1989

Assessing Soccer Referee Performance Using Work Sample and Conventional Testing Methods

Robert L. Kuhnle
Old Dominion University

Assessing Soccer Referee Performance
Using Work Sample and Conventional
Testing Methods

by

Robert L. Kuhnle
B.S. June 1963, U.S. Coast Guard Academy
M.S. June 1973, U.S. Air Force Institute
M.S. June 1973, U.S. Air Force Institute

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

PSYCHOLOGY

OLD DOMINION UNIVERSITY

Approved by:

____

Terry L. Dickinson (Director)

____

____

____

ABSTRACT

ASSESSING SOCCER REFEREE PERFORMANCE
USING WORK SAMPLE AND CONVENTIONAL TESTING METHODS

Robert L. Kuhnle
Old Dominion University, 1989
Director: Dr. Terry L. Dickinson

A game simulation consisting of game segments filmed from two camera angles, a behavioral event interview (BEI), a written test, and physical performance test battery were compared for testing college soccer referees as linesmen. A content-oriented strategy (Alba & Dickinson, 1985) was used to prepare the tests. Sixty-one referees from two testing sites were assigned to one of two conditions of physical demand and one of three experience groups. Strong evidence of criterion-related validity was found for the game simulation from the press box camera angle when game simulation scores were compared with peer ratings and assessment scores. Mixed results were found for measures of construct validity. Although some encouraging results were found, convergent and discriminant validity were low. Method bias was low except in the high experience group, where method bias was moderate. The effects of testing site and physical demand on game simulation score were not significant. The effects of experience and camera angle were significant. Game simulation scores from the press box camera angle increased with total senior level soccer experience. Comparisons of the tests showed that the testing of linesmen can be best accomplished with a combination of methods that includes the game simulation, the BEI, a written test about fouls and misconduct, and a physical performance test. Results also showed that scoring the game simulation was not influenced by the soccer-related experience of the scorers. Questionnaires were used to assess the acceptability of the game simulation, the BEI, and the written tests. All three were generally acceptable to the participants, but the BEI was significantly more acceptable than the game simulation. The combined evidence from this research suggests that the content-oriented strategy produced valid, reliable, and acceptable tests of linesman performance.

Dedication

To Janet Kuhnle, my wife and best friend for the past 25 years, and our children, Kristi, Scott, Kurt, and Kellianne. Without them, this would mean nothing. Because of their patience, understanding, and encouragement, this dream has become a reality.

To my parents, Audrey and Bob Kuhnle who guided me during my formative years and instilled in me an insatiable thirst for knowledge.

Finally, to soccer, the ultimate sport. This is my effort to repay the game for the years of pleasure that it has brought me.

ACKNOWLEDGEMENTS

This document and the hundreds of hours that it represents would not have been possible without the help of many individuals. A complete list of those who have guided this research, my professional development, and my existence over the past eight years is too long to be included here.

I would like to thank my dissertation committee members, Drs.: Terry Dickinson, Glynn Coates, Robert McIntyre, and Roy Yarbrough for their support throughout the research . To Dr. Coates goes my thanks for his help during the analysis phases of this research, despite his questions that occasionally served as “heart- stoppers.” Your patience with my seemingly endless stream of statistical questions is a skill that I hope to learn. To Dr. McIntyre, thanks for the “pep-talks.” Your boundless energy is contagious and served to restore my level of enthusiasm at times when the end could not be seen. Thanks also for your guidance and eye for detail early in this research. Finally, thanks to Dr. Yarbrough for his sense of humor and for allowing me to be a part of his NISOA physical fitness staff. Without that opportunity, this research would not have been possible.

A special thanks to Dr. Terry Dickinson who chaired my dissertation committee and guided my development throughout my doctoral program. You have been a constant source of expertise, guidance, and support. I will always hold our relationship as a model toward which to strive in my future as a mentor for others. I look forward to continuing our friendship in the years to come.

