THE APPLICATION OF MULTIPLE MATRIX SAMPLING TO KSAO-TASK LINKAGES

Sharisse Balanon Dy
B.A., California State University, Sacramento, 2007

THESIS

Submitted in partial satisfaction of
the requirements for the degree of

MASTER OF ARTS

in

PSYCHOLOGY
(Industrial and Organizational Psychology)

at

CALIFORNIA STATE UNIVERSITY, SACRAMENTO

SUMMER
2010
THE APPLICATION OF MULTIPLE MATRIX SAMPLING TO KSAO-TASK LINKAGES

A Thesis

by

Sharisse Balanon Dy

Approved by:

__________________________, Committee Chair
Gregory Hurtz, Ph.D.

__________________________, Second Reader
Lawrence Meyers, Ph.D.

__________________________, Third Reader
Amy Kohls, M. A.

__________________________
Date
Student: Sharisse Dy

I certify that this student has met the requirements for format contained in the University format manual, and that this thesis is suitable for shelving in the Library and credit is to be awarded for the thesis.

__________________________, Graduate Coordinator  
Jianjian Qin, Ph.D.  Date

Department of Psychology
Abstract

of

THE APPLICATION OF MULTIPLE MATRIX SAMPLING TO KSAO-TASK LINKAGES

by

Sharisse Balanon Dy

The linkage process of job analysis requires subject matter experts (SMEs) to rate an overwhelming number of linkages between tasks and KSAOs. Archival data was utilized to explore the capacity of multiple matrix sampling (MMS) theory to reduce SME rater workload while maintaining the desired accuracy and reliability of obtained ratings. SMEs ($N = 28$) were divided into pairs and each pair was assigned to rate subsets of linkages. MMS helped reduce the total amount of ratings required from each individual. A generalizability study supported the use of MMS such that the variance in the ratings generally reflected true differences in the linkages themselves. The SME raters, their classifications, and the various pairs contributed relatively little to the linkage variance. This study provides a solid basis from which future studies can continue to explore the feasibility of MMS in improving the linkage process.

Greg Hurtz, Ph.D.

Date
DEDICATION

I would like to dedicate this thesis to my parents, Elena and Elmer, my sister, Shaina, and my boyfriend, Jing.

Mom and Dad, I cannot thank you enough for your constant encouragement, support, and love throughout my entire academic career. You have dealt with my countless hours studying and preparing for this accomplishment and I thank you for always being there for me. I truly appreciate all you have done for me.

Sis, thanks for cheering me on, even from halfway around the world. Though we are miles apart, I appreciate you thinking of me and continuing to support me. I look forward to seeing you as I walk across the stage.

Jing, I deeply thank you for your amazing support during this important journey of my life. Thank you for your time and efforts to help me succeed. Thank you for caring about my work and being so willing to understand the content of my thesis. Thanks to your unfailing faith in me, I have made it this far. I am forever grateful to you.
ACKNOWLEDGEMENTS

I would like to acknowledge the hard work and dedication my thesis committee has given me.

Dr. Hurtz, I graciously thank you for providing me with immense knowledge, guidance, and support throughout this process. You have been nothing but helpful and patient, ensuring all my questions were answered and my concerns addressed. I have truly enjoyed studying and working with you since the days of my undergrad. I sincerely appreciate all you have done for me. Thank you for helping me with this wonderful achievement.

Dr. Meyers, thank you for helping me exceed my own expectations, both in school and in the workplace. Thank you for being confident in me and always trusting in my work. I have always admired your remarkable wisdom and your dedication to your students. I am truly grateful to have worked with you.

Amy, thank you for staying committed to my thesis, especially from the other side of the country. I appreciate all the time and effort you sacrificed. Your positive words have been very encouraging. Thank you for sticking with me as I completed this milestone in my life.

Lastly, I would like to thank Karen, Catherine, Kasey, and Larry for being so willing to assist me whether it was with the actual content of my thesis, formatting issues, or simply other obstacles I faced unexpectedly. In various ways, the help each of you have given me has made this challenging process possible. I thank you for genuinely caring.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dedication</td>
<td>v</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>ix</td>
</tr>
<tr>
<td>List of Figures</td>
<td>x</td>
</tr>
<tr>
<td>Chapter</td>
<td></td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>The Role of Job Analysis</td>
<td>1</td>
</tr>
<tr>
<td>Preparing for the Job Analysis</td>
<td>2</td>
</tr>
<tr>
<td>Identifying the Tasks Performed on the Job</td>
<td>12</td>
</tr>
<tr>
<td>Inferring the Required Knowledge, Skills, Abilities, and Other Characteristics</td>
<td>19</td>
</tr>
<tr>
<td>Using the Task and KSAO Results</td>
<td>23</td>
</tr>
<tr>
<td>The Linkage Process</td>
<td>23</td>
</tr>
<tr>
<td>Multiple Matrix Sampling Theory</td>
<td>29</td>
</tr>
<tr>
<td>Traditional Sampling Plans</td>
<td>29</td>
</tr>
<tr>
<td>Applications of MMS</td>
<td>39</td>
</tr>
<tr>
<td>Addressing the Needs of Job Analysis with MMS</td>
<td>44</td>
</tr>
<tr>
<td>Reliability of Linkage Ratings Obtained through MMS</td>
<td>45</td>
</tr>
<tr>
<td>Generalizability Theory</td>
<td>47</td>
</tr>
<tr>
<td>An Extension of Reliability</td>
<td>47</td>
</tr>
<tr>
<td>Using G Theory to Evaluate Job Analysis Ratings</td>
<td>48</td>
</tr>
<tr>
<td>Framework of G Theory</td>
<td>50</td>
</tr>
<tr>
<td>Summary</td>
<td>57</td>
</tr>
<tr>
<td>2. METHOD</td>
<td>59</td>
</tr>
<tr>
<td>Participants</td>
<td>60</td>
</tr>
<tr>
<td>Procedure</td>
<td>62</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>67</td>
</tr>
<tr>
<td>3. RESULTS</td>
<td>71</td>
</tr>
<tr>
<td>Common KSAO-Task Linkages</td>
<td>71</td>
</tr>
<tr>
<td>Descriptive Statistics</td>
<td>71</td>
</tr>
</tbody>
</table>
LIST OF TABLES

1. Table 1 Expected Mean Squares and Random Effects Variance Components for the Fully Crossed Design ................................................................. 53
2. Table 2 Expected Mean Squares and Random Effects Variance Components for the Partially Nested Design ................................................................. 56
3. Table 3 Expected Mean Squares and Random Effects Variance Components for the Multiple Matrix Sampling Design ......................................................... 57
4. Table 4 Characteristics of the SME Sample ........................................................................... 61
5. Table 5 Expected Mean Squares and Random Effects Variance Components for the Common Linkage Ratings ......................................................................... 69
6. Table 6 Expected Mean Squares and Random Effects Variance Components for the Non-Common Linkage Ratings ................................................................. 70
7. Table 7 Frequency Count of Common Linkage Ratings ............................................................ 72
8. Table 8 Mean Common Linkage Ratings for each SME and Pair ........................................... 74
9. Table 9 Mean Common Linkage Ratings for each Classification ............................................. 76
10. Table 10 G Study Results for the Common Linkage Ratings .................................................. 78
11. Table 11 Frequency Count of Non-Common Linkage Ratings ................................................ 83
12. Table 12 Mean Non-Common Linkage Ratings for each SME and Pair ................................. 85
13. Table 13 G Study Results for the Non-Common Linkage Ratings .......................................... 87
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Figure 1 Census-Taking</td>
<td>31</td>
</tr>
<tr>
<td>2.</td>
<td>Figure 2 Examinee Sampling</td>
<td>33</td>
</tr>
<tr>
<td>3.</td>
<td>Figure 3 Item Sampling</td>
<td>34</td>
</tr>
<tr>
<td>4.</td>
<td>Figure 4 Multiple Matrix Sampling</td>
<td>36</td>
</tr>
<tr>
<td>5.</td>
<td>Figure 5 SME-by-Linkage Matrix: Modified Sampling Plan</td>
<td>64</td>
</tr>
<tr>
<td>6.</td>
<td>Figure 6 KSAO-Task Linkage Rating Scale</td>
<td>67</td>
</tr>
</tbody>
</table>
Chapter 1
INTRODUCTION
The Role of Job Analysis

Job analysis is important for understanding the nature of a job and establishing selection and training standards for an organization. It is essentially the “process of discovering the nature of a job” (Brannick, Levine, & Morgeson, 2007, p. 7). Job analysis relates these standards to the job and enables an organization to meet mandated requirements established by the Equal Employment Opportunity Commission and outlined in the Uniform Guidelines on Employee Selection Procedures (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978). Job analysis is important for the development of various workplace needs, such as selection tests, training modules, recruitment, placement, job counseling, performance appraisals, job design and redesign (Berry, 2003; Goldstein & Ford, 2002; Goldstein, Zedeck, & Schneider, 1993; Murphy & Cleveland, 1995). This study focuses on the role job analysis plays in developing selection and training standards in particular.

Job analysis plays a critical role in establishing content validity for selection and training purposes (Goldstein et al., 1993). In order to obtain the information needed to support content validity, the tasks performed on the job and the essential knowledge, skills, abilities, and other characteristics (KSAOs) needed for performing them must be identified (Berry, 2003; Brannick et al., 2007; Goldstein & Ford, 2002). Through this systematic study, organizations gain evidence of the criticality and job-relatedness of the
KSAOs used for the development and validation of selection and training standards that comply with the law.

In addition to establishing selection and training standards, job analysis serves several other purposes. Many researchers describe these goals, which revolve around the development of job descriptions, job classifications, job evaluations, job design and redesign, performance appraisal, worker mobility, workforce planning, efficiency, safety, and legal and quasi-legal requirements (Berry, 2003; Brannick et al., 2007). Different needs require the gathering of specific types of information. For example, the need for employee selection measures requires the gathering of worker requirements for performance, whereas employee training needs require the gathering of task and work context information. Job analysis ensures organizations that their purpose and goals are evident through the work performed in its jobs.

Preparing for the Job Analysis

A job analyst, the individual analyzing a job, must specify the goal of the job analysis in advance (Brannick et al., 2007). At this phase, the job analyst determines the needs of the organization and specifies the intended use of the information obtained from the job analysis. Knowing the organization’s needs allow the job analyst to establish criteria for evaluating whether the goal of the job analysis is being met (Primoff & Fine, 1988).

Typically, a well-trained job analyst works with an organization to conduct the job analysis and implement the resulting information. The job analyst consults different sources of information as an initial phase for gathering information, such as printed or
electronic resources, observations of a job being performed, or discussions with individuals who know the job (Berry, 2003). An organization’s documents typically contain a lot of information about the job. These resources include current job announcements, training manuals, and previously conducted job analyses. Published research on job analyses is also helpful by providing procedures for conducting a job analysis.

Collecting information about the job. Job descriptive information can be collected through a variety of methods, such as observations, interviews, questionnaires, and written records (Berry, 2003; Brannick et al., 2007). “Simple deduction based on assumed knowledge” can also be used (Primoff & Fine, 1988, p. 14). A job analyst can directly observe the job for a period of time, such as a single work shift or a few hours during which an individual performs the job being studied. The job analyst can ask questions of an individual on the job in order to gain more details about what is being performed.

Interviews help job analysts gather information about the tasks performed on the job. Job incumbents and supervisors serve as subject-matter experts (SMEs) because they are effective sources for job analysts to learn what is done on the job and what skills are needed to perform those activities successfully (Berry, 2003). The goal of an interview is to obtain as much information as possible. Several interviews are often conducted and can take place at various stages of the job analysis. More than one individual can also be interviewed at the same time. Groups of SMEs to be interviewed should consist of a relatively small number of job incumbents and/or supervisors. These individuals are
gathered and guided to provide important job information. These individuals can also refine preexisting lists of task and KSAOs gathered at a previous time.

Although data from observations and interviews provide substantial job information, questionnaires are central to the job analysis, providing the most complete information about the job (Berry, 2003). Job analysts design questionnaires using the information they learn from observations and interviews. Task statements are placed into a questionnaire as items and rating scales are developed to obtain responses to each statement. Various rating scales are utilized for questionnaires. For example, respondents can simply be asked to indicate whether a particular task is performed on a job or not. They can also indicate how critical or important a task is to the job, how much time is spent performing a task, or how frequently a task is performed. Questionnaires are beneficial for obtaining data from a large number of job informants.

Job analysts also obtain job information by collecting written records about the job. These records include documents describing what is performed on the job, such as current job descriptions or previously conducted job analyses (Berry, 2003). Other background materials can be obtained from published sources, such as the Dictionary of Occupational Titles (DOT, 1991) and the Occupational Information Network (O*NET, 2001).

Beyond these initial methods of gathering job information, specialized job analysis techniques and instruments are used to collect job information. Task analysis (Brannick et al., 2007) and the task inventory method (Morsh, 1964) are work-oriented methods that primarily focus on exactly what a worker does. The Occupation Analysis
Inventory (Cunningham, Boese, Neeb, & Pass, 1983) inventories work elements on the job, such as information received by a jobholder, mental activities, and work goals. The job element approach and critical incidents technique are other specialized job analysis methods (Berry, 2003). Other methods of collecting job information are worker-oriented and broader in their application. A classic example is the Position Analysis Questionnaire (McCormick, Jeanneret, & Mecham, 1972), which was developed through a survey of over 4,000 jobs in an attempt to produce a checklist of worker traits that is generic to almost all jobs. The significance of the Position Analysis Questionnaire is that it stands as a standardized list of actions and behavior common to almost all jobs. While these work- and worker-oriented techniques and instruments are useful for certain situations and organizational needs, a combination of both types is effective for developing selection and training standards in the current study.

Few methods combine work-oriented and worker-oriented approaches. Brannick et al. (2007) explain the Combination Job Analysis Method. The Position Analysis Questionnaire and Combination Job Analysis Method methods involve a combination of work-oriented and worker-oriented strategies, which is essentially the analysis of both tasks and KSAOs. These methods are similar to those employed in this study, where the tasks of the job were first generated to provide the basis from which the KSAOs were inferred.

Selecting SMEs. Job incumbents and supervisors are appropriate for supplying important job information, especially since they are individuals with a large amount of knowledge and expertise about the job. These SMEs possess experiences performing the
job and/or observing subordinates. The use of SMEs is central to job analysis. The validity of conclusions drawn from job analysis are dependent on the accuracy of the information obtained (Green & Stutzman, 1986). SMEs are essential to providing this information. According to Landy and Vasey (1991), SMEs provide quality job analysis ratings because of their understanding of the job’s work behaviors.

Gathering groups, or samples, of SMEs requires a job analyst to consider the potential effects of their various personal characteristics and background experience that may present a source of bias in the job analysis data. While it is assumed that SMEs provide accurate job analysis information, research that examines the characteristics and background experience of SMEs is of great practical and theoretical relevance to establish an empirical basis for SME selection. The job analyst must identify potential systematic factors that influence the job analysis rating provided by SMEs, such as their gender, age, current work assignments, and number of years of experience. It is important that particular subgroups of SMEs are not underrepresented in order to avoid systematic differences in results (Landy & Vasey, 1991).

As with any study, the sample used should be representative of the population. However, research shows that differences in experience can elicit different perceptions of the job, therefore influencing job analysis ratings (Harvey, 1991; Nathan & Alexander, 1985; Rothstein & Jackson, 1980). Human judgments are subject to inaccuracies due to various influences, whether individuals are aware or unconscious of these biases. Nathan and Alexander (1985) and Rothstein and Jackson (1980) attribute such discrepancies in job analysis ratings to the inferential leap used when making judgments. Primoff and Fine
(1988) point out that the shortcoming of task analysis is its reliance on human judgment. However, research has indicated that errors are more likely to occur for ratings of KSAOs than tasks (Morgeson, Delaney-Klinger, Mayfield, Ferrara, & Campion, 2004).

