-.07</td>
<td>.18</td>
<td>.14</td>
</tr>
<tr>
<td>PAQ 15</td>
<td>.20*</td>
<td>.17</td>
<td>-.07</td>
</tr>
</tbody>
</table>

**Note.**  \( N = 105 \) jobs. PAQ dimensions are: 17 = Communicating judgments, 18 = General personal contact, 19 = Supervising and coordinating, 20 = Exchanging job-related information, 21 = Public and outside personal contacts, 13 = Performing controlled manual & related activities, 15 = Performing handling & related manual activities.

\(* p < .05. \quad ** p < .01. \quad *** p < .001.\)