This research would not have been possible if Dr. Raymond Bernabei, the Executive Director of the National Intercollegiate Soccer Officials Association, and Mr. Bob Sumpter, then the National Director of Referee Instruction for the United States Soccer Federation and NISOA, had not been willing to listen to my proposal and provide their support for this project in 1983. It would certainly not have been possible if Ab Leonard, the NISOA National Camp Director, had not been willing to allow me to disrupt the

schedule of his national camps with my testing. Finally, this research could not have been done without the referees of NISOA who volunteered to be the participants, raters, assessors, and experts in this project. You must remain anonymous, but you will not be forgotten.

Thanks, also, to my colleagues at Thomas Nelson Community College who have provided me with considerable support, especially during the past year. A special thanks to Marie Tyler, the Chairman of the Business Sciences Division and my immediate supervisor, for allowing me the freedom to complete my course work and this research.

Lastly, and most importantly, I acknowledge the sacrifices of my family over the past eight years. You have "done without" what other families take for granted so that I might finish this project. I know that I can never fully repay you for those sacrifices.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF TABLES

LIST OF TABLES

LIST OF FIGURES

ASSESSING SOCCER REFEREE PERFORMANCE USING WORK SAMPLE AND CONVENTIONAL TESTING METHODS

I. INTRODUCTION

> The human animal is an extraordinary species. Of all the events in human history, the one to attract the largest audience was not a great political occasion, nor a special celebration of some complex achievement of the arts or sciences, but a simple ball-game—a soccer match (Morris, 1981, p. 7).

Sports have played a significant role in the history of humanity. From an anthropological perspective, they have served as ritualistic hunts, stylized battles, status displays from the personal to the national level, and semi-religious ceremonies (Morris, 1981). In recent years, sports have served to emphasize (e.g., the 1980 Olympics) and to bridge cultural and philosophical differences.

Today, sports are also big business. Sports personalities often earn thousands, even millions of dollars in a single event. The average salary of players in major professional sports in the United States exceeds $250,000. Gate revenue at a single football game can approach one million dollars.

Media coverage of sporting events has increased greatly in recent years. Today, millions of spectators may view one event. As the amount of coverage has increased, the focus of that coverage has shifted. No longer is media attention limited to the event itself. The focus has shifted to other aspects of the lives of those involved in playing, coaching, or officiating the game, including patterns of past performance and their strengths and weaknesses. It is virtually impossible for players, coaches, and game officials (e.g., an umpire in baseball; the referee in football; or the referee or linesman in soccer) to avoid the watchful eye of the media and viewing audiences.

In some instances, a single photograph or one frame of a videotape captures the difference between winning and losing. Often, the split second which has been frozen forever shows the brilliance of one combatant or the error of another. More and more, these critical incidents involve game officials and their decisions.

With the outcome potentially riding on each decision rendered by a game official, there is both a need and an opportunity for industrial/organizational (I/O) psychologists to contribute their expertise about selection testing and performance assessment in work and training settings. Work by I/O psychologists in these areas is considerable; but, application of existing theory or research to sports has been focused mainly on players and coaches. Few applications have involved game officials.

The Game of Soccer

Soccer has enjoyed worldwide popularity for the past one hundred years. World Cup matches often draw over 100,000 spectators and over one billion viewers around the world. Competition in the World Cup has risen from less than 30 countries to over 150 countries in less than 25 years. Soccer is among the fastest growing youth sports in this country; however, our players, teams and officials have received little international recognition. I have attended national level college soccer referee clinics in each of the past four years where college coaches from around the country have been invited to address the participants. The coaches have consistently pointed out that the development of soccer officials has not kept pace with development in other areas of the game. These observations suggest that the systems for selecting (i.e., certifying) and developing officials in the United States are both ineffective and rudimentary.

Soccer Official Selection Systems

Three organizations govern soccer in the United States. The National Federation of State High School Associations governs high school soccer. About 10,000 members of the National Federation Interscholastic Officials Association (NFIOA) officiate high school soccer. The National Collegiate Athletic Association governs college soccer for men, providing referees through the National Intercollegiate Soccer Officials Association (NISOA) and its 2,000 referee members. College soccer for women, most amateur soccer, and professional soccer are governed by the United States Soccer Federation (USSF). Approximately 35,000 individuals are registered as USSF referees.

The three organizations have very different goals and function independently especially where referee certification and upgrade are concerned. Referee certification is based on pencil-and-paper tests alone or in combination with physical performance tests. Although rules of play are nearly identical for the organizations, their written tests vary in style, content, procedures for administration, and standards for passing.