Several researchers have examined the potential influence of SME experience. Research on the Position Analysis Questionnaire, for example, revealed that expert ratings tend to be more reliable than naïve ones (DeNisi, Cornelius, & Blencoe, 1987; Friedman & Harvey, 1986). However, these findings were based on laboratory experiments using undergraduate students. The question is therefore raised of whether or not these results can be generalized to supervisors and incumbents. Supervisors can be considered experts due to their responsibilities of overseeing several individuals perform their jobs. In contrast, incumbents have only their personal experiences performing their individual positions to serve as the information they use to make job analysis ratings. Tross and Maurer (2000) were able to identify differences in frequency and importance ratings of task statements due to SME job experience. Job incumbents in the high tenure group tended to provide higher task frequency and skill importance ratings as opposed to the low tenure groups. The researchers offered the explanation that incumbents with varying levels of experience may simply perceive the same job differently. However, the question is raised of whether more experienced incumbents actually perform tasks more often or are more likely to misrepresent their judgments of task frequency due to pressures of self-image and recognition of the benefits of self-enhancement.
In a laboratory experiment, Richman and Quiñones (1996) found that participants who performed a task were more accurate in their task frequency ratings than participants who observed the task being performed. The authors raised an important concern regarding the generalizability of these findings to incumbents and supervisors. While it is logical that supervisors may provide more accurate job information, their increased experiences may lead to more automatized performance, which may reduce their cognitive processing in rating task frequency. Moreover, participants in this study with less experience performing or observing the task were more accurate in their task frequency ratings than participants with more experience. These findings may be attributable to differences in encoding the frequency of tasks into memory due to the varying levels of experiencing the tasks.

Landy and Vasey (1991) also explored SME experience, as well as SMEs’ gender, race, and education, and the influence of these demographic characteristics on task frequency ratings. Only the experience factor was significantly related to the ratings, a result that stresses the importance of including SMEs of varying levels of experience in samples. The gender variable also had a substantial impact, but was confounded with experience. Court cases often stress the racial and educational make-up of SME samples. While SME samples should be representative of their stakeholder groups, Landy and Vasey’s (1991) study demonstrates why the impact of these factors on ratings is not cause for concern. Instead, it is more crucial to investigate the experience levels of SMEs and ensure this variable is not confounding ratings.
In addition to SME experience, personal characteristics tend to result in different task ratings. In terms of work attitudes, for example, Conte, Dean, Ringenbach, Moran, and Landy (2005) found that job incumbents who reported higher job satisfaction, job involvement, and organizational commitment tended to rate tasks as higher in frequency and more important. Thus, job analysts must be wary of their selection of SMEs. The association of work attitudes to job analysis ratings may lead to inflated ratings. Carefully planned sampling procedures may help reduce the potential for this problem.

Morgeson and Campion (1997) identified several social and cognitive sources of inaccuracy in ratings provided by SMEs. The researchers contended that social sources of inaccuracy were potentially due to social influence and self-presentation processes. They also stated that cognitive sources of inaccuracy were due to the limitations and biases in information-processing. Differences due to social influences were further explored in a field study conducted by Morgeson et al. (2004). The researchers identified inflation in ability statements, but not in task statements. The inflation was attributed to self-presentation motives, which refer to attempts to present oneself in a more acceptable and socially desirable manner.

In evaluating methods for selecting job analysis respondents, Green and Stutzman (1986) developed what they called a “carelessness index,” which refers to an individual’s tendency to endorse tasks that are unrelated to the job of interest as important. Carelessness was revealed in over half of incumbents, due to their responses to “bogus” tasks. SMEs were rating tasks as important even though the tasks were known to be unrelated to the job. Consistent with the self-presentation motives identified in the
research of Morgeson and Campion (1997), individuals with high carelessness index scores may be attempting to project positive images of themselves.

These findings serve as a warning to job analysts to exercise caution in selecting SMEs based on the assumption that they can provide accurate job analysis information. Careful SME selection is essential and some suggestions to avoid or alleviate the problems raised by SME differences include screening individuals to determine the degree to which their attitudes, experiences, and personal characteristics may result in inaccurate ratings.

*Increased SME involvement.* Job analysts utilize task analysis information to infer the KSAOs required for performing the identified tasks (Landy & Vasey, 1991). Being trained in industrial psychology, job analysts possess an understanding of the relationship between tasks and KSAOs, as well as the rating process. Accompanied with knowledge of the job, job analysts may provide sufficient task frequency and importance ratings (DeNisi et al., 1987).

Although job analysts typically make inferences about the KSAOs needed for performing the tasks of job, more contemporary practices are making use of SME involvement for completing these judgments (Hughes & Prien, 1989; Landy & Vasey, 1991). Research supports the interchangeability of job analysts and SMEs (Baranowski & Anderson, 2005; Maurer & Tross; 2000). SMEs are useful for making these inferences and then linking the tasks to the related KSAOs. Groups of SMEs complete a judgmental process to infer the KSAOs needed to perform the tasks identified as significant (Berry, 2003). SME involvement in this process enhances the defensibility of the established
linkages and the subsequent linkage ratings for organizations. Despite the expense and intrusiveness of having SMEs perform the inferences and linkage process, organizations must ensure the linkage ratings that result from the job analysis comply with legal requirements such as content validity. SMEs are the individuals who possess the experience in performing the tasks and are therefore credited with truly understanding what it takes for successful performance.

Landy and Vasey (1991) describe the advantages of more SME involvement in contrast with typical procedures that utilize job analysts for a majority of the job analysis. They explain that SMEs have more involvement in the process by considering exhaustive lists of tasks, whereas job analysts only have the information they gather from interviews, observations, and documents to make their judgments. Moreover, job analysts can only estimate the relative importance and frequency of tasks by combining information using their professional judgment and past experience. This quasi-clinical method is not as good as the statistical analyses that can be carried out on the ratings provided by SMEs. Using SMEs allows for decisions about selection and training standards “based on an empirical rather than a clinical basis” (Landy & Vasey, 1991, p. 30). While an advantage of having job analysts perform the linkage ratings is their better theoretical understanding of the process, research by Baranowski and Anderson (2005) indicates SMEs, both incumbents and supervisors, provide sufficient ratings. They note that rater type is does not significantly affect the reliability of linkage ratings and suggest the use of SMEs is a promising approach.
The extent of SME involvement with regard to asking them to complete all job analysis and linkage activities is of concern by some (Hughes & Prien, 1989; Sanchez & Fraser, 1994). Asking SMEs to make judgments regarding the KSAOs needed for performing the tasks of a job is a demand that may elicit negative reactions. Additionally, SMEs may not be fully capable of making such inferences, which may result in a waste of cost and effort.

Identifying the Tasks Performed on the Job

What exactly is performed on the job is an important question to answer. There is much to describe when discovering what is done on the job. The activities that comprise a job refer to the groups of physical motions or sensory processes performed to fulfill some worker requirement (Brannick et al., 2007). A given job can consist of hundreds of activities, and even more if the job is complex. Tasks refer to collections of activities that are intended to achieve specific job objectives (Brannick et al., 2007). Anywhere from 30 to 100 or more tasks or more can be identified in a job analysis. To organize tasks, duties can be established, each of which consists of related tasks that are all directed at some general goal of the job (Berry, 2003). A job analysis may result in five or more duties.

Task statements, in particular, begin with action verbs to convey the “observable nature of a task” (Berry, 2003, p. 52). Some verbs are clearly observable since they reflect actual physical performance, such as “record” and “type.” However, many other verbs reflect cognitive performance, such as “evaluate” and “predict.” Thus, task analysis has evolved to allow for these task statements that are cognitive in nature. To ensure cognitive task statements are still observable and verifiable, they include observable
terms that suggest observable outcomes. Berry (2003) recommends that task statements convey exactly what is performed on the job. Abstract verbs should be replaced with clearer ones.

*Generating the tasks.* Task analysis is a work-oriented method that has been developed to identify the essential activities performed on the job (Berry, 2003; Brannick et al., 2007; McCormick, 1976). This technique results in a complete inventory of these activities in the form of task statements (Drauden, 1988). This approach allows the job analyst to understand what is done on the job, as well as document and communicate that understanding.

Typically, an analyst initially gathers information regarding what exactly is done on the job. The job analyst can obtain this information through direct observation of job performers and/or individual or group interviews with SMEs (Berry, 2003; Brannick et al., 2007). However, more common practices involve the gathering of a sample of job experts at a location away from the workplace. This group should include incumbents and supervisors, men and women, individuals of various ethnic backgrounds, and individuals possessing various experiences and years on the job. These SMEs are first asked to generate as many tasks as they can individually (Brannick et al., 2007). Hundreds of tasks on the job can be identified during this time (Drauden, 1988). The group is then asked to review each list and add missing tasks or delete duplicates. Tasks can also be clarified at this point. The job analyst then carefully screens the final list, edits the tasks, and produces a list of about 30 to 100 or more tasks in draft form, sorted into five or more major duty categories.
Once all the tasks of the job have been identified, a second group of SMEs is assembled to determine which of these tasks is essential to the job (Brannick et al., 2007). Some individuals in the group may be participants from the initial task generation meeting. The group reviews the task list and determines if it is sufficient, if certain tasks need to be separated into two or more tasks, or if two or more tasks need to be combined into one. The duty labels are also reviewed to determine if they are appropriate and encompass the correct tasks. After revisions to the list of tasks are made, which are usually minor, the SMEs are asked to complete a task analysis questionnaire, on which the tasks identified on the job are the items (Berry, 2003).

**Rating the tasks.** Depending on the purpose of the job analysis, the rating scales of the task questionnaire can vary. Rating scales can ask respondents to indicate the relative importance, difficulty, or criticality of a task (because of the consequence of error), or the relative time spent on a task in relation to other tasks (Berry, 2003; Brannick et al., 2007; Landy & Vasey, 1991). Respondents can also be asked to indicate if a task is or is not required on the job. It is important that SMEs understand ratings must be based on the job in general, not on their individual positions or the positions they directly supervise. Many questionnaires consist of more than two rating scales; however, research has indicated that several scales are not necessary (Sanchez & Fraser, 1992). Equivalent information is obtained with fewer scales, which translate into less work analyzing job analysis ratings. No more than two rating scales are sufficient for producing the needed job information. Moreover, Berry (2003) asserts these various rating scales essentially describe the
importance of a task. Therefore, it appears best for a questionnaire to ask respondents
whether or not each task is performed on the job and how important each task is.

*Analyzing the tasks.* Responses on the task questionnaire can be analyzed
statistically to gain a sense of how tasks on the job are perceived by the SMEs (Landy &
Vasey, 1991). Descriptive statistics include the percentage of respondents indicating they
perform a task on the job and the mean ratings and standard deviation for each task. This
information can indicate the average degree of importance a task has on the job. Task
importance can be computed in other ways, such as the sum or product of task difficulty
and criticality ratings (Brannick et al., 2007). A task importance value is computed for
each rater and the mean task importance value is obtained over all raters for an overall
task importance value for a given task. This process would be repeated for all tasks rated.

A final product of the task analysis is typically a written document consisting of a
list of the tasks identified as an essential part of the job (Brannick et al., 2007). The tasks
should be organized within duty categories and in order of the most important to the least
important.

*Assessing reliability.* Results from the task questionnaire provide the job analyst
with standardized information regarding the tasks. Standardizing the information helps to
control measurement error and improve reliability (Berry, 2003). Reliability is an
important consideration to be made when interpreting job analysis data. Reliability refers
to the “amount of error or disagreement contained in the responses” (Brannick et al.,
2007, p. 272). It indicates the extent to which judgments made during the rating process
were consistent (Berry, 2003). The idea is that while some differences in ratings are due
to real differences among the SMEs, some part of the differences is due to measurement error. When subjective judgments are made, some amount of uncontrolled error is present.

It is important to evaluate the reliability of task ratings to determine the amount of error present in the data. By estimating reliability, the job analyst knows how much confidence can be placed in the data. Decisions based on erroneous data are likely to be wrong and contain a great deal of error. Thus, it is worth the effort to collect job information that has as minimal error as possible. Brannick et al. (2007) emphasize both the scientific and practical importance of estimating the reliability of job analysis ratings.

There are various sources of error in job analysis data, including human judgment, the specific item content, and changes in job content over time. However, human judgment is the most commonly studied source of error, especially since individuals may differ in their perceptions and experiences on the job (Brannick et al., 2007). As noted earlier, these differences influence job analysis ratings and must be taken into consideration. By measuring the reliability of job analysis data, the job analyst can judge the precision of the data and decide if more or better data must be collected, or if some other corrective action must be taken.

**Interrater agreement.** Interrater agreement can be measured by calculating $r_{wg}$, which allows for comparison of the agreement of ratings with the agreement that would have been obtained if the ratings were made at random (James, Demaree, & Wolf, 1984). Calculating the $r_{wg}$ compares the observed standard deviation with the standard deviation that would have been obtained with random ratings. A standard deviation refers to the
spread of ratings in relation to their mean. A smaller spread of the observed ratings in comparison to the spread of the theoretical distribution based on chance allows the job analyst to conclude there is relatively good agreement among the raters (Brannick et al., 2007).

*Interrater reliability.* While “agreement” refers to how similar or different a set of ratings are to each other, “reliability” refers to the variance due to random errors and the variance due to systematic differences among the items of interest. Interrater reliability can be evaluated by computing the Pearson product-moment correlation coefficient, $r$, with ratings from at least two raters (Berry, 2003). This coefficient is the simplest true index of reliability. Large correlations generally indicate little error and, thus, good reliability. This method has been criticized because it fails to account for differences in means across raters (Brannick et al., 2007). A high correlation coefficient may indicate high agreement when agreement is actually lacking. It measures the pattern of relations for sets of ratings, which can yield a high correlation if the sets are similar, but have actual ratings that are very different. For example, on a 5-point rating scale of task importance, Rater A may provide ratings of 3, 4, and 5 to three tasks respectively and Rater B may provide ratings of 1, 2, and 3 to the same tasks. While there is no agreement between the ratings, there is the same pattern of responding present in both sets. This relationship corresponds to a correlation coefficient of 1.0, but does not indicate the distinct differences among the actual ratings. Therefore, the use of intraclass correlation is more appropriate for assessing interrater reliability.
With ratings from three or more raters, the intraclass correlation can be computed. The intraclass correlation is an index of the reliability of ratings that produces an indication of how consistently raters provided task judgments (Berry, 2003; Lahey, Downey, & Saal, 1983). Intraclass correlation allows the job analyst to deal with differences in the means across raters (Cornelius, 1988) and is a convenient way of estimating the number of raters needed to obtain a desired level of reliability (Brannick et al., 2007). Using the analysis of variance (ANOVA) model, the intraclass correlation is usually expressed as the ratio of the variance associated with targets, or individuals, over the sum of the variance associated with targets plus error variance (Lahey et al., 1983). This ratio results in an index which is used to determine the degree of consistency between the ratings provided by multiple raters.

Intraclass correlation is associated with generalizability theory, a broad and flexible theory of measurement which will be discussed later. The idea is to take the correlation between measures taken under various conditions and examine how one set generalizes to the other. Brannick et al. (2007) provided an example of comparing job analyst ratings to job incumbent ratings. If the correlation between the two sets of ratings is high, then there is confidence that one set of ratings will predict the other and either set could be used. A low correlation between the two sets, however, would indicate that different information is provided from the two sources.