Systems for classifying referees into grades exist, but vary considerably in terms of procedures and criteria, (e.g., experience in years, number of games officiated). Upgrading involves some combination of written tests, physical performance tests, and field assessments in actual games. Certification and upgrading processes do not differentiate between the qualifications for referees and linesmen, which are two very different jobs. Written tests are not job specific. Validity and test reliability data have not been reported. Work sample tests are not used even though they have been used in other occupational areas.

Two organizations require members to complete physical performance tests. The USSF requires three physical tests (i.e., an endurance run, a sprint, and a shuttle run) and NISOA requires four (i.e., an aerobic endurance run, a sprint, an agility run, and an anaerobic endurance run). When the present research began in 1985, only the aerobic run and the sprint were used by the USSF and NISOA. In 1989, the USSF added the agility run suggested by Kuhnle and Yarbrough (1986) to the test battery for their National Referees. Except for the agility run, published standards are different in the two organizations, even for the tests that are similar. Test reliability and validity data have not been reported.

The generally accepted criterion measure of soccer official performance is the field assessment, where experts observe and evaluate the officiating of an actual game. Weighted performance scores from up to 10 performance dimensions are combined to produce a single composite assessment score.

There are problems when field assessments are used as criterion measures. First, games and game events are not standardized. A wide range of situational variables accompany each field assessment. Easy games may contain situations (e.g., penalty kicks, violent conduct) requiring critical decisions. An inconsequential mid-season game may be excellent for assessment, while a potentially critical tournament game may be without challenge and unsuitable for assessment. The number and interaction of situational variables make comparisons of field assessment scores difficult, if not meaningless.

Problems also exist with the published performance dimensions. The relatedness of the dimensions to performance has not been shown. In addition, only dimension titles with five-point rating scales and adjectival anchors—excellent, very good, good, fair, and poor—are used. Neither behavioral descriptions of the dimensions nor behavioral anchors for the rating scales are provided. Behavioral examples are essential to provide unambiguous interpretations of dimensions. By themselves, dimension titles are insufficient. As anecdotal evidence, I can report that, at workshops for assessors and at training seminars for officials between 1985 and 1988, participants and instructors informally voiced a wide range of opinions about what behaviors constitute exceptional, average, and unacceptable performance for each performance dimension.

Finally, the reliability of individual dimensions and overall ratings is highly suspect. In a workshop I attended in 1985, 25 experienced assessors were asked to view game films and complete a referee assessment form to rate the performance of a senior American referee. Scores ranged from 61 to 88 ($\underline{M}$ = 73.8, $\underline{SD}$ = 6.67). Of the 25 scores, 7 were below 70 and 6 were above 80. Similar results were obtained when assessors rated the performance of a senior level referee in an actual game. Top-level assessors considered this variability of scores to be excessive.

In summary, the selection and development system for soccer officials suffers from several classic problems. First, acceptable criterion measures of performance do not exist. Second, evidence of selection measure validity has not been shown. Third, the reliability of field assessment ratings is questionable. The present research investigated a method to overcome these problems.

Sports Literature

Articles from sports psychology (e.g., Hanin, 1977; Morgan, 1980) and recent texts (e.g., Cox, 1985) were reviewed to guide the current research and to identify successful and unsuccessful applications of testing and selection techniques in sports-related situations. Several articles involved testing of athletes; however, few instances of research involving referees in any sport were found. Schurr and Phillips (1971) investigated the frequency and characteristics of successful women sports officials. Alker, Straub, and Leary (1973) and Fratzke (1975) reported mixed results for predicting the success of basketball referees from biographical data and pencil-and-paper tests.

Kuhnle and Yarbrough (1986) sought to identify the physical requirements of soccer officiating. Task analysis results were combined with the results of film and on-site game analyses to determine the scope of the physical requirements of soccer officiating. Seventeen physical performance tests were administered to 80 college soccer officials. Four dimensions were identified as critical to the physical performance of soccer officials (i.e., speed, agility, aerobic endurance, and anaerobic endurance). Four tests, one per dimension, were recommended, but evidence was not obtained for their validity. Of importance to the present research was their finding of no significant decrement in physical performance with age among college soccer referees. The finding is in contrast with findings about aerobic endurance (Robinson, 1938; Astrand & Christensen, 1946; and Londeree & Moeschberger, 1982) and muscle strength (Astrand & Rodahl, 1986; and Grimby & Saltin, 1983) for samples from the general population.