Research on task ratings. Several researchers have investigated potential factors that influence job analysis ratings. Extensive research has examined ratings of tasks regarding task significance and importance, frequency, and time spent (Conte et al., 2005;
Goffin & Woycheshin, 2006; Green & Stutzman, 1986; Landy & Vasey, 1991; Lindell, Clause, Brandt, & Landis, 1998; Richman & Quiñones, 1996). Results from these studies have identified potential differences in SME characteristics and experience which contribute to task ratings, which were discussed previously. Types of task ratings have also resulted in differences among SMEs. For example, Lindell et al. (1998) found more interrater agreement among ratings of the time spent performing tasks than importance ratings. Additionally, if SMEs do not understand what task “criticality” means or do not agree with each other about its meaning, then results from the data provided by these SMEs are likely to be faulty. In a meta-analysis of over 100 studies, Dierdorff and Wilson (2003) found that task rating scales typically have a reliability of .77. Their research indicated that scales of frequency (.70) and reliability (.77) tended to be more reliable than scales of difficulty (.63) and time spent (.66). Generally, Brannick et al. (2007) suggest reliable task ratings can be obtained with large samples.

Inferring the Required Knowledge, Skills, Abilities, and Other Characteristics

Although a task analysis questionnaire is important for obtaining information about the tasks performed on the job, it does not provide information about the attributes needed for job performance. Instead, the questionnaire serves as a basis from which these attributes can be inferred (Berry, 2003). Brannick et al. (2007) asserts that the information provided by the analysis of these attributes is more crucial than that of tasks. Thus, after the essential tasks performed on the job are identified, the potential KSAOs needed for successful performance of the tasks is determined (Brannick et al., 2007). Since KSAOs are more difficult to observe and verify, KSAO statements are more
complex to develop than task statements (Berry, 2003). Whereas tasks can be directly observed and confirmed, KSAOs typically involve cognitive processes and are subject to potential sources of inaccuracy.

*Generating the KSAOs.* A group of SMEs are gathered to first generate the KSAOs needed on the job. The SMEs conduct a judgmental process to infer these KSAOs (Drauden, 1988). Some of the SMEs may be from the previous task analysis groups. Additionally, more supervisors than incumbents may be utilized because of supervisors’ greater familiarity with the individuals who are successful on the job (Brannick et al., 2007). The SMEs review the final product developed from the task phase, which is the list of tasks and their importance values. The group provides feedback about the completeness and accuracy of the list, including their thoughts regarding the duty labels and importance values.

To begin generating KSAOs, the SMEs receive definitions and examples of KSAOs. The SMEs are then asked to produce lists of basic sensory and physical requirements of the job. They must be careful to avoid the exclusion of people unnecessarily, such as generating KSAOs of fully functioning individuals and not of individuals with disabilities (Brannick et al., 2007). After lists of the sensory and physical requirements of the job are produced by each SME, the group considers one duty category at a time to generate the KSAOs needed to perform all the tasks in each duty. About 30 to 100 or more KSAOs may be generated. The KSAOs are organized into related groups, which help individuals who complete the questionnaires recognize the usefulness of each KSAO on the job (Berry, 2003).
Rating the KSAOs. A list of all KSAOs, including the various sensory and physical requirements, is provided to a new group of SMEs in draft form (Brannick et al., 2007). The group reviews the list, discussing any necessary revisions. Next, the SMEs receive instructions for completing ratings scales which they use to provide information about the KSAOs. The rating scales can ask respondents to indicate if a KSAO is required by employees upon entry, the extent to which trouble is likely if a KSAO is ignored in selection, or the extent to which different levels of a KSAO distinguish superior from average workers (Brannick et al., 2007). Each SME then rates each KSAO using each rating scale.

Analyzing the KSAOs. After SMEs have generated and rated the KSAOs needed for successful job performance, the ratings can be analyzed. Descriptive statistics and the average rating across all SMEs for each KSAO are computed to determine how they are generally perceived (Brannick et al., 2007). Finally, a list of all KSAOs and their corresponding ratings is developed and verified by the SMEs.

Just as it is important to assess the reliability of task ratings, the job analyst must evaluate the reliability of KSAO ratings to determine the extent to which information gained from the ratings may be interpreted accurately (Brannick et al., 2007). High interrater agreement, more specifically, is desired because it indicates a high degree of similarity between raters of KSAOs. Research indicates that individual raters do not produce very reliable KSAO ratings. Research by Hughes and Prien (1989) indicated that the average agreement of KSAO importance was only .31, suggesting a problem with interrater reliability. Cain and Green (1983) also found low reliability of trait ratings in a
study of the ratings provided by the *Dictionary of Occupational Titles*. Panels of expert raters, rather than individual raters, can significantly control the reliability of ratings due to differences in raters (Van Iddekinge, Putka, Raymark, & Eidson, 2005). Brannick et al. (2007) suggest that the reliability of KSAO ratings is expected to increase “to the extent that the KSAOs are relatively concrete and closely tied to specific tasks” (p. 236). For example, KSAO ratings for “ability to handcuff” may be much more reliable than “skill in verbal reasoning” for a correctional officer.

*Research on KSAO ratings.* While there is an abundance of research on task ratings, less research has investigated KSAO ratings. This research has sought to identify factors that influence KSAO ratings of importance, frequency, and usefulness (Hughes & Prien, 1989; Morgeson et al., 2004; Sanchez & Fraser, 1994; Tross & Maurer, 2000). However, as noted earlier, KSAOs are more difficult than tasks to examine because they are harder to observe and verify. As proposed by Morgeson and Campion (1997) and supported by research from Morgeson et al. (2004), KSAO statements are susceptible to self-presentation tactics. SMEs may have desires of portraying themselves in a socially acceptable manner or want to receive credit for appearing to possess particular skills. These motives tend to lead to inflation of KSAO ratings as compared with task ratings. Thus, job analysts must interpret KSAO ratings with the awareness of such self-presentation motives.

Primoff and Fine (1988) assert that SMEs should be the source of job information. However, Goffin and Woycheshin (2006) raise the concerns of having SMEs make judgments when inferring KSAOs. As noted earlier, human judgments are subject to error
due to social and cognitive sources of inaccuracy (Morgeson & Campion, 1997; Nathan & Alexander, 1985; Rothstein & Jackson, 1980).

**Using the Task and KSAO Results**

Identifying the important tasks performed on the job and the KSAOs needed to perform them is critical for developing and evaluating selection and training standards. A major advantage of the task and KSAO analysis is that these processes provide evidence that the KSAOs are critical and job-related (Berry, 2003; Landy & Vasey, 1991). Task analysis allows an organization to obtain detailed task information, such as the importance of a given task relative to another task. This information is valuable for the specific uses of developing selection tests and training programs. Task analysis alone is not sufficient for broader purposes of comparing jobs or assessing selection and training standards across jobs. It does not directly provide the information needed for these purposes. With only knowledge of tasks, job analysts can only make inferences about the KSAOs needed to perform them in order to develop selection and training standards. However, KSAO analysis provides a more direct link to build selection and training standards.

**The Linkage Process**

After SMEs make judgments to determine the KSAOs needed to accomplish the important tasks of the job, they need to make judgments regarding the relationship between the KSAOs and tasks (Berry, 2003). The content obtained from the task and KSAO analyses is the basis for this judgmental process whereby individual KSAOs are linked to individual tasks, which results in the creation of KSAO-task linkages (Goldstein
et al., 1993). Through a “linkage process,” SMEs link the KSAOs to the tasks and rate the relevance of each KSAO to performing each task. SMEs receive instructions on how to perform this linkage process (Goldstein & Ford, 2002). Through the linkage process, the connection between the KSAOs and tasks of the job is established. SMEs only link the tasks and KSAOs that have met some previously established criteria, such as being a crucial or frequently occurring part of the job.

Once the linkages are established, SMEs review and rate each KSAO-task linkage and rate the usefulness or importance of a KSAO to the task with which it is linked. Raters can receive various requirements when evaluating linkage ratings. For example, they can be asked to indicate if a KSAO is needed to perform a task or to identify the extent to which the KSAO is needed to perform the task (Baranowski & Anderson, 2005). Raters may also be asked to indicate whether the KSAO is essential, helpful, or irrelevant to its paired task. The average rating for each linkage can be obtained by averaging the ratings provided across the SMEs. For each task, these values result in a list of the KSAOs needed to perform the task.

The linkage ratings provided by the SMEs are important for establishing selection and training standards. For example, a performance test designed for selection purposes may be developed and evaluated for its capacity to assess a range of important job skills that were identified in a task analysis. In another example, information from a task analysis may define the training needs, content, and objectives for a program designed to meet the needs of newly hired individuals (Berry, 2003). By demonstrating to analysts the needed tasks of a job and the needed KSAOs to perform those tasks, the ratings of the
KSAO-task linkages provide useful information for designing effective selection practices and training programs, as well as justification and validation for their use (Berry, 2003; Goldstein & Ford, 2002).

Research on linkage ratings. Despite the research on task and KSAO ratings, there is a serious lack of research on the ratings of the linkages between tasks and KSAOs. Hughes and Prien (1989) examined the judgment process of linking skills to tasks for the purpose of establishing test specifications and found sufficient interrater agreement among the SMEs who made the linkages. Sanchez and Fraser (1994) attempted to empirically link tasks to KSAOs through regression analyses, which appeared to be an adequate method of establishing the linkages. However, these studies did not involve human judgment for linking tasks to KSAOs.

Baranowski and Anderson (2005) conducted a study of the linkage process to examine the influence of rater type on the linkage ratings they provide. Job incumbents, project job analysts (job analysts knowledgeable of the job), and nonproject job analysts (job analysts with little or no knowledge of the job) were recruited from nine different jobs to participate in the study. Despite the limited knowledge possessed by the nonproject job analysts, they were familiar with the linkage rating process. The participants rated linkages that consisted of work behaviors and KSAOs.

Work behavior statements consisted of a set of related tasks. Linkages from four of the jobs were rated on a five-point rating scale while linkages from the other five jobs were rated on a two point scale. The five-point rating scale consisted of the following question, “How important is this knowledge, skill, or ability for performing this work
behavior?” and response options ranged from 1, *Not important*, to 5, *Extremely important*. The two-point rating scale consisted of the following question, “Is this knowledge, skill, or ability used to perform this work behavior?” and response options were either 0, *No*, or 1, *Yes*.

The main finding reached by Baranowski and Anderson (2005) was the large portion of variance in ratings associated with the individual KSAO-task combinations in contrast with the relatively low variance associated with rater type. This finding suggests that one can expect different types of raters to provide similar ratings, thus lending support to the reliability of the ratings they provide. Job analysts, however, provided more reliable ratings overall than job incumbents. Baranowski and Anderson (2005) explain that analysts may provide more reliable ratings due to their better understanding of the linkage process and the theoretical relationship between tasks and KSAOs. Job analysts also possess neutral positions about the job that are free of individual experiences and perceptions, thus making them less prone to the impression management tendencies noted earlier (Green & Stutzman, 1986; Morgeson & Campion, 1997) because they do not possess self-presentation motives and are not vulnerable to social influences. However, it is standard practice to use job incumbents for obtaining job analysis information in order to determine legal defensibility. Job incumbents provide direct information about the job that trained job analysts then revise and edit to prepare the information for job analysis purposes. The results obtained by Baranowski and Anderson (2005) support the use of job incumbents due to the fact that their ratings were sufficiently similar to those of job analysts.
Problem with KSAO-task linkages. The steps of the typically occurring task analysis represent a method analogous to a “crossed design,” a sampling method commonly used in behavioral sciences. In a crossed design, individuals in a sample of the population of interest each respond to the entire set of tasks of the measurement strategy being analyzed (Shavelson & Webb, 1991). These tasks can come in the form of test items, physical actions, or other activities required by examinees. After each examinee completes all items of interest, the resulting responses would be averaged across the examinees to estimate the item parameters, such as item means, variance, and standard errors.

In terms of task analysis, the structure of a crossed design would be imitated by requesting a group of SMEs to each provide ratings to all KSAO-task linkages. Although structuring a task analysis like a crossed design would ensure that every single linkage is rated by every single SME, this process can be overwhelming and even impractical. Although certain criteria allow only the crucial or frequently occurring KSAOs and tasks to be linked, the number of KSAOs and tasks that survive these criteria can still be extremely large. The production of extensive and detailed lists of tasks is the strength of using a task-based method for obtaining job information. However, the number of linkages becomes overwhelming for SMEs to handle (Goldstein & Ford, 2002). For instance, 50 KSAOs and 150 tasks would require SMEs to make 7,500 (50 KSAOs x 150 tasks) linkage ratings. SMEs can be cognitively overloaded with the overwhelming amounts of information they must process. This strain can, in turn, reduce interrater reliability and the validity of the KSAOs inferred (Morgeson & Campion, 1997). As with
any measurement requiring an unreasonable amount of time and energy of its examinees, SMEs would be subject to fatigue and frustration that may result in erroneous judgments. Such errors and other losses in quality would have an impact on the job analysis process since the ratings SMEs provide ultimately guide the development of selection and training standards, among other things. Therefore, this undertaking is impractical and nearly impossible due to limited resources, such as the time and energy of SMEs, efforts of job analysts, and the expenses needed to execute the process.

For practical reasons, and to prevent SMEs from fatigue and frustration, a strategy must be adopted that can reduce the length of the linkage process and the efforts needed to complete it, while maintaining the number of ratings and the accuracy of each rating desired. Goldstein and Ford (2002) have offered a recommendation to help make the rating process more manageable. They suggest that linkages first be made between individual KSAOs and broader levels of tasks, such as task clusters or duties. Then, individual KSAOs can be linked to the individual tasks within the cluster the KSAOs were originally assigned to. The authors reason that this process makes it easier for SMEs to handle the amount of linkages to be rated because for each task cluster, only the KSAOs that were first linked to that cluster need to be dealt with. It is important to determine the most important KSAOs and tasks to the job, as well as determine how helpful each of those important KSAOs is to each important task.

Although Goldstein and Ford (2002) present a way to reduce the overwhelming number of linkages SMEs must rate, their suggestion still results in a time-consuming activity for SMEs because the number of linkages to rate remains extremely large. More
importantly, there needs to be a better way to remedy the problem without losing critical information between certain KSAO-task linkages just for the sake of increasing efficiency. Crucial information about specific linkages desired by the job analyst is lost by reducing the total number of linkages each SME must rate. It may seem possible to simply increase efforts of job analysts and organizations in order to reduce the time and effort needed to complete the linkage process. However, it is more important to obtain all desired linkage ratings while maintaining the needed quality of the judgments. The use of multiple matrix sampling (MMS) theory is proposed as a potential solution to this problem. MMS theory can be applied to the linkage process as a way to address the challenge of the overwhelming number of linkages faced while allowing job analysts to achieve the needed accuracy and efficiency of the obtained ratings. Other than the attempt made by Goldstein and Ford (2002), there has been a lack of literature addressing the problem with the linkage rating process.

Multiple Matrix Sampling Theory

Traditional Sampling Plans

Census taking. Occasionally, researchers are interested in small populations, with which they are able to administer all items of interest to each individual in the population. Whereas the term “population” typically refers to the entire group of individuals a researcher is interested in, the term “universe” is often used to refer to the group of items of interest. A researcher, for instance, may be interested in a population of 10 individuals and a universe of five items each. It would not take long to administer each of these five items to each of the 10 individuals. The researcher would obtain 50 examinee-item
responses (10 individuals, or examinees, x 5 items), a method referred to as “census
taking” that allows the researcher to directly obtain population parameters (Gold &
Basch, 1984). The procedure of census taking is desirable because population parameters
can be computed directly. Figure 1 provides an illustration of census taking. The
population of $N$ examinees the researcher is interested in is to the left of the figure and
the population of $K$ items of interest are at the top. The shaded area reflects the
administration of each item to each individual in the population.
When estimating population parameters, researchers are more frequently interested in larger populations and a larger number of items. Moreover, this extensive process of obtaining information about each item of interest from each individual in a population results in information about individual performance differences, while researchers are primarily concerned with measuring group differences. Differences in individual performance are of little or no importance (Shoemaker, 1973); it is the information about the items themselves that researchers want. When a researcher is interested in a larger population and a larger universe of items, the use of census taking is
impractical. For example, if 300 students each completed a 50 item exam, 15,000 examinee-item responses (300 students x 50 items) would be obtained. This process could take weeks to complete, an amount of time schools are often unwilling to sacrifice or support.