Two studies of soccer official performance were done in foreign countries. Brodie (1981) focused on the physical activities of referees in England. He provided a method for charting referee movement. From an analysis of the movements of top level referees in 10 games, he drew conclusions about the physical demands of refereeing.

Vikhrov (1978) did multi-stage research of soccer officials in the Soviet Union. Only the summary of this unpublished research has been translated into English. In the summary, Vikhrov describes his focus on referee mental activity, specifically the accuracy of decisions of referees and linesmen. He provided (a) an alternative method for assessing referee game performance, (b) a means for determining the most important characteristics of a top level soccer officials, and (c) a set of tests for assessing those characteristics.

He judged how closely the actions of the official corresponded with (a) the rules of the game, (b) the accepted referee mechanics, and (c) the procedural instructions provided by the referee before the game (i.e., pre-game instructions). He classified referee errors (deviations) as consequential, substantial, or minor and developed a ratio of total consequential and substantial errors to the number of games officiated in a season as his criterion measure of referee performance.

Selection Research

Developing a valid selection system involves three fundamental steps. First, job requirements must be identified. Next, methods to assess and predict job performance must be developed. Finally, the psychometric properties of the measures must be evaluated.

Job requirements. For the present research, job requirements were specified in terms of critical tasks to be accomplished and critical knowledges and skills required of referees in performing those tasks. Both types of elements were important to the present research, because both ends and means are important to referee performance.

Levine (1983) suggested the use of a task importance value, where importance is a function of the time spent (T) doing that task, the perceived difficulty (D) of the task, and the criticality (C) of the task. Difficulty and criticality are assessed by incumbents using Likert-type scales. Time spent is expressed as the percent of the available time spent performing that task. A task importance value (TIV) is computed using the following formula:

$$\underline{TIV} = \underline{T} + (\underline{D} \times \underline{C}) \tag{1}$$

Although empirical evidence of the reliability and validity of task importance values was not reported by Levine (1983), the formula is logical. Task importance increases as time, difficulty, or criticality increase.

An alternative approach for assessing job element importance uses the content-validity ratio (*CVR*) (Lawshe, 1975). A job is described in terms of its elements, and these elements are rated by job experts in terms of importance. Then, *CVR* values, ranging from +1 to -1, are computed for each job element by the following formula:

$$\underline{CVR} = \frac{\underline{N(i)} - \underline{N(u)}}{\underline{N(i)} + \underline{N(u)}} \tag{2}$$

In this formula, *N(i)* is the number of experts who rated the element as important, and *N(u)* is the number of experts who rated the element as unimportant.

Ford and Wroten (1984) used the *CVR* technique to describe the importance of job tasks in a police officer training course. Ratings from a 5-point rating scale were trichotomized (i.e., important, unimportant, and neutral) to produce *CVR* values ranging from -.85 to +.93. High interrater agreement was found for three groups of raters (i.e., city police, sergeant, and outside city police).

No research has compared the two techniques, but the *CVR* method appears to be faster and easier to use. Respondents answer only one question for each task as opposed to three questions in Levine's *TIV* method.

Schmitt and Ostroff (1986) suggest that knowledges, skills, and abilities (KSAs) are also important, especially when establishing criteria for job entry. They suggested that a job be rated in terms of (a) the necessity of the KSAs if newly hired employees are to be successful, (b) the risk involved if KSAs are overlooked at selection, and (c) the extent that the presence or absence of KSAs distinguishes between superior and average performers. Approaches to measuring the importance of KSAs in the selection process have not been reported.

Assessing and Predicting Performance

Effective performance assessment systems involve (a) a variety of methods, (b) observation, and (c) integration of information (Cronbach, 1960, p. 582). While this research focused on performance tests and particularly on a work sample as a criterion measure of job performance, data from other sources (i.e., biodata, peer ratings, expert assessments of game performance) were used to provide evidence of validity.

Biodata. Biodata has predicted job performance quite well. Average validity coefficients of .32 to .46 between biodata and various job-related criteria were reported by Reilly and Chao (1982). The average correlations of biodata with ratings (r = .36) and objective measures of job performance (r = .46) were important to the present research, because biodata and objective measures were used to evaluate test scores.

Peer ratings. Lewin and Zwany (1976) and Kane and Lawler (1978) reviewed peer rating research with positive findings, reporting that the reliability, validity, and freedom from bias of peer ratings were acceptable for a variety of applications (Reilly & Chao, 1982). Peer evaluations are more likely to differentiate effort from performance, focus on relevant characteristics (Klimoski & London, 1974; Zammuto, London, & Rowland, 1982), and maintain stability over time (Williams & Leavitt, 1947; Kane & Lawler, 1978). Further, Wherry and Fryer (1949) reported that peer ratings do not tend to be popularity contests, and McEvoy and Buller (1987) noted that peer ratings tend to be more readily accepted by those rated than other ratings (e.g., supervisor ratings).

Kavanaugh, Borman, Hedge, and Gould (1986) suggested that raters from different organizational vantage points are needed to provide sufficient information to evaluate an individual's performance. Zammuto et al. (1982, p. 645) suggested that "each rater occupies a different vantage point vis-a-vis the ratee." Fiske and Cox (1960) concluded that these differences in rater frame of reference contribute most to the differences in ratings that an individual receives. Thus, consistent evaluations from a peer viewpoint should be viewed as useful data.

Expert judgments. Formalized evaluations under actual work conditions by job experts (e.g., supervisors, senior soccer officials) provide another measure of performance. A major problem with such judgments (e.g., game assessments) is the situational specificity of validity when work conditions are not standardized (Schmitt & Hunter, 1984). When work conditions and the conditions under which judgments are gathered are standardized, job expert judgments have been highly accepted as accurate indicators of performance.

Work sample tests. Asher and Sciarrino (1974) reported that work-sample tests, where individuals perform one or more tasks normally required of job incumbents, are among the best for predicting job proficiency. Successful test completion is evidence of the participant's demonstrated ability to do the job. A properly validated work-sample test can also be used as a criterion variable against which to compare other predictors of job performance.

A related testing method is the situational or behavioral-event interview (Latham, Saari, Pursell, & Campion, 1980; Reilly & Chao, 1982; Alba & Dickinson, 1985; Schmitt & Ostroff, 1986). This method requires the construction of a structured interview from job analysis data. The interview questions describe typical job situations. Respondents are rated based on the actions they describe for those situations. Latham et al. (1980) reported interrater reliabilities of .76 and .79 for such interviews. Reilly and Chao (1982) reported an average predictive validity of .33 for 56 interviews.

The behavioral event interview (BEI) is particularly useful for replicating situations that are difficult to simulate or duplicate in work-sample tests (e.g., situations involving emergencies such as lightning, serious injury, or power failure during a night game). The BEI is part of the screening process for soccer officials in England. In a one-hour structured interview, referees are presented with oral descriptions of game situations and asked to describe what actions, if any, would be taken. Responses are recorded or matched with a checklist of possible responses. Interviewers use a branching network to ask follow-up questions.

Walk-Through-Performance-Test. A Walk-Through-Performance-Test (WTPT) is a standardized, task specific technique developed by the U.S. Air Force (Alba & Wilcox, 1985; Alba & Dickinson, 1985). The WTPT includes two components: a work-sample test and a BEI. It provides strong content-related evidence of validity because both the work-sample and BEI are selected from the job domain. A properly constructed WTPT has the psychometric qualities necessary to be used for selection purposes in specific settings (Alba & Wilcox, 1985). No research was found that showed the acceptability of the WTPT to the testing of sports officials. The demonstration of its applicability to the testing of senior level soccer referees was the main goal of the present research.

In the WTPT, the tasks, the test format, and the weighting of possible responses are identified with input from a panel of job experts to ensure standardization of procedures for administering and scoring tests (Standards for Educational and Psychological Testing, 1985). The testing environment is made to approximate the work environment as closely as possible.