In the job analysis context, if there are only a few tasks, KSAOs, or KSAO-task linkages to be judged, or if the time and effort needed from SMEs is not a constraint, the method of census taking is optimal (Norcini, Shea, & Ping, 1988). Unfortunately, however, organizations are increasingly faced with limited resources, a constraint that makes the collection of all SME judgments in one group impractical.

Examinee and item sampling. Instead of measuring each individual’s answer to every item of interest, population parameters have been estimated traditionally and most often with “examinee sampling,” a method that involves administering all of the items of interest to a sample of the population of interest (Gold & Basch, 1984; Shoemaker, 1973; Shavelson & Webb, 1991; Sirotnik, 1974). For example, if a researcher was interested in obtaining information from the performance of the 300 students on the 50-item exam, examinee sampling could be used by obtaining a sample of the population of students, say 100 of the students, and asking them each to respond to all 50 items. As a result, 5,000 examinee-item responses (100 students x 50 items) would be required, which is relatively less than the 15,000 examinee-item responses required with census taking. The mean and standard deviation of the population would be estimated with the examinee-item responses obtained through examinee sampling. Figure 2 provides an illustration of
examinee sampling. The shaded area again represents the examinee-item responses obtained through this method.

Figure 2. Examinee Sampling. This figure, adapted from Garg (1987), illustrates the method of examinee sampling.

Another approach to sampling, although less commonly used, “item sampling,” entails randomly selecting a sample of items and administering them to each individual in the population of interest (Gold & Basch, 1984; Sirotnik, 1974). That is, 25 of the 50 items would be administered to each of the 300 students, which would result in 7,500 examinee-item responses (25 items x 300 students). Again, the mean and standard
deviation of the population would be estimated with the examinee-item responses obtained. However, this plan is used rarely because researchers are usually more interested in the data of the items rather than that of the examinees (Sirotnik, 1974).

Figure 3 provides an illustration of item sampling. The shaded area again represents the examinee-item responses obtained through this method.

![Figure 3. Item Sampling. This figure, adapted from Sirotnik (1974), illustrates the method of item sampling.](image-url)
MMS. While examinee sampling is used most often in research, it is still an overwhelming process for the examinees involved and requires resources such as time and money that are not always easy to obtain. MMS is a method that combines both examinee sampling and item sampling, providing researchers with a sampling method that reduces the number of examinee-item responses required (Lord & Novick, 1968; Sirotnik, 1974). MMS entails obtaining subsets from the population of interest and administering to each subset a sample of items of interest. A “matrix” would thus be comprised of a sample of the 300 students to each take a sample of the 50 items. For example, a subset of 50 students could each be administered a subset of 10 items, which would require 500 examinee-item responses. The assignment of a subset of the population to a subset of the universe of items results in the matrix. Several other matrices could be obtained until the items of interest have been designated to some subset of examinees. The expression, “multiple matrix sampling” is reflected in the repetition of this process of obtaining matrices by sampling from both examinees and items (Shoemaker, 1973). The data obtained by this procedure would be used to estimate what the mean and standard deviation of the scores would have been had the population of examinees each responded to the population of items (Sirotnik, 1974). Figure 4 provides an illustration of MMS for comparing it with census taking, examinee sampling, and item sampling.
If MMS can be used to estimate population parameters without the need for an overwhelming amount of examinee-item responses, it is beneficial to examine whether or not MMS can be applied to the job analysis context. By establishing evidence that MMS could be used to increase efficiency of the linkage process, the amount of resources could be reduced while maintaining the desired quality and accuracy of estimating population parameters (Gold & Basch, 1984; Norcini et al., 1988).

MMS is a sampling technique that includes the advantages of the traditionally and most commonly used examinee sampling procedure while “allowing for the possibility of
greater scope of coverage” (Gold & Basch, 1984, p. 136). The theory involves an
effective strategy of assessing the performance or behavior of a population of interest,
while being as efficient and useful as examinee sampling for estimating group parameters
(Lord & Novick, 1968; Shoemaker, 1973). MMS allows researchers to estimate the mean
and standard deviation of the population without having to use the same amount of time
and effort required by examinee sampling (Lord & Novick, 1968; Norcini et al., 1988;
Shoemaker, 1973). Lord and Novick (1968) first outlined this procedure in response to
the need for developing test norms that were representative of the population, but
demanded less time per examinee. Subsequent efforts were made (Shoemaker, 1973;
Sirotnik, 1974; Sirotnik & Wellington, 1977) to explain MMS for practitioner use.

The illustrations of census taking, examinee sampling, and item sampling allow
for a visual comparison of these sampling methods with MMS in terms of the way items
are allocated to examinees. In Figure 4, a single shaded box, or “matrix,” represents the
responses obtained by administering a subset of the universe of items to a subset of the
population of examinees. The repeated process of obtaining matrixes results in the
remaining shaded boxes in the figure. Once the various examinee subsets respond to their
respective item subsets, parameters of the examinee population and item universe can be
estimated, which reflect the results that would have been obtained if all items of interest
were administered to each examinee in the population.

The advantage of MMS over the other sampling methods can be observed by
comparing the amount of shaded areas within each figure’s examinee-by-item
arrangement. Although these figures provide illustrations only, one can assume that the
The examinee-by-item arrangement of each figure represents the same examinee population and item universe. Thus, a comparison of the shaded areas in each sampling strategy suggests MMS is the method that requires the least amount of examinee-item responses. The smaller amount of shaded area in MMS as compared with the other sampling plans translates into less effort, time, and money needed as opposed to the other methods.

MMS is also advantageous over examinee sampling, even when each plan results in the number of examinee-item responses. To illustrate this point, Sirotnik (1974) compared the MMS and examinee sampling methods using a hypothetical examinee-by-item arrangement of 25 examinees and 15 items. The examinee sampling plan required each of the 15 items to be administered to a subset of five examinees, whereas the MMS plan required five randomly selected matrices of five examinees to each respond to three of the items. Both sampling plans required the same amount of examinee-item responses, 75 examinee-item responses (15 items x 5 examinees) for the examinee sampling plan, and 75 examinee-item responses (5 matrixes x 3 items x 5 examinees) for the MMS plan (Sirotnik, 1974). However, the matrix samples were more representative of the entire examinee-by-item arrangement because each item is responded to by some examinees and each examinee responds to some items. Five estimates of the population mean were computed for each sampling plan. Because a hypothetical population was used, the actual population mean could be computed, which was .507 (Sirotnik, 1974). The mean of the mean estimates was .499 for the MMS plan and .483 for examinee sampling, with standard deviations of the mean estimates of .076 and .092, respectively (Sirotnik, 1974). These results demonstrated that MMS estimates the mean better than examinee sampling.
and, more importantly, there was less variability among the MMS estimates than among those of examinee sampling.

Applications of MMS

Several studies have demonstrated that MMS is a promising alternative to traditional sampling plans (Lord & Novick, 1968; Shoemaker, 1973; Sirotnik, 1974). MMS theory has been successfully applied in a number of settings, including test development (Gressard & Loyd, 1991), program evaluation (Garg, 1987; Gold & Basch, 1984; Liang, 2003), and setting performance standards (Berk, 1986; Norcini et al., 1988). Results from these studies have supported the capacity of MMS to sufficiently estimate population parameters. Many researchers have explored MMS theory as a way to reduce testing and administration time per examinee. However, MMS has not yet been applied to job analysis, particularly with the linkage process. Although increasing the number of SMEs involved in the process may improve the quality of job analysis data (Berry, 2003), this solution is not always practical or possible. There is a need to explore the feasibility of MMS for addressing the issue of extremely large numbers of KSAO-task linkage ratings in job analysis. The research conducted on MMS in other contexts will be reviewed before discussing the potential of applying MMS to the linkage process.

Test development. In the realm of education, it would seem ideal for an entire population of examinees to respond to an entire universe of items. However, time and money constraints make this desire impractical. Understanding this problem, Gressard and Loyd (1991) attempted to decrease administration time of achievement tests used in educational settings. The researchers developed various sampling plans in which a
proportion of examinees each responded to only a subset of test items. They found that
MMS provided precise mean estimates with only a moderately sized sample.

Research conducted by Gao, Shavelson, and Baxter (1994) supported the
efficiency gained by using MMS for measuring school achievement, in terms of the time
and cost of administration, responding, and scoring. The researchers obtained samples of
students from 40 different schools for a total of 600 students. Each student performed
randomly chosen tasks of a science performance assessment. There were multiple forms
of the tasks and each student received one of these forms. The researchers found that
MMS allowed more tasks to be used, which translated into fewer tasks for each student to
complete and less of a response burden. By using MMS, the time and cost of
administration, responding, and scoring of the science performance assessment were
more efficient. Although a large number of tasks and students were needed, there were
fewer tasks assigned to each student than if each student completed all tasks.

Garg, Boss, & Carlson (1986) also explored the capacity of MMS for estimating
item parameters by comparing MMS with examinee sampling. The researchers used a
Monte Carlo approach to simulate examinee data. Results of their study indicated MMS
is an acceptable alternative for providing item parameter estimates. The researchers noted
that when the number of items of interest is large and it is difficult to obtain a
representative sample of examinees of adequate size to respond to all of the items, MMS
provides a sufficient alternative. When a large number of observations is obtained, such
as an examinee sample size that is 10 times the number of items, the item parameter
estimates obtained through MMS are as good as those obtained by examinee sampling. If
this criterion is applied to job analysis, obtaining a substantially smaller sample size may lead a researcher to underestimate parameters of job analysis ratings or obtain misleading results.

Program evaluation. MMS has also been used to evaluate program effectiveness. Liang (2003) recognized the challenge in determining a group’s voice instead of individual attitudes and opinions, especially since the collection of a large number of participants can be a costly and timely undertaking. However, MMS was helpful in increasing the efficiency of data collection while maximizing the coverage of test items. Despite the risk of gaining efficiency at the cost of accuracy, the use of MMS confirmed this was not a problem. Liang (2003) noted the importance of minimizing intrusion into students’ learning; reducing administration time of an assessment or other measurement tool facilitates increased willingness from schools to participate in such research.

Using a Monte Carlo simulation, Garg (1987) explored examinee sampling and MMS. Population parameters were estimated 50 times and results were compared with a true population of 5,000 examinees. There were no systematic differences identified between examinee sampling and MMS. In fact, MMS was found to be as efficient and useful as examinee sampling. Therefore, since examinee sampling is often a difficult method to pursue in terms of time and resources, MMS provides a satisfactory alternative for completing the linkage process in the current study. Based on Garg’s (1987) results, it is possible that accurate estimates of linkage parameters could likewise be obtained through MMS more efficiently than traditional methods.
Gold and Basch (1984) attempted to replace traditional sampling with MMS for the collection of data relevant to health education. The researchers conducted a field study to compare the efficiency of MMS to evaluate a health education program with that of examinee sampling. Results indicated that MMS provided sufficient estimates of population parameters that were as useful as those obtained from examinee sampling. No significant differences were found in the estimates from both sampling strategies.

*Setting performance standards.* MMS has been applied to the process of setting performance standards, namely “cutting scores,” which refer to scores that distinguish the difference between performance that is minimally acceptable and performance that is not acceptable (Norcini et al., 1988). Typical methods for setting cutting scores in competency testing involve gathering SMEs and obtaining judgments from each individual on each item of interest (Angoff, 1971; Berk, 1986). The average of these judgments usually results in the cutting score. For example, a group of SMEs would identify a group of examinees whose performance is considered “borderline” and make judgments as to the percentage of this group that should provide correct responses to one or more items. The cutting score would be based upon these estimates of borderline examinee performance.

With large scale testing, the use of MMS presents a more efficient method of producing a quality cutting score under time and financial constraints (Norcini et al., 1988). Applying MMS to the task of setting performance standards would entail having subsets of SMEs estimate the performance standards for various subsets of items. The expected mean over the groups of SMEs and items would then be the basis for the cutting
score. This process is analogous to having groups of examinees complete subsets of items in test development. However, rather than obtaining averages of examinee performance to estimate population parameters, averages of SME judgments are collected. Since job analysis ratings of the current study involve SME judgments, the method utilized by Norcini et al. (1988) is given careful attention. There is a lack of research on the effectiveness of applying MMS to methods requiring SME judgments.

Norcini et al. (1988) explored the use of MMS as an alternative to typical methods for setting performance standards by using a variant of Angoff’s method (1971). From one particular module on a medical recertifying examination, 190 items were chosen for the study. These items were divided into approximately five equal groups of 31 to 43 items. Thirty-six physician-experts served as SMEs and were each randomly assigned to one of the five groups of items. Reliable results were obtained despite the fact that only a relatively few number of SMEs were used, which did not fit Garg’s (1987) criterion for sample size. Each SME group was asked to provide an estimate of internist borderline performance for their subset of items. For the first item, SMEs were asked to make individual judgments about the proportion of borderline internists who they felt would provide a correct response. The SMEs with the highest and lowest estimates gave justifications of their decisions to the group and a general discussion followed. After this discussion, each SME was allowed to change their estimates. Each group followed this process for the same 67 items, which were not included in the study. These items were used as practice and a way to evaluate the consistency of SMEs in their judgments. The
remaining items, including the 190 analyzed in the study, along with instructions were taken by SMEs to complete the judgments at home.

The expected mean of SME judgments were taken as the cutting score for the examination (Norcini et al., 1988). There was some variability in the cutting scores estimated by each group. However, these estimates were reasonably stable as they ranged from .587 to .654, which translated into raw test scores of 111.5 to 124.3. The standard error of the cutting scores among the five groups ranged from 4.2 items to 5.5 items, which reflected a 90% confidence interval of only 1.3 items. The Coefficient alpha of the final cutting score was .98. The results of this study were encouraging and supported the use of MMS to obtain accurate and useful standard-setting data.

Addressing the Needs of Job Analysis with MMS

MMS provides a potential solution to obtaining the necessary ratings of the numerous KSAO-task linkages produced by job analysis. Considering the efficiency and benefits of applying MMS in other settings, MMS may reduce the overwhelming amount of time and effort currently needed to complete the linkage process. Moreover, it addresses the drawback of Goldstein and Ford’s (2002) compromise of having SMEs rate linkages between KSAOs and certain task clusters, rather than each individual task. Just as it is unreasonable to ask a sample of examinees to each respond to all items of interest, it is clear that requesting a sample of SMEs to rate all KSAO-task linkages for a given job is just as impractical. If MMS can be successfully applied to the linkage process, subsets of SMEs may only need to each rate subsets of linkages. MMS could help job
analysts obtain the desired number of specific KSAO-task linkage ratings while maintaining the desired accuracy of those ratings.

Since MMS has not yet been applied to the linkage process, it should be explored to determine if its application would be fitting in such a context. Positive findings would have important implications for organizations. It is of theoretical and practical importance to determine if MMS is a feasible strategy to apply to the linkage process.

In order to apply MMS to the linkage process, the entire set of linkages that would need to be rated would be divided into subsets. A sample of SMEs would then be obtained and each SME randomly assigned to rate one of the linkage subsets. The ratings would be used to estimate the parameters of the linkages that would have been obtained if the same group of SMEs had each rated all linkages.

When applied to other contexts, MMS has enabled researchers to acquire estimates of item parameters that are comparable to estimates obtained with traditional sampling procedures (Garg et al., 1986; Gressard & Loyd, 1991; Norcini et al., 1988). Therefore, it is expected that applying MMS to the linkage process will result in estimates of linkage ratings that are comparable to linkage ratings obtained through traditional procedures. In order to evaluate the effectiveness and accuracy with which MMS can be applied to the linkage process, the reliability of linkage ratings obtained through this sampling procedure can be assessed.

Reliability of Linkage Ratings Obtained through MMS

The precision with which population parameters are estimated is based largely upon the degree to which a sampling procedure is appropriate and the measurement
procedure is reliable (Gold & Basch, 1984, p. 136). Despite all efforts to obtain information in an accurate and thorough manner, the information obtained from a measure is subject to error to some degree (Brennan, 2001a). Thus, the use of any measurement procedure deserves careful attention. For instance, measuring the speed of a runner is fallible. The measurements obtained may vary depending on various measurement conditions, such as the stopwatch used, the person who records the measurements, or the weather conditions.

One way of increasing the precision of measurement is to obtain an average of measurements over some subset of predefined conditions of measurement (Brennan, 2001a). An average measurement can be obtained through this process, which can serve as an estimate of an ideal measurement that would be hypothetically obtained over all conditions of the measurement. Another way of increasing measurement precision is to fix one or more conditions of measurement (Brennan, 2001a). Doing so decreases the variability in obtained information that is attributable to the fixed variable. This, however, restricts the set of measurement conditions to which the results of the measurement procedure can be generalized. Thus, while fixing a condition of measurement may increase precision, the interpretation of the measurement is confined.

Error does not suggest a mistake has occurred. Rather, it is an aspect of measurement that must be quantified and for which conditions attributable to it must be specified (Brennan, 2001a). This perspective of measurement falls under the classical
theory of reliability. Reliability generally involves “the quantifying of consistencies and inconsistencies” (Brennan, 2001a, p. 2). It is easy for researchers to be deceived when obtaining information from only one element among a large set of interrelated elements.

Generalizability Theory

An Extension of Reliability

Generalizability (G) theory offers a larger perspective within which reliability analysis should be viewed (Cronbach & Shavelson, 2004). G theory is a statistical theory that enables researchers to evaluate the dependability of behavioral measurements (Brennan, 2001a; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). G theory allows researchers to carefully examine different variables of a measurement procedure that may contribute to measurement error. The advantage of G theory over classical test theory, which is traditionally used to evaluate measurement issues, is the opportunity for extensive assessments of the components involved in research studies.

According to the classical theory of reliability, only one source of variance of a given measurement tool can be evaluated at a time (Furr & Bacharach, 2008; Kline, 2005). According to classical test theory, each observation obtained from a measurement procedure can only be broken down into a “true” score and a single, undifferentiated random error term. Consequently, multiple sources of error cannot be separated through a single application of the classical test theory model. The error term is seen as one value that encompasses measurement error due to all variables. If a researcher’s study involves items and observers as variables, the error due to items would be indistinguishable from
the error due to observers. Thus, the application of classical test theory is limited for the purpose of evaluating multiple sources of error. In the context of obtaining job analysis ratings, it would then be insufficient to apply classic test theory when various sources of error are likely to contribute to the rating.

G theory extends the concept of reliability by further investigating the effects of each variable included in a measurement procedure (Brennan, 2001a; Cronbach et al., 1972; Furr & Bacharach, 2008; Shavelson & Web, 1991). G theory extends classical test theory by applying certain ANOVA procedures to measurement tools. A G study entails the partitioning of variance, a procedure similar to the ANOVA. When multiple sources of error may potentially contribute to results of a measurement procedure, G theory enables a researcher to identify those various sources of error. G theory has the ability to differentiate between different variables, which are referred to as “facets.” For example, error in a measurement tool may be due to a single facet, the interaction of two facets with each other, or combinations of numerous facets. Therefore, rather than resulting in only one error term, G theory disentangles the various sources of error, allowing the researcher to fully evaluate measurement procedure.

Using G Theory to Evaluate Job Analysis Ratings

Information obtained through job analysis is subject to error because it is mainly based on human judgment (Nathan & Alexander, 1985; Rothstein & Jackson, 1980). SMEs are considered experts of the job and assumed to be capable of providing accurate and complete information regarding the KSAs and tasks under consideration (Landy & Vasey, 1991). However, the dependence on SMEs raises questions about their status as
experts since their individual experiences and perceptions of the job have been demonstrated to influence the job analysis ratings they provide, even if unintentionally or unconsciously (e.g., Morgeson & Campion, 1997; Tross & Maurer, 2000). Additionally, SMEs who share the same job may still provide different ratings due to their current assignments (Landy & Vasey, 1991).

As with any measurement procedure, the accuracy and reliability of applying MMS to the linkage process must be evaluated. The amount of measurement error that affects the MMS process serves as an indicator of this accuracy. The presence of measurement error can be carefully assessed to determine which sources influence the KSAO-task linkage ratings. Information obtained from the linkage process ultimately guides the development of selection and training standards. As a critical part of the job analysis, it is important to determine the extent to which the actual linkage ratings are as useful as linkage ratings obtained if a group of SMEs were to each rate an entire set of linkages.

G theory can be used to evaluate the application of MMS to the linkage process. Potential facets that contribute to the variance in linkage ratings can be analyzed to evaluate the overall effects of these facets on themeasurement procedure of MMS. In the context of the linkage process, SMEs and actual linkages between the KSAOs and tasks, for instance, would represent these various facets.

Murphy and DeShon (2000b) promote the use of G theory for analyzing job performance ratings, reasoning that SME raters naturally have varying experiences and perceptions of the workplace due to the unique roles and responsibilities of their
positions. The classic theory of reliability assumes measurement error is uncorrelated to true scores, but this may not be the case with job analysis ratings, which have been found to reflect true differences in raters (Harvey, 1991; Nathan & Alexander, 1985; Rothstein & Jackson, 1980). Thus, it is insufficient to use the classical theory of reliability to evaluate job analysis ratings.

Murphy and DeShon (2000a) emphasize that G theory was developed in order to estimate the reliability of ratings obtained in contexts where multiple systematic sources of variance in ratings may be present. The authors advocate the application of G theory methods in order to obtain desired estimates of the reliability of job performance ratings. Analyses based on G theory are more useful and informative than those performed with the classic theory of reliability.

Baranowski and Anderson (2005) used G theory to examine the sources of variation in ratings of linkages between work behaviors to KSAOs and better understand how the ratings varied. Their analyses suggested that the variability in ratings was largely due to the combinations of work behaviors to KSAOs whereas relatively low variance was attributed to the actual raters themselves. This finding suggested that the differences in ratings were due to differences in the work behaviors and KSAOs rather than the individuals who provided the ratings. Using G theory, the researchers were able to identify and evaluate potential sources of error.

Framework of G Theory

(Universe of admissible observations and G studies. When developing a measurement procedure or instrument to measure a certain construct, a researcher first
identifies various facets, or sources, of potential measurement error that can be used to evaluate the construct (Brennan, 2001a). The researcher does not necessarily need to commit to actually using any facet in particular or any number of conditions for a certain facet. The researcher simply needs to characterize the facets. Doing so indicates that the researcher simply considers acceptable any one condition of any given facet. The researcher is establishing that any one condition of a given facet constitutes an “admissible condition” of measurement for that facet. Thus, the researcher’s “universe” of admissible observations contains each facet the researcher is considering for the measurement procedure.

Suppose a researcher is interested in a construct such as the writing ability of college freshmen. In developing a measurement procedure, the researcher may decide to evaluate writing ability in students by asking them to complete various tasks (\(t\)), such as responding to writing prompts, and have their performance assessed by trained raters (\(r\)). The tasks and raters would represent two facets in the researcher’s measurement procedure since these variables may potentially contribute to the error found in the students’ writing ability scores. Although the students, or persons (\(p\)), participating in the study can be considered another source of error, they would be referred to as the “object of measurement.” In G theory, while the term “universe” is used when describing conditions of measurement, the term “population” is used to describe the objects of measurement. Individuals being administered the measurement procedure are typically objects of measurement, but other entities can be the objects of measurement, such as test items.
**Fully crossed design.** Scores of writing ability in the researcher’s study would be the result of pairing each person to each task and to each rater. If the research considers the scores of writing ability from each person in the study as acceptable, then the design of this study would be described as “crossed.” This design can be represented in the following formula: \( p \times t \times r \), which indicates that each person is crossed, or evaluated by, each task and each rater, to result in each student’s score of writing ability. The main effects of this design are \( p, t, \) and \( r \), and the interaction effects are \( pt, pr, tr, \) and \( ptr \). Any student’s observable score for a single condition in the first facet and single condition in the second facet can be represented as:

\[
X_{ptr} = \mu + v_p + v_t + v_r + v_{pt} + v_{pr} + v_{tr} + v_{ptr} \tag{1}
\]

where \( \mu \) is the grand mean in the population of college freshmen, and \( v \) is any one of the seven effects for this design. The variance of the scores given by this equation, over the population of individuals and the conditions in the universe of admissible observations is:

\[
\sigma^2(X_{ptr}) = \sigma^2(p) + \sigma^2(t) + \sigma^2(r) + \sigma^2(pt) + \sigma^2(pr) + \sigma^2(tr) + \sigma^2(ptr), \tag{2}
\]

where \( \sigma^2 \) is a G study random effects variance component for any one of the seven effects. This equation illustrates that the total observed score can be broken down into seven independent variance components. Assuming the population and both facets in the universe of admissible observations are very large and theoretically approach infinity, the variance components are called “random effects” variance components.

After the researcher specifies the population and universe of admissible observations for the \( p \times t \times r \) design, data need to be collected and analyzed to estimate the variance components. The researcher must conduct a G study by obtaining a sample
of \( n_p \) students in the population to whom \( n_t \) conditions of the first facet and \( n_r \) conditions of the second facet can be administered. This particular design is considered a two-facet design because the objects of measurement, which typically are the individuals being assessed by the measurement procedure, are not usually referred to as facets.

To estimate the random effects variance components in the \( p \times t \times r \) design, the expected mean square, \( E(\text{MS}) \), of each effect in Equation 1 must be determined using the equations in the first column of Table 1. The equations in the second column of Table 3 are the intermediary estimators of variance components in terms of mean squares from an ANOVA or other variance component estimation method.

Table 1

*Expected Mean Squares and Random Effects Variance Components for the Fully Crossed Design*

<table>
<thead>
<tr>
<th>Effect</th>
<th>( E(\text{MS}) )</th>
<th>( \hat{\sigma}^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p )</td>
<td>( \sigma^2(p) + n_p \sigma^2(pr) + n_r \sigma^2(pt) + n_p n_r \sigma^2(p) )</td>
<td>( \frac{\text{MS}(p) - \text{MS}(pt) - \text{MS}(pr) + \text{MS}(ptr)}{n_p n_r} )</td>
</tr>
<tr>
<td>( t )</td>
<td>( \sigma^2(ptr) + n_p \sigma^2(pt) + n_r \sigma^2(pr) + n_p n_r \sigma^2(t) )</td>
<td>( \frac{\text{MS}(t) - \text{MS}(pt) - \text{MS}(tr) + \text{MS}(ptr)}{n_p n_r} )</td>
</tr>
<tr>
<td>( r )</td>
<td>( \sigma^2(ptr) + n_p \sigma^2(pr) + n_r \sigma^2(pt) + n_p n_r \sigma^2(r) )</td>
<td>( \frac{\text{MS}(r) - \text{MS}(pr) - \text{MS}(tr) + \text{MS}(ptr)}{n_p n_r} )</td>
</tr>
<tr>
<td>( pt )</td>
<td>( \sigma^2(ptr) + n_p \sigma^2(pt) )</td>
<td>( \frac{\text{MS}(ptr) - \text{MS}(pr)}{n_r} )</td>
</tr>
<tr>
<td>( pr )</td>
<td>( \sigma^2(ptr) + n_r \sigma^2(pr) )</td>
<td>( \frac{\text{MS}(ptr) - \text{MS}(pr)}{n_t} )</td>
</tr>
<tr>
<td>( tr )</td>
<td>( \sigma^2(ptr) + n_p \sigma^2(tr) )</td>
<td>( \frac{\text{MS}(ptr) - \text{MS}(tr)}{n_p} )</td>
</tr>
<tr>
<td>( ptr )</td>
<td>( \sigma^2(ptr) )</td>
<td>( \text{MS}(ptr) )</td>
</tr>
</tbody>
</table>

*Note.* Adapted from Brennan (2001a).
The estimated variance components are estimates of the actual variances in Equation 2. For example, $\hat{\sigma}^2(p)$ is an estimate of the variance component, $\sigma^2(p)$. This estimate, as well as the estimates for the other main effect variance components, $\hat{\sigma}^2(r)$ and $\hat{\sigma}^2(t)$, can be interpreted in the following manner. If each person’s mean score on the measurement procedure could be obtained over all levels of the facet $r$ and all levels of the facet $t$ in the universe of admissible observations, $\sigma^2(p)$ would be the variance of these mean scores over the entire population of persons (Brennan, 2001a). Thus, $\hat{\sigma}^2(p)$ would be an estimate of $\sigma^2(p)$. Interaction variance components are interpreted as the extent to which differences in the relative ordering of a given facet are due to the various levels of another facet (Brennan, 2001a). For example, $\hat{\sigma}^2(pt)$ can be interpreted as the extent to which the ordering of persons differs due to differences in the facet, $t$.

**Measuring of reliability.** Variance components other than $\sigma^2(p)$ contribute to one or more different types of error variance, such as absolute and relative error variances (Brennan, 2001a; Shavelson & Webb, 1991). Absolute error variance refers to the difference between a person’s observed score and universe score, or the absolute level of individuals’ scores. Relative error variance refers to the difference between a person’s observed deviation score and universe deviation score, or the relative standing or ranking of individuals.

Absolute and relative error variance can be used to determine two types of G theory coefficients, a generalizability coefficient and index of dependability, which are similar to reliability coefficients (Brennan, 2001a). These values have a range of zero to one, with higher values indicating a higher degree of consistency among the ratings given
by the SMEs than lower values. A generalizability coefficient, \( E \rho^2 \), is the ratio of the universe score variance to the sum of the universe score variance and relative error variance. The generalizability coefficient is analogous to the reliability coefficient in classical test theory (Brennan, 2001a; Shavelson & Webb, 1991). The index of dependability, \( \Phi \), is the ratio of the universe score variance to the sum of the universe score variance and absolute error variance. The index of dependability is more analogous to a measure of interrater agreement.

*Partially nested designs.* Due to practical constraints, a “partially nested” design may be necessary, in which a researcher may only be able to have each level of one facet evaluated by a certain number of levels of another facet (Cronbach et al., 1972). For example, the desire to have each rater evaluate each task completed by each student is limited due to financial restraints and the amount of time needed to complete such a study. The researcher may then decide that, for each task, only a certain number of raters will evaluate each student. This partially nested design would be represented as a \( p \times (r:t) \) design, where “:” is read “nested within” (Brennan, 2001a). The main effects of this design are \( p \), \( t \), and \( r:t \), and the interaction effects are \( pt \), and \( pr:t \). Any observable score in this design can be broken down into the various effects of the \( p \times (r:t) \) design:

\[
X_{ptr} = \mu + v_p + v_t + v_{rt} + v_{pt} + v_{prt},
\]  

[3]

The concepts and procedures used for the crossed design apply to nested designs. For the \( p \times (r:t) \) design, the total variance is the sum of the five independent variance components:

\[
\sigma^2(X_{ptr}) = \sigma^2(p) + \sigma^2(t) + \sigma^2(r:t) + \sigma^2(pt) + \sigma^2(pr:t).
\]  

[4]
Table 2 presents the equations for estimating the expected mean square $E(MS)$ of each effect in Equation 3 and estimators of the random effects variance components for the $p \times (r:t)$ design.