The WTPT also provides a method for comparing performance in the work sample and the BEI. At least one-fourth of the work sample and the BEI must be judged as equivalent by a panel of experts. High correlation of recorded performance ratings on these components is evidence of high validity of the components and the overall WTPT.

To be useful in the assessment of soccer official performance, a work sample must overcome the problems associated with the dynamic nature of an actual game. Soccer involves 22 players in constant motion. Rarely do game situations repeat, even under the most controlled circumstances. Orchestration of standardized situations with actual players during testing would be impractical. Instead, videotapes could be used. Segments of the desired game situations could be extracted from game films. These films are readily available; however, the perspective provided by these films contrasts sharply with the field level perspective seen by referees during a game. Most videotapes of games are made from the press box area located well above field level. No research was found that compared the effects of camera angle on response reliability or accuracy of assessing soccer official performance. The present research was designed to permit a preliminary assessment of the effects of camera angle on the accuracy of referee decisions.

Also of interest in this research was whether it was necessary to assess referee performance under near game physical demands or whether "classroom" testing was sufficient. Schmitt and his colleagues (e.g., Schmitt & Ostroff, 1986; Schmitt, Gooding, Noe and Kirsh, 1984) expressed the need for test conditions to replicate actual work conditions. When test conditions accurately reflect job conditions, test scores should be correlated with job performance. No research was found that reported the effect of physical demand on work-sample performance.

Psychometric Properties

Psychometric soundness must be shown to establish the usefulness of criterion measures. Criterion measures must meet standards of relatedness, reliability, and acceptability (Cascio, 1982).

Relatedness. Relatedness or validity "... refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (p. 9) in the 1985 Standards. Evidence of validity is demonstrated by means of content-related, criterion-related, or construct-related strategies. Cronbach (1960) and Ronan and Prien (1966) argued strongly for multiple evidence of validity to show relatedness.

Criterion-related evidence of validity is provided when test scores are systematically related to one or more outcome criteria (Standards, 1985, p. 10). From their meta-analysis comparing criteria for use in test validation, Nathan and Alexander (1988) reported that ratings and work-sample performance were "highly predictable" criteria regardless of the test used in the validation research. They noted that the lack of evidence to support the claim that objective measures were more predictable than subjective evaluations. Their research suggested that work samples are the best criterion for use in assessing tests and that other measures, including subjective ratings, could be used effectively as criteria by which to establish criterion-related evidence of validity for all types of tests.

Although empirical strategies are often used to provide evidence of validity for criteria, appropriate and reliable measures are not always available. In instances where empirical evidence cannot be demonstrated, content-related strategies have been used (Alba & Dickinson, 1985; Schmitt & Ostroff, 1986; Sackett, 1987). Content-relatedness is demonstrated when the sample of tasks or test questions is representative and clearly defined for a domain of job performance. Expert judgments should play an integral part in developing the definition of relatedness (Standards, 1985, p. 10-11). Further, content-related strategies are most effective when the job behavior of interest is observable (Lawshe, 1985).

Content-relatedness is demonstrated by meeting six conditions (Guion, 1978). First, dimensions to be measured must be defined behaviorally. Second, the definitions must be free of ambiguity. Third, the dimensions must be relevant to the job domain in that they reflect the most common and most important aspects of the job (Schmitt & Ostroff, 1986). Fourth, qualified judges must agree that the job domain has been adequately sampled. Fifth, observations of behavior must be measurable and reliable. Sixth, score variance must be attributed to exercises (methods) and not to contaminating and other situational factors.

Construct-related strategies can be used to provide evidence of validity when measures of performance are available for multiple job traits using multiple methods (Campbell & Fiske, 1959; Kavanaugh, MacKinney, & Wolins, 1971). Multitrait-multimethod results can be analyzed by means of analysis of variance techniques

(Boruch, Larkin, Wolins, & MacKinney, 1970). Table 1 shows the interpretation of the sources of variation from a complete analysis of variance of measures (Dickinson, 1987). The first three terms represent fixed effects of little interest in research studies. These effects are eliminated from consideration through analysis of the measures after transforming them to z-scores (i.e., *M* = 0.0, *SD* = 1.0). The remaining terms reflect on construct validity.

Convergent validity of measures is demonstrated when there is sufficient variation in measures to permit ordering of participants. The Participants (P) source of variation reflects the convergent validity of measures. Discriminant validity is demonstrated when job dimensions produce unique ordering of participants. The Participants x Traits (P x T) source of variation is used to show evidence of discriminant validity of measures. Method bias is demonstrated when participants are ordered differently because of the method. This undesirable characteristic is indicated by the Participants x Method (P x M) source of variance. Finally, Error variance is evidence that ratee differences are not accounted for by methods or traits.

Assessing the significance of these effects with F-ratios alone can lead to interpretations that have little practical significance when the number of degrees of freedom is large (Dickinson, 1987). The use of weighted sums of mean squares (i.e., variance components) is a more appropriate strategy for comparing the relative sizes of effects (Vaughn & Corballis, 1969). However, standardized variance components in the form of intraclass correlation coefficients show the percent of variance accounted for and permit comparison of results across research studies.

Reliability. Shrout and Fleiss (1979) noted that measurement error is an unavoidable part of judgments about human performance. Because measurement error can seriously impact statistical analyses and the interpretations of those analyses, assessment of that error in the form of an appropriate reliability index is important.

Measurement error can be introduced during data collection. Consistency of

Table 1

Psychometric Interpretation of Analysis of Variance Summary Table Entries for Multitrait-Multimethod Designs

Source	Psychometric Interpretation
Traits (T)	Trait Bias
Methods (M)	Scale Bias
T x M	Trait x Scale Bias
Participants (P)	Convergent Validity
P x T	Discriminant Validity
P x M	Method Bias
Error	Unexplained Variance

observation and recording of data is a critical part of efforts to reduce measurement error. All performance data for this research were gathered by a single individual. Therefore, reliability for data collection was not assessed.

Measurement error can also occur when performance is assessed by raters or scorers. The sources of variation associated with such assessments is shown in Table 2. From the mean squares of these sources, intraclass correlation coefficients can be computed to reflect indices of reliability. The appropriate index of reliability depends on the method used to gather participant scores (Shrout & Fleiss, 1979). When all participants are rated by all judges, the reliability in ordering participants is given by the following intraclass correlation coefficient (ICC) formula:

$$\mathrm{ICC} = \frac{\mathrm{BMS} - \mathrm{EMS}}{\mathrm{BMS} + (k-1)\ \mathrm{EMS}} \qquad (3)$$

Although no measure of interrater reliability exists when multiple judges rate a unique group of participants, a measure of differences between judges can be obtained.

Table 2

Analysis of Variance Summary Table Used To Test For Consistency of Scores Among Judges of Performance

Source	df	MS
Between Participants	n-1	BMS
Within Participants	n(k-1)	WMS
Between Judges	k-1	JMS
Residual	(n-1)(k-1)	EMS

Note: Abbreviations: BMS, Between-participant mean squares; WMS, Within- participant mean squares; JMS, Between-judges mean squares; and EMS, Error mean squares.

In those applications where judges are expected to produce consistent ratings (i.e., patterns of ratings with equal means) and the process of assigning judges and ratees to rating situations is random, no differences in mean ratings should be found.

Acceptability. Central to the success of a selection measure is the acceptability of the measure to participants and management. Dipboye and Pontbraind (1981) suggested that the opinions of a performance measurement system may be as important to the long-term effectiveness of the system as validity and reliability.

Questionnaires have been administered to measure satisfaction with the performance assessment or appraisal process. For example, Dipboye and Pontbraind (1981) found employee opinions tend to be positive when the employees perceived that they were evaluated on relevant job factors. Landy, Barnes, and Murphy (1978) found that employee reactions to performance appraisal tended to be favorable when (a) employees participated in a feedback session, (b) goals and plans were discussed before the appraisal, and (c) appraisals were made for relevant aspects of the job (i.e., when the measurement device was perceived as job-related).