### Table 2

**Expected Mean Squares and Random Effects Variance Components for the Partially Nested Design**

<table>
<thead>
<tr>
<th>Effect</th>
<th>$E(MS)$</th>
<th>$\widehat{\sigma}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td>$\sigma^2_{(pr:t)} + n \sigma^2_{(pt)} + n_p n_t \sigma^2_{(p)}$</td>
<td>$\frac{MS(p) - MS(p)}{n_t}$</td>
</tr>
<tr>
<td>$t$</td>
<td>$\sigma^2_{(pr:t)} + n \sigma^2_{(pt)} + n_p \sigma^2_{(r:t)} + n_p n_t \sigma^2_{(t)}$</td>
<td>$\frac{MS(t) - MS(pr:t) - MS(pt) + MS(pr:t)}{n_p}$</td>
</tr>
<tr>
<td>$r:t$</td>
<td>$\sigma^2_{(pr:t)} + n_r \sigma^2_{(r:t)}$</td>
<td>$\frac{MS(r:t) - MS(pr:t)}{n_r}$</td>
</tr>
<tr>
<td>$pt$</td>
<td>$\sigma^2_{(pr:t)} + n \sigma^2_{(r:t)}$</td>
<td>$\frac{MS(pt) - MS(pr:t)}{n_r}$</td>
</tr>
<tr>
<td>$pr:t$</td>
<td>$\sigma^2_{(pr:t)}$</td>
<td>$MS(pr:t)$</td>
</tr>
</tbody>
</table>

*Note.* Adapted from Brennan (2001a).

**MMS Design.** Additionally, an MMS design may be favorable if the size of a researcher’s population of interest is too large. For example, the researcher may want to apply MMS to obtain accurate estimates of writing ability without having to gather an overwhelming number of students and ask them to each complete an extremely large number of tasks. Table 3 presents the equations for estimating the mean squares and random effects variance components for the MMS design, $r:(p \times t)$, where $p$ is crossed with $t$ and nested within $r$. According to this design, different raters are each being assigned to evaluate specific tasks completed by different groups of students. This design
reduces the number of scores of writing ability that the researcher needs to obtain without affecting the accuracy of estimating the population parameters.

Table 3

*Expected Mean Squares and Random Effects Variance Components for the Multiple Matrix Sampling Design*

<table>
<thead>
<tr>
<th>Effect</th>
<th>$E(\text{MS})$</th>
<th>$\hat{\sigma}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
<td>$\sigma^2(pt:r) + n_p \sigma^2(t:r) + n_t n_r \sigma^2(p:t) + n_p n_t n_r \sigma^2(r)$</td>
<td>$\frac{\text{MS}(r) - \text{MS}(p:r) - \text{MS}(t:r) + \text{MS}(pt:r)}{n_p n_t}$</td>
</tr>
<tr>
<td>$p:t$</td>
<td>$\sigma^2(pt:r) + n_p \sigma^2(p:t)$</td>
<td>$\frac{\text{MS}(p:t) - \text{MS}(pt:r)}{n_t}$</td>
</tr>
<tr>
<td>$t:r$</td>
<td>$\sigma^2(pt:r) + n_p \sigma^2(t:r)$</td>
<td>$\frac{\text{MS}(t:r) - \text{MS}(pt:r)}{n_p}$</td>
</tr>
<tr>
<td>$pt:r$</td>
<td>$\sigma^2(pt:r)$</td>
<td>$\text{MS}(pt:r)$</td>
</tr>
</tbody>
</table>

*Note.* Adapted from Brennan (2001a).

**Summary**

MMS and G theory provide a promising solution to needing an overwhelming number of linkage ratings during job analyses. In order to evaluate the application and efficiency of these theories, data gathered from a state correctional agency will be used. A previous job analysis was conducted for this agency, in which SMEs were asked to make linkages between the important KSAOs and tasks of the job.

If the application of MMS and G theory to this context is feasible, it could solve the dilemma of obtaining overwhelming numbers of ratings, one of the biggest problems faced in the linkage process. Time and effort could be reduced while preserving the accuracy necessary to interpret ratings for the development of standards for selection and
training. Positive results of this study would add knowledge to the literature surrounding these topics, especially since MMS has not yet been applied to the linkage process. G theory would provide an improved strategy for evaluating the reliability of this method and confirming its benefit to the job analysis process.

As mentioned earlier, Goldstein and Ford (2002) have so far responded to this challenge by reducing the process to the linking of individual KSAOs to only tasks within task clusters determined to be important to those KSAOs. The application of MMS, on the other hand, presents a better alternative by sustaining the detail of linkages between each KSAO to each task. With MMS, all desired ratings could be made without asking SMEs to take on such a difficult task. Organizations would gain an efficient method for developing selection and training standards, while being ensured of the quality and accuracy they desire, all without the extensive and impractical amount of time, effort, and resources demanded by current practices.
Chapter 2

METHOD

Archival data gathered from a job analysis conducted by the Corrections Standards Authority (CSA, 2007) were used. The CSA is a state correctional agency that is responsible for developing and monitoring selection and training standards for designated state correctional peace officer classifications. Of the 47 state peace officer classifications this responsibility encompasses, a job analysis study was conducted for three of them, the Correctional Officer (CO), Youth Correctional Officer (YCO), and Youth Correctional Counselor (YCC). The CO is the largest correctional peace officer classification in the state corrections system and is primarily responsible for public protection. The YCO is the largest correctional peace officer classification in the Juvenile Justice division, followed by the YCC classification. The primary responsibility of the YCO is for security of juvenile facilities and supervision of juvenile offenders. The YCC is responsible for counseling, supervision, and maintaining custody of juvenile offenders.

The intention of the job analysis the CSA conducted was to determine a baseline for setting the selection and training standards for the CO, YCO, and YCC classifications (CSA, 2007). The CSA staff chose to include these three classifications in the same job analysis study on the hypothesis that there is overlap in their tasks performed on the job (CSA, 2007).
Participants

Twenty-eight individuals throughout the state of California served as SMEs to perform the linkage process stage of the job analysis (CSA, 2007). The SMEs were at the level of supervisor or higher except for one individual, who was representing a supervisor unable to participate in the linkage process. These individuals were selected based on the reasoning that they possess a broader view of the vision, mission, and values of the correctional agency and are in the best position to accurately link tasks and KSAOs. The SMEs were chosen to represent the three correctional job classifications for which the job analysis was conducted. This sample correctly represented the populations of the three classifications throughout the State. Table 4 provides characteristics of the SME sample.
<table>
<thead>
<tr>
<th>Characteristics of the SME Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification</strong></td>
</tr>
<tr>
<td>Correctional Officer (CO)</td>
</tr>
<tr>
<td>( n_{CO} = 15 ) (53.6%)</td>
</tr>
<tr>
<td>Youth Correctional Officer (YCO)</td>
</tr>
<tr>
<td>( n_{YCO} = 6 ) (21.4%)</td>
</tr>
<tr>
<td>Youth Correctional Counselor (YCC)</td>
</tr>
<tr>
<td>( n_{YCC} = 7 ) (25.0%)</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>( n_{male} = 21 ) (75.0%)</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>( n_{female} = 7 ) (25.0%)</td>
</tr>
<tr>
<td><strong>Age</strong></td>
</tr>
<tr>
<td>( M )</td>
</tr>
<tr>
<td>43.3 years</td>
</tr>
<tr>
<td>( SD )</td>
</tr>
<tr>
<td>8.1 years</td>
</tr>
<tr>
<td><strong>Ethnic Background</strong></td>
</tr>
<tr>
<td>Asian</td>
</tr>
<tr>
<td>( n_{asn} = 3 ) (10.7%)</td>
</tr>
<tr>
<td>Black/African American</td>
</tr>
<tr>
<td>( n_{blk} = 5 ) (17.9%)</td>
</tr>
<tr>
<td>Hispanic</td>
</tr>
<tr>
<td>( n_{his} = 9 ) (32.1%)</td>
</tr>
<tr>
<td>White</td>
</tr>
<tr>
<td>( n_{wht} = 11 ) (39.9%)</td>
</tr>
</tbody>
</table>
Procedure

*MMS Plan*

During the task analysis stage of the job analysis, a different group of SMEs generated 285 tasks (CSA, 2007). The 28 SMEs of the current study reviewed these tasks and identified 122 KSAOs required for successful performance of each task. Each SME was provided with a pre-assigned group of tasks to rate according to an MMS plan developed for the linkage process of the current study.

Based on the MMS plan developed by Norcini et al. (1988), a modified MMS plan was developed for the current study. The 28 SMEs were divided into 14 pairs of raters. Each pair was assigned to rate one of 14 sets of KSAO-task linkages (CSA, 2007). Therefore, each SME was assigned to rate approximately 42 tasks along the 122 KSAOs (approximately 5,124 linkages each), of which 23 were common tasks were assigned to all SMEs (2,806 linkages). These 23 common tasks were rated by each SME in order to evaluate rater consistency. If the SMEs rated these 23 common tasks consistently, it
would be reasonable to assume the remaining tasks divided among the SME pairs would be rated in a similar manner.

Following the examinee-by-item matrix structure for typical MMS sampling plans in Figure 2, the current study utilized a KSAO-task linkage-by-SME matrix structure. Figure 5 illustrates the modified MMS plan. The shaded area on the left of the matrix represents the common linkages rated by each SME. The smaller shaded areas represent the pairs of SMEs and their respective set of non-common linkages they were assigned to rate.
Figure 5. SME-by-Linkage Matrix: Modified Sampling Plan. This figure illustrates the SME-by-linkage matrix sampling plan for the current study. SMEs each rated 2,806 common linkages; then pairs of SMEs rated subsets of the remaining non-common linkages.

This plan was developed in order to be feasible within the operational constraints of the correctional agency. While the use of two SMEs to rate each linkage was less than originally desired, it was necessary for fitting the process within time constraints faced by the agency (CSA, 2007). Assuming SMEs would each rate four linkages per minute, 5,124 linkage ratings (42 tasks each linked to 122 KSAOs) were believed require a total
of 1,270 minutes, or approximately 21 hours, to complete. Thus, assigning more SMEs per task was determined to not be feasible.

The use of two SMEs per linkage rating was justified on three grounds (CSA, 2007). First, Baranowski and Anderson (2005) concluded from their study of the linkage process that it is unnecessary to use a large number of individuals to ensure linkage ratings are reliable. The samples in their study frequently consisted of two to three SMEs. Second, two SMEs per linkage was actually a minimum sample size in the current study (CSA, 2007). Data were screened live as they were collected, which flagged discrepancies in ratings. The linkages with discrepancies were assigned to a third SME to be rated. Thus, each linkage was actually rated by two or three SMEs, except for the common tasks which were rated by all SMEs. Additionally, samples of this size were consistent with the SME samples used in Baranowski and Anderson’s (2005) methodology, which were approximately the same size and yielded acceptable reliabilities. Third, systematic differences between the small SME samples were assessed to identify potential biases in linkage ratings. These differences were assessed with the 23 common tasks, which were rated by each SME. As previously noted, these common tasks were included to determine if the SMEs were rating linkages consistently, which would provide support for the assumptions that the SMEs were rating the remaining tasks in a similar manner.
Rating the Linkages

The SMEs received training on the linkage process. For the common KSAO-task linkages, SMEs were instructed to work independently. For the non-common linkages, the SME pairs were instructed to continue working independently, and then discuss each rating with their partner to reach a consensus rating, when possible.

For each linkage, SMEs were instructed to indicate the relevance of each KSAO for successful performance of each task. The SMEs were given rating forms to record their ratings for their assigned KSAO-task linkages. Figure 6 provides the scale SMEs responded to for each linkage (CSA, 2007). For each KSAO-task linkage, SMEs were asked to indicate if the KSAO was “not relevant” to the associated task, “helpful” for performing the task successfully, or “essential” for performing the task successfully; these ratings were coded as a 0, 1, or 2, respectively, in order to be evaluated statistically.
### KSAO-Task Linkage Rating Scale

<table>
<thead>
<tr>
<th>Rating</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Not relevant</td>
</tr>
<tr>
<td>1</td>
<td>Helpful</td>
</tr>
<tr>
<td>2</td>
<td>Essential</td>
</tr>
</tbody>
</table>

*Figure 6. KSAO-Task Linkage Rating Scale. This figure provides the rating scale options SMEs used to rate the KSAO-task linkages (CSA, 2007).*

SMEs were also asked to indicate their familiarity with each task by indicating whether or not they were able to link rate the task either because they have performed it or because they have supervised someone performing it (CSA, 2007). SMEs were instructed to only rate tasks for which they answered positively on this certification statement in order to prevent SMEs from rating tasks with which they were unfamiliar.

**Data Analysis**

A G study was performed to disentangle the various facets in the designs for the common and non-common linkages. The results of the G study were obtained using the software program, urGENOVA (Brennan, 2001b). urGENOVA is a program for estimating variance components of both balanced and unbalanced designs. urGENOVA was especially created to handle designs in which different numbers of items are nested.
within each level of a particular facet, as well as crossed or nested designs with missing data. Furthermore, MMS designs can be analyzed by urGENOVA. Therefore, this program was very appropriate because the designs in the current study involved nesting and MMS.

Analysis of common items. The variance components of the design for the common KSAO-task linkage ratings were assessed to determine how consistently the SMEs were rating the linkages. If a consensus was identified among the SMEs across these common linkages, it would be assumed that the SMEs would rate the remaining non-common linkages in the same manner.

Although people are typically the object of measurement in a G study, the KSAO-task linkages (t) were the objects of measurement in the current study. The facets in the common linkage design were SME rater (s) and job classification (c). Thus, the design of the analysis for the common tasks was specified as a t x (s:c) design. The notation “s:c” indicates the SME raters were “nested within” classification, which is due to the fact that each rater belonged to only one of the three job classifications. This design can be read as “KSAO-task linkage by SME rater nested within classification.” Table 5 presents the equations for deriving the mean squares and variance components for the common KSAO-task linkage ratings.
Table 5
Expected Mean Squares and Random Effects Variance Components for the Common Linkage Ratings

<table>
<thead>
<tr>
<th>VC</th>
<th>$E(\text{MS})^b$</th>
<th>$\hat{\sigma}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>$\sigma^2(ts:c) + n_s \sigma^2(tc) + n_p \sigma^2(t)$</td>
<td>$\frac{\text{MS}(t) - \text{MS}(t)}{n_p n_s}$</td>
</tr>
<tr>
<td>$c$</td>
<td>$\sigma^2(ts:c) + n_s \sigma^2(tc) + n_p \sigma^2(s:c) + n_p n_s \sigma^2(c)$</td>
<td>$\frac{\text{MS}(c) - \text{MS}(s:c) - \text{MS}(tc) + \text{MS}(ts:c)}{n_p n_s}$</td>
</tr>
<tr>
<td>$s:c$</td>
<td>$\sigma^2(ts:c) + n_s \sigma^2(s:c)$</td>
<td>$\frac{\text{MS}(s:c) - \text{MS}(ts:c)}{n_p}$</td>
</tr>
<tr>
<td>$tc$</td>
<td>$\sigma^2(ts:c) + n_s \sigma^2(s:c)$</td>
<td>$\frac{\text{MS}(tc) - \text{MS}(ts:c)}{n_s}$</td>
</tr>
<tr>
<td>$ts:c$</td>
<td>$\sigma^2(ts:c)$</td>
<td>$\text{MS}(ts:c)$</td>
</tr>
</tbody>
</table>

Note. $c =$ classification; $s =$ SME rater; $t =$ KSAO-task linkage. $^a$Variance component. $^b$Expected mean square. $^c$Expected variance component.

Although there was some missing data for particular common linkages because some SMEs indicated they were unable to rate them, urGENOVA was created to handle these missing data. Thus, all linkages were included in the analyses, regardless of whether or not they received ratings from each SME.

Analysis of non-common items. The non-common KSAO-task linkage ratings were also analyzed in a G study to evaluate the success of applying MMS to the rest of the linkages. The facets in these non-common linkage ratings were evaluated, which corresponded to: pair ($p$) and SME rater ($s$). SME classification was not included at this point in the analysis because the results from the analysis of the common linkages revealed that classification contributes almost nothing to the differences in ratings. Thus, there was confidence that the SMEs from various classifications can provide equally valid linkage ratings. The design of the analysis for the non-common tasks was specified
as a \((t \times s):p\) design. This design indicates that for each pair, KSAO-task linkages were rated by a specific number of SMEs. Table 6 presents the equations for deriving the mean squares and variance components for the non-common KSAO-task linkage ratings.

There was missing data for particular non-common linkages because some SMEs indicated they were unable to rate them. Although urGENOVA was created to handle missing data, the design for the non-common linkages resulted in an extremely large amount of missing data. Particularly, if at least one SME in a given pair did not provide ratings for a given linkage, only one data point would be left in the analysis for that linkage, which is the data point of the other SME in the pair. Thus, linkages with any missing ratings from at least one SME were removed from the analysis.

Table 6

*Expected Mean Squares and Random Effects Variance Components for the Non-Common Linkage Ratings*

<table>
<thead>
<tr>
<th>VC(^a)</th>
<th>(E(MS)^b)</th>
<th>(\hat{\sigma}^2c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p)</td>
<td>(\sigma^2(ts:p) + n_t \sigma^2(s:p) + n_s \sigma^2(t:p) + n_t n_s \sigma^2(p))</td>
<td>(\frac{MS(p) - MS(ts:p) - MS(s:p) + MS(ts:p)}{n_t n_s})</td>
</tr>
<tr>
<td>(t:p)</td>
<td>(\sigma^2(ts:p) + n_s \sigma^2(t:p))</td>
<td>(\frac{MS(t:p) - MS(ts:p)}{n_s})</td>
</tr>
<tr>
<td>(s:p)</td>
<td>(\sigma^2(ts:p) + n_t \sigma^2(s:p))</td>
<td>(\frac{MS(s:p) - MS(ts:p)}{n_t})</td>
</tr>
<tr>
<td>(ts:p)</td>
<td>(\sigma^2(ts:p))</td>
<td>(MS(ts:p))</td>
</tr>
</tbody>
</table>

Note. \(p\) = pair; \(s\) = SME rater; \(t\) = KSAO-task linkage. \(^a\)Variance component. \(^b\)Expected mean square. \(^c\)Expected variance component.
Chapter 3

RESULTS

Common KSAO-Task Linkages

*Descriptive Statistics*

Table 7 presents the number of times each linkage rating was chosen for a given common linkage across all SMEs. Some SMEs were unable to rate some linkages due to the fact that they had not performed the task or supervised someone performing the task. Of the 74,723 ratings that were provided by the SMEs as a whole, a rating of “essential” was chosen almost half of the time (45.05%). This finding was expected since the KSAOs chosen to be linked were already deemed important by previous SMEs. However, there was variance in the ratings as whole because the other two ratings, “not relevant” and “helpful,” were chosen a moderate amount of times. This variance confirms that the SMEs in fact used the entire range of response options on the linkage rating scale. Therefore, G theory was appropriate for evaluating the sources that contributed to this variance. If a particular rating was chosen more than half of the time or if a rating was not used at all, suggesting little or no variance in the linkage ratings, it would be difficult to determine which sources contributed the most to the differences in the linkage ratings.
Table 7

Frequency Count of Common Linkage Ratings

<table>
<thead>
<tr>
<th>Rating</th>
<th>Description</th>
<th>( n^a )</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0: Not relevant</td>
<td>This KSAO is not needed to perform the task/activity. Having the KSAO would make no difference in performing this task.</td>
<td>15330</td>
<td>20.52%</td>
</tr>
<tr>
<td>1: Helpful</td>
<td>This KSAO is helpful in performing the task/activity, but is not essential. These tasks could be performed without the KSAO, although it would be more difficult or time-consuming.</td>
<td>25728</td>
<td>34.43%</td>
</tr>
<tr>
<td>2: Essential</td>
<td>Without the KSAO, you would not be able to perform these tasks.</td>
<td>33665</td>
<td>45.05%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>74723</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

*Note.* \(^a\)Number of linkages.

Interrater reliability was evaluated using an intraclass correlation. A high intraclass correlation (.94) was revealed for the common tasks, suggesting the SMEs overall provided linkage ratings in a consistent manner. Table 8 provides the mean common linkage rating and standard deviation given by each SME and each pair, as well as the intraclass correlation for each pair. The mean and standard deviation reported for each SME is based on the total number of linkages for which the SME provided a rating. SMEs sometimes provided less than 2806 common linkage ratings since some of them had not performed a given task. The mean, standard deviation, and reliability coefficient
reported for each pair is based on the total number of linkages for which both SMEs in
the pair provided a rating. Thus, even if a linkage rating was rated by only one SME in a
given pair, the linkage was dropped from the analysis for that pair. Comparing the
equivalence of the common linkage ratings provided by each pair indicated the degree of
confidence with which the non-common linkage ratings can be interpreted.
<table>
<thead>
<tr>
<th>SME</th>
<th>Linakge Ratings</th>
<th>Mean&lt;sup&gt;a&lt;/sup&gt;</th>
<th>SD&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Pair</th>
<th>Common Ratings</th>
<th>Mean&lt;sup&gt;b&lt;/sup&gt;</th>
<th>SD&lt;sup&gt;b&lt;/sup&gt;</th>
<th>r&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2806</td>
<td>1.21</td>
<td>0.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2560</td>
<td>1.23</td>
<td>0.73</td>
<td>1</td>
<td>2560</td>
<td>1.22</td>
<td>.00</td>
<td>.50</td>
</tr>
<tr>
<td>3</td>
<td>2561</td>
<td>0.78</td>
<td>0.69</td>
<td>2</td>
<td>2192</td>
<td>.86</td>
<td>.11</td>
<td>.57</td>
</tr>
<tr>
<td>4</td>
<td>2315</td>
<td>0.92</td>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2562</td>
<td>1.53</td>
<td>0.68</td>
<td>3</td>
<td>2562</td>
<td>1.47</td>
<td>.08</td>
<td>.74</td>
</tr>
<tr>
<td>6</td>
<td>2806</td>
<td>1.38</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2806</td>
<td>0.92</td>
<td>0.86</td>
<td>4</td>
<td>2684</td>
<td>.94</td>
<td>.00</td>
<td>.95</td>
</tr>
<tr>
<td>8</td>
<td>2684</td>
<td>0.95</td>
<td>0.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2799</td>
<td>0.97</td>
<td>0.76</td>
<td>5</td>
<td>2555</td>
<td>.99</td>
<td>.00</td>
<td>.58</td>
</tr>
<tr>
<td>10</td>
<td>2562</td>
<td>1.00</td>
<td>0.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2554</td>
<td>1.50</td>
<td>0.67</td>
<td>6</td>
<td>2554</td>
<td>1.50</td>
<td>.00</td>
<td>.74</td>
</tr>
<tr>
<td>12</td>
<td>2805</td>
<td>1.45</td>
<td>0.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2806</td>
<td>1.54</td>
<td>0.69</td>
<td>7</td>
<td>2805</td>
<td>1.54</td>
<td>.00</td>
<td>.97</td>
</tr>
<tr>
<td>14</td>
<td>2805</td>
<td>1.55</td>
<td>0.68</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2802</td>
<td>1.34</td>
<td>0.61</td>
<td>8</td>
<td>2800</td>
<td>1.39</td>
<td>.08</td>
<td>.56</td>
</tr>
<tr>
<td>16</td>
<td>2804</td>
<td>1.44</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>2806</td>
<td>1.14</td>
<td>0.65</td>
<td>9</td>
<td>2806</td>
<td>1.17</td>
<td>.04</td>
<td>.87</td>
</tr>
<tr>
<td>18</td>
<td>2806</td>
<td>1.19</td>
<td>0.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>2316</td>
<td>.93</td>
<td>0.76</td>
<td>10</td>
<td>2194</td>
<td>.95</td>
<td>.04</td>
<td>.57</td>
</tr>
<tr>
<td>20</td>
<td>2683</td>
<td>.96</td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Pairs 4, 7, and 13 had almost perfect Pearson correlation coefficients. There was some degree of inconsistency among the rest of the pairs, but the intraclass correlation of all SMEs (.94), as well as subsequent urGENOVA analyses, indicates the SMEs provided ratings that were sufficiently reliable.

Table 9 presents the mean common linkage rating and standard deviation given by each of the three correctional job classifications, based on the total number of common linkage ratings provided by all SMEs in each classification. Additionally, the intraclass correlation of each classification is provided in the table. The intraclass correlations presented in the table suggest there was sufficient consistency in the ratings provided by each classification. However, the intraclass correlations of the YCO and YCC classifications were not as high as that of the CO classification. Because the CO
classification had almost twice as many SMEs as the other classifications, the Spearman-
Brown formula (Furr & Bacharach, 2008) was used to calculate the intraclass correlation
for the YCOs and YCCs as if there were as many SMEs in these classifications as there
were COs.

Table 9

Mean Common Linkage Ratings for each Classification

<table>
<thead>
<tr>
<th>Classification</th>
<th>Number of SMEs</th>
<th>Number of Linkages</th>
<th>Mean</th>
<th>SD(^a)</th>
<th>ICC(^b)</th>
<th>R(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO</td>
<td>15</td>
<td>1693</td>
<td>1.23</td>
<td>.30</td>
<td>.90</td>
<td>-</td>
</tr>
<tr>
<td>YCO</td>
<td>6</td>
<td>2409</td>
<td>1.33</td>
<td>.26</td>
<td>.76</td>
<td>.89</td>
</tr>
<tr>
<td>YCC</td>
<td>7</td>
<td>2189</td>
<td>1.26</td>
<td>.22</td>
<td>.83</td>
<td>.91</td>
</tr>
</tbody>
</table>

Note. \(^a\)Standard deviation. \(^b\)Intraclass correlation coefficient. \(^c\)Spearman-Brown

correction: the reliability of ratings for the YCO and YCC classifications as if there were
as many SMEs as there were for the CO classification.

The corrected intraclass correlations for the YCO and YCC classifications
compare very well with the intraclass correlation of the CO classification. The Spearman-
Brown correction allowed for the projection that additional SMEs in the YCO and YCC
classifications would increase the intraclass correlations of these groups. This consensus
among the classifications lent confidence with which the non-common linkage ratings
could be evaluated. Since classification was not a major source of variability in the
linkage ratings, the removal of this factor from the design of the non-common linkages was justified.

G Study Results

The G study design for the common KSAO-task linkages was analyzed by urGENOVA. Table 10 presents the mean squares, estimated variance components, and percentage of variance in the linkage ratings for the various facets of measurement. Each facet, or source, and the interactions between them were evaluated to determine the extent they contributed to the variance in linkage ratings.

Because the estimated variance components are difficult to interpret (Brennan, 2001a) and there is a lack of a standard for determining a large versus a small variance component, calculating the percentage of each variance component over the total amount of variance helps to make this judgment. These percentages allows for a comparison of the relative magnitude of each variance component in relation to all others.
Table 10

_G Study Results for the Common Linkage Ratings_

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>MS&lt;sup&gt;a&lt;/sup&gt;</th>
<th>$\sigma^2$&lt;sup&gt;b&lt;/sup&gt;</th>
<th>PV&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>17.289</td>
<td>0.525</td>
<td>16.21%</td>
</tr>
<tr>
<td>$c$</td>
<td>240.147</td>
<td>-0.002&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.00%</td>
</tr>
<tr>
<td>$s:c$</td>
<td>281.921</td>
<td>0.105</td>
<td>3.24%</td>
</tr>
<tr>
<td>$tc$</td>
<td>3.137</td>
<td>0.074</td>
<td>2.28%</td>
</tr>
<tr>
<td>$ts:c$</td>
<td>2.536</td>
<td>2.537</td>
<td>78.33%</td>
</tr>
</tbody>
</table>

**Note.** $t$ = KSAO-task linkage; $c$ = classification; $s$ = SME rater. <sup>a</sup>Mean square. 
<sup>b</sup>Estimated variance component. <sup>c</sup>Percentage of total variance. <sup>d</sup>Negative estimates of variance components interpreted as 0.

Reflecting the object of measurement for the common linkages design, the variance due to the KSAO-task linkages ($t$) reflected the true variance, or actual differences in the linkages. The variance due to the linkages varied considerably in the degree of relevance between each KSAO and task, accounting for 16.21% of the total variance in ratings. A large amount of variance among the linkage ratings due to the KSAO-task linkages relative to the variance attributable to other facets was expected, especially since different tasks can be linked to different KSAOs, as well as different numbers of KSAOs. Therefore, the KSAO-task linkage facet was reflective of its actual, or true score, variance; the differences in the obtained linkage ratings can be attributed to
their actual differences. Consequently, error refers to the degree to which this true score variance varies across other facets, as well as unidentified sources of error.

The variance due to the classification \((c)\) facet yielded a negative variance component estimate. When a negative estimated variance component is near zero, the approach of Shavelson and Webb (1991) can be followed, which suggests this variance component be set to zero. In the event a negative variance component is much larger in relative magnitude, there may be a misspecification in the design of the analysis. Thus, the classification facet accounted for 0% of the total variance in ratings. Since classification hardly contributed to the variance in the linkage ratings, the average ratings provided by the three classifications were extremely consistent. This suggests the three classifications perceived the linkages in the same manner and were therefore interchangeable during the linkage process. In other words, an SME representing the CO classification can be expected to provide linkage ratings equally valid to an SME representing the YCO or YCC classification. Further, this finding confirms that the pairing of SMEs by classification for rating the non-common linkages did not threaten the generalizability of any linkage ratings across the three classifications.

The variance due to the SME rater \((s:c)\) facet accounted for 3.24% of the total variance in ratings. Human subjectivity is always expected to contribute to some of the error in ratings. However, this facet was not a major source of variance in the linkage ratings.
The variance among the two-way interaction of KSAO-task linkage by classification ($tc$) accounted for 2.28% of the variance. This further supports that classification was not a major source of variance in the linkage ratings.

The last facet listed in the table reveals that the two-way interaction of linkage by SME raters ($ts:c$) and all other unspecified and random sources of error accounted for 78.33% of the total variance. This variance component was relatively large compared to the others because it includes the residual, or unspecified and random sources of error.

*Measures of Reliability*

To assess the level of dependability, or consistency, among the common linkage ratings across all SMEs, the generalizability coefficient and index of dependability were calculated. In order to calculate these coefficients, the absolute and relative error variances of the common linkages were computed. The absolute error of the common linkages was computed using the following formula:

$$\sigma^2(\Delta) = \sigma^2(c)/n_c + \sigma^2(s:c)/n_sn_c + \sigma^2(tc)/n_c + \sigma^2(ts:c)/n_sn_c,$$

which was $\sigma^2(\Delta) = .056$. The relative error of the common linkages was computed using the following formula,

$$\sigma^2(\delta) = \sigma^2(tc)/n_c + \sigma^2(ts:c)/n_sn_c,$$

which was $\sigma^2(\delta) = .055$. For relative decisions when the rank ordering of the linkage ratings is of central concern, the generalizability coefficient was computed with the following formula,
\[ E\rho^2 = \frac{\sigma^2(t)}{\sigma^2(t) + \sigma^2(\delta)}, \]  
which was \( E\rho^2 = .91 \). For absolute decisions when the relevance of each KSAO to each task is of concern, regardless of all other linkages, the index of dependability was computed using the following formula,

\[ \Phi = \frac{\sigma^2(t)}{\sigma^2(t) + \sigma^2(\Delta)}, \]  
which was \( \Phi = .90 \). These reliabilities suggest there was a high degree of consistency among the common linkage ratings. This finding was expected since the variance components obtained from the G study in Table 10 indicated that the classification and SME rater facets contributed relatively little to the variance in the common linkage ratings as opposed to the actual linkages themselves. Although the interaction between the linkages, SME raters, classifications, and all other unspecified sources of error contributed significantly to the variance in the ratings, the ratings were still considered very reliable as indicated by the computed generalizability coefficient and index of dependability.

Non-Common KSAO-Task Linkages

Descriptive Statistics

Table 11 presents the number of times each linkage rating was chosen for a given non-common linkage across all SMEs. Some SMEs were unable to rate some linkages due to the fact that they have not performed the task or supervised someone performing the task. Of the 50,176 ratings that were provided by the SMEs as a whole, a rating of “essential” was chosen almost half of the time (41.64%). Like with the common linkage
ratings, this finding was expected since the KSAOs chosen to be linked were already
deemed important by previous SMEs. However, there was variance in the ratings as
whole because the other two ratings, “not relevant” and “helpful,” were chosen a
moderate amount of times. This variance also confirms that the SMEs in fact used the
entire range of response options on the linkage rating scale. Therefore, G theory was
appropriate for evaluating the sources that contributed to this variance. If a particular
rating was chosen more than half of the time or if a rating was not used at all, suggesting
little or no variance in the linkage ratings, it would be difficult to determine which
sources contributed the most to the differences in the linkage ratings.
Table 11

*Frequency Count of Non-Common Linkage Ratings*

<table>
<thead>
<tr>
<th>Rating</th>
<th>Description</th>
<th>N</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0: Not relevant</td>
<td>This KSAO is not needed to perform the task/activity. Having the KSAO would make no difference in performing this task.</td>
<td>11025</td>
<td>21.97%</td>
</tr>
<tr>
<td>1: Helpful</td>
<td>This KSAO is helpful in performing the task/activity, but is not essential. These tasks could be performed without the KSAO, although it would be more difficult or time-consuming.</td>
<td>18256</td>
<td>36.38%</td>
</tr>
<tr>
<td>2: Essential</td>
<td>Without the KSAO, you would not be able to perform these tasks.</td>
<td>20895</td>
<td>41.64%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>50176</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

In order to investigate the degree of consistency between each pair of SMEs, the mean non-common linkage rating and standard deviation given by each pair, as well as the intraclass correlation, was computed and is provided in Table 12. The mean and standard deviation reported for each pair is based on the total number of linkages for which both SMEs in that pair provided a rating. Although each SME in a given pair was assigned to rate 1,952 to 2,440 linkages (16 to 20 non-common tasks x 122 KSAOs), some SMEs were unable to rate some linkages because they have not performed the task.
or supervised someone performing the task. Thus, SMEs sometimes provided less than
1,952 to 2,440 linkage ratings to their assigned non-common linkages.

The mean, standard deviation, and intraclass correlation reported for each pair is
based on the total number of linkages for which both SMEs in the pair provided a rating.
Thus, even if a linkage rating was rated by only one SME in a given pair, the linkage was
dropped from the analysis for that pair.
Table 12

Mean Non-Common Linkage Ratings for each SME and Pair

<table>
<thead>
<tr>
<th>SME</th>
<th>Mean(^a)</th>
<th>SD(^a)</th>
<th>Pair</th>
<th>Linkage</th>
<th>Mean(^b)</th>
<th>SD(^b)</th>
<th>r(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.22</td>
<td>.64</td>
<td>1</td>
<td>1585</td>
<td>1.24</td>
<td>.03</td>
<td>.49</td>
</tr>
<tr>
<td>2</td>
<td>1.26</td>
<td>.68</td>
<td>2</td>
<td>1464</td>
<td>.87</td>
<td>.10</td>
<td>.54</td>
</tr>
<tr>
<td>3</td>
<td>.80</td>
<td>.70</td>
<td>3</td>
<td>1952</td>
<td>1.34</td>
<td>.09</td>
<td>.73</td>
</tr>
<tr>
<td>4</td>
<td>.94</td>
<td>.83</td>
<td>4</td>
<td>1950</td>
<td>.87</td>
<td>.01</td>
<td>.98</td>
</tr>
<tr>
<td>5</td>
<td>1.44</td>
<td>.71</td>
<td>5</td>
<td>1946</td>
<td>.95</td>
<td>.05</td>
<td>.56</td>
</tr>
<tr>
<td>6</td>
<td>1.31</td>
<td>.76</td>
<td>6</td>
<td>1951</td>
<td>1.44</td>
<td>.00</td>
<td>.84</td>
</tr>
<tr>
<td>7</td>
<td>.86</td>
<td>.86</td>
<td>7</td>
<td>2069</td>
<td>1.54</td>
<td>.01</td>
<td>.96</td>
</tr>
<tr>
<td>8</td>
<td>.88</td>
<td>.88</td>
<td>8</td>
<td>2069</td>
<td>1.54</td>
<td>.01</td>
<td>.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>-----</td>
<td>---</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1.21</td>
<td>.63</td>
<td>8</td>
<td>1708</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1.39</td>
<td>.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>1.10</td>
<td>.62</td>
<td>9</td>
<td>2196</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>1.14</td>
<td>.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>.96</td>
<td>.75</td>
<td>10</td>
<td>1708</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>.95</td>
<td>.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>1.05</td>
<td>.89</td>
<td>11</td>
<td>1934</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>.99</td>
<td>.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>1.15</td>
<td>.70</td>
<td>12</td>
<td>1706</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>.98</td>
<td>.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>1.53</td>
<td>.75</td>
<td>13</td>
<td>1460</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>1.53</td>
<td>.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>1.46</td>
<td>.69</td>
<td>14</td>
<td>1458</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>1.57</td>
<td>.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.** aBased on the total number of linkages for which the SME provided a rating. bBased on the total number of linkages for which both SMEs in the pair provided a rating. cPearson product-moment correlation coefficient.

**G Study Results**

The design for the non-common KSAO-task linkages was also analyzed by urGENOVA. Table 13 presents the mean squares, estimated variance components, and percentage of variance in the linkage ratings for the various facets of measurement and
the interactions between them. Each facet, or source, and the interactions between them were evaluated to determine the extent they contributed to the variance in linkage ratings.

Table 13

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>MS(^a)</th>
<th>(\hat{\sigma^2})</th>
<th>PV(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p)</td>
<td>220.692</td>
<td>0.059</td>
<td>9.82%</td>
</tr>
<tr>
<td>(t:p)</td>
<td>0.940</td>
<td>0.401</td>
<td>66.72%</td>
</tr>
<tr>
<td>(s:p)</td>
<td>7.540</td>
<td>0.004</td>
<td>0.67%</td>
</tr>
<tr>
<td>(ts:p)</td>
<td>0.137</td>
<td>0.137</td>
<td>22.80%</td>
</tr>
</tbody>
</table>

Note. \(p\) = pair; \(t\) = KSAO-task linkage; \(s\) = SME rater. \(^a\)Mean square. \(^b\)Estimated variance component. \(^c\)Percentage of total variance.

The variance due to the pair \((p)\) facet accounted for 9.82% of the total variance in ratings. While it appears that a reasonable amount of variance was attributable to the pairs of SMEs, this percentage of variance is still relatively small compared with the other facets in the design, with the exception of the SME rater facet.

The variance due to the SME rater \((s:p)\) facet accounted for 0.67% of the total variance in ratings. This amount of variance indicated that error in the linkage ratings was hardly due to the SME raters themselves.

Reflecting the object of measurement for the non-common linkages design, the variance due to the KSAO-task linkages \((t:p)\) accounted for 66.72% of the total variance.
in ratings. As with the common linkages, the variance in linkages was expected due to the fact that tasks vary in their relevance to various KSAOs.

The variance among the two-way interaction between the linkages and SME raters \((t_s\cdot p)\), as well as all other unspecified and random sources of error, accounted for 22.80% of the total variance in ratings. The percentage of variance of this source was considerable, yet not as large as the variance due to the linkages themselves. Thus, while other unidentified sources of error may be attributing to the differences in the linkage ratings, the true differences in the linkages contributed to the variance in the ratings.

**Measures of Reliability**

To assess the level of dependability, or consistency, among the non-common linkage ratings across all SMEs, the generalizability coefficient and index of dependability were calculated. In order to calculate these coefficients, the absolute and relative error variances of the non-common linkages were computed. Using Equation 5, the absolute error of the non-common linkages was computed, which was \(\sigma^2(\Delta) = .009\). Using Equation 6, the relative error of the non-common linkages was computed, which was \(\sigma^2(\delta) = .005\). For relative decisions when the rank ordering of the linkage ratings is of central concern, the generalizability coefficient was calculated using Equation 7, which was \(E\rho^2 = .99\). For absolute decisions when the relevance of each KSAO to each task is of concern, regardless of all other linkages, the index of dependability was calculated using Equation 8, which was \(\Phi = .98\). These reliabilities suggest there was an extremely high degree of consistency among the non-common linkage ratings. The
reliabilities were expected given the variance components provided in Table 13 and the reliabilities of .90s for common linkages. The pair facet contributed relatively little to the variance in the linkage ratings. Thus, the pairing of the SMEs was inconsequential. Additionally, the SME raters themselves were again a negligible source of error in the linkage ratings.
Chapter 4

DISCUSSION

In response to the lack of published research to effectively improve the linkage process during job analysis, this study examined the feasibility of applying MMS to address the overwhelming number of KSAO-task linkages that often need to be rated. The success of using MMS in various other contexts (Berk, 1986; Garg, 1987; Gold & Basch, 1984; Gressard & Loyd, 1991; Liang, 2003; Norcini et al., 1988) led to the thinking that MMS has the potential to be as successful in the context of the linkage process. Results from the current study have supported the capacity of MMS in alleviating a significant issue faced in the linkage process. While it would be ideal for each linkage to be rated by each SME, this endeavor is impossible and impractical. To confront the overwhelming number of KSAO-task linkages identified during the linkage process, MMS presents a solution to reduce the work and effort needed from SMEs and the time and financial resources put forth by organizations paying for the job analysis while allowing all the desired number of linkage ratings to be made. The linkages can instead be divided into subsets using MMS, which are then rated by subsets of SMEs.

After an MMS design was created and used to obtain the necessary linkage ratings, G theory was used to evaluate the reliability of these ratings. G theory was appropriate for this analysis as it can disentangle various sources of error, unlike classical test theory, which cannot make this distinction (Brennan, 2001a; Shavelson & Webb, 1991). The G study conducted on the common linkage ratings revealed that the SMEs
were providing ratings in a sufficiently reliable manner. The variance in the linkage ratings primarily reflected the true differences in the linkage ratings, whereas the SMEs themselves and their various classifications attributed relatively little error into the rating process. Although the SMEs were subject to human error (Harvey, 1991; Nathan & Alexander, 1985), they accounted for a relatively small amount of variance in the ratings. The actual classifications of the SMEs may have contributed almost nothing to the variance in the ratings because the classifications were operating as similar entities. In fact, the final results of the job analysis indicated that 48%, or 138, of the 285 tasks that were evaluated have been found to be core tasks for all three classifications (CSA, 2007). This finding supports the hypothesis that there is overlap among the three classifications. Thus, the SMEs representing different classifications were viewed as interchangeable in the linkage process, such that the different classifications did not affect the validity of the obtained linkage ratings. While the classifications are still different in over half of the tasks, enough similarity between them exists that enabled the SMEs in the study to evaluate the linkages in a sufficiently similar manner.

The consensus among the SMEs lent confidence with which the remaining linkages could be rated. These remaining linkages were divided into groups to be rated by the SME pairs. The G study conducted on these non-common linkage ratings also revealed that the true differences in the linkages were the main contributors to the variance in the ratings. The SME raters themselves contributed almost nothing to the variance in the ratings and the actual pairs contributed relatively little. The fact that the
pairs could discuss their ratings with each other may explain why the percentage of variance due to SMEs and the interaction between SMEs and linkages were so much smaller in the non-common linkage ratings than in the common linkage ratings. The discussion between SMEs may have reduced the error variance in the non-common linkage ratings. Although the variance due to the pairs was relatively small, it still accounted for almost 10% of the variance in the ratings. However, it is unclear whether this variance resulted from having only two individuals provide ratings for each non-common linkage or from having each pair represent the same classification.

These findings provide support that MMS can be successfully applied to the linkage process by reducing the number of linkage ratings required by each SME. G theory enabled the reliability of these ratings to be assessed and confirmed that MMS maintained the desired accuracy of the linkage ratings. While the study conducted by Norcini et al. (1988) demonstrated the feasibility of MMS to setting performance standards and the current study to the linkage process, both studies involved human judgments and still obtained acceptable results. Thus, the current study adds to the research literature surrounding the use of MMS beyond test development, setting standards, and program evaluation. It adds to the use of MMS in obtaining human judgments, namely KSAO-task linkage ratings, and serves as a foundation for future research to take place.

The current study also adds to the research literature by revealing that rater type is a minimal contributor of variance in job analysis ratings. Despite the use of various types
of SME raters in both the current study and in that of Baranowski and Anderson (2005), job analysis ratings were sufficiently similar across the individuals. In particular, similar ratings were obtained in the current study among three different correctional officer job classifications and in the study conducted by Baranowski and Anderson (2005) among job incumbents and project and non-project job analysts. Future studies should seek to expand the current study by assessing the ability of non-correctional raters (such as other staff within a correctional organization) to provide reliable job analysis ratings. By determining if individuals who are not familiar with the job can provide ratings as sufficiently as corrections SMEs, this investigation would then be analogous to that of Baranowski and Anderson (2005). Rater type could be incorporated into another MMS design as another facet and a G study could be conducted to determine if this facet adds significant variance in the linkage ratings. A finding that non-correctional raters are able to provide the desire ratings would result in a huge cost savings for organizations, such as reducing the resources needed to obtain SME raters for extended periods of time. It is important for future studies to continue examining the linkage process, particularly by determining factors that contribute significantly to the variance in linkage ratings. While rater type may not be a major contributor of variance to job analysis ratings, obtaining accurate ratings depends largely on identifying other potential sources of error that are in fact influential.
**Future Direction**

The findings obtained from this study provide a starting point from which additional research on the linkage process can be conducted. Because this study was the first to apply MMS to the linkage process, future research has several directions to take. Other potential facets beyond those included in the current study should be examined to further evaluate which sources of error contributed to the variance in the linkage ratings. The error terms in the G studies were somewhat substantial, suggesting that other unspecified and random sources of error are contributing to the variance in the linkage ratings. Unspecified sources of error may include the current area of focus at work, of each SME. For example, SMEs who currently are assigned to adult corrections may view the job differently from SMEs who are currently assigned to juvenile corrections. The particular background experiences of the SMEs may be another random source of error. For instance, the views of SMEs who work in small, rural locations may differ from those of SMEs working in large, urban locations. Future studies should seek to investigate these facets as potential sources of error.

While it seems that simply increasing the number of SMEs to provide ratings will improve the reliability of ratings, doing so requires an increase in the needed resources, costs, and time. Nevertheless, different sample sizes of SMEs should be considered to determine a minimum number of SMEs needed to yield reliable linkage ratings. It may be the case that fewer total SMEs may be needed to successfully complete the linkage
process. In contrast, more SMEs per group may be needed to reduce the variance in the linkage ratings that is due to the actual SME pairs.

It is likely that MMS will fare well when used for different jobs, especially since MMS appears to be a versatile theory when applied in contexts other than job analysis. However, other correctional job classifications should be explored since this study only examined three classifications. While it was feasible to examine three classifications in one job analysis, other classifications may not be as interchangeable as those in the study. Other classifications may not overlap in similar tasks as much as those in the current study. If other classifications are not as similar, but are included in the same job analysis, the obtained linkage ratings may differ substantially due to the different classifications. Future studies should also examine the application of MMS to different types of jobs to determine if the actual job content does not affect MMS designs. The field of corrections may have classifications such as the CO, YCO, and YCC that overlap enough to allow for one job analysis. However, other job types may be too unique that a MMS design must be developed for each of them separately.

**Implications for Organizations**

An efficient and reliable linkage process is necessary, particularly as organizations seek to justify use of their selection tests and training programs to establish standards for entrance into their organizations. Carefully planned steps for conducting the linkage process, completed by properly trained and monitored SMEs, satisfies this need. The current study provides an advantageous basis for organizations from which future
studies can continue to explore the feasibility of MMS. The current study demonstrates how organizations can use MMS to develop selection and training standards without exhausting SMEs and using unnecessary amounts of resources. With additional studies supporting the capacity of MMS to improve the linkage process, organizations can be insured that they can save resources and increase the efficiency of job analysis, without losing the detail and specificity needed during the linkage process.
REFERENCES