Research Hypotheses

This research had four main goals. First, it was undertaken to assess a game simulation of linesman performance in college soccer. Second, it was undertaken to assess the impact of physical demands, experience, and camera angle on game simulation scores. Third, it was undertaken to assess the suitability and acceptability of WTPT components for use in the selection of soccer referees. Finally, it was undertaken to assess whether a content-related strategy could be used to develop tests of linesman performance. In keeping with these goals, the following hypotheses were made:

Goal 1: Relatedness of Work Sample to Other Measures of Job Performance.

Hypothesis 1: Scores from the work-sample test will be highly correlated with game assessment scores by experts and subjective ratings of performance by peers. High evidence of rater consistency will be shown for assessment scores and peer ratings.

Hypothesis 2: Evidence of high convergent validity of measures from the Walk-Through-Performance-Test will be demonstrated. The intraclass correlation coefficient for the Participants (P) main effect in the multitrait-multimethod analysis of variance will be large.

Hypothesis 3: Evidence of high discriminant validity of measures from the Walk-Through-Performance-Test will be demonstrated. The intraclass correlation coefficient for the Participant x Trait interaction in the multitrait-multimethod analysis of variance will be large.

Hypothesis 4: Participant ordering will not be due to method. The intraclass correlation coefficient for the Participant x Method interaction in the multitrait-multimethod analysis of variance will be small.

Goal 2: Game Simulation Scores

Hypothesis 5: Significant differences in game simulation scores will be observed for participants placed under high physical demands during the game simulation compared to those placed under low physical demand.

Hypothesis 6: Game simulation scores for top performers will be significantly greater than for average performers.

Hypothesis 7: The angle from which the game is viewed will significantly impact game simulation scores. Videotapes viewed from field level and press box level will produce different orders of participants' scores.

Goal 3: Test Appropriateness and Acceptability

Hypothesis 8: Suitable combinations of conventional tests (i.e., written tests and physical performance tests) will be found for testing of linesmen.

Hypothesis 9: A scoring scheme will be developed to produce reliable scores for the WTPT components.

Hypothesis 10: Post questionnaire results will show that (a) oral instructions were realistic, (b) the game simulation and the BEI will be viewed as being better in terms of their "realism" and the extent that they evaluate linesman ability than a knowledge test or a physical performance test, and (c) the game simulation will be judged as simulating game conditions, including the emotion and pressure normally experienced and the typical flow of events in a college game. The quality of the videotape used for the WTPT will be judged as not significantly interfering with or detracting from referee performance.

Goal 4: Content-related Strategy For Test Construction

Hypothesis 11: Evidence will show that a content-related strategy can be used to develop tests for use in testing soccer linesmen.

II. METHOD

Participants and Setting

Sixty-one soccer referees served as volunteer participants in this research without compensation. Part of the participant group (n = 30) was selected from attendees (n = 42) and staff members (n = 4) at a referee clinic that extended over five days at a central location (i.e., Site 1). Site 1 participants were registered NISOA referees from many states across the country. The remaining participants (n = 31) were registered NFIOA, NISOA, or USSF referees from a single state (i.e., Site 2). They were contacted individually and completed tests by appointment.

Of the 61 participants, 8% (n = 5) were women and 92% (n = 56) were men. Ages ranged from 22 to 54 years (M = 39.86, SD = 7.89). The median age was 41. Except for one Black and one Oriental, participants were Caucasian (96.8%).

Most participants (n = 55) were registered with NISOA. The remaining six referees were registered with the USSF or NFIOA and were either past college referees or potential candidates for entry into NISOA in 1989.

Total soccer experience (i.e., the sum of playing, coaching, and refereeing experience) ranged from 8 to 87 years (M = 36.60, SD = 18.40). Playing experience ranged from 0 to 36 years (M = 8.32, SD = 9.05). Coaching experience ranged from 0 to 29 years (M = 4.53, SD = 5.69). Refereeing experience ranged from 1 to 48 years (M = 23.86, SD = 11.59). College refereeing experience ranged from 0 to 11 years (M = 3.43, SD = 2.65). The median college referee experience was 3 years.

Recruitment of job experts. Instructors, assessors, and senior officials of NISOA were canvassed by mail for 30 volunteers to serve as job experts. Letters described the research and explained that volunteers would complete two mail surveys about the importance of elements of the job of linesman at the college level. They were told that NISOA management supported the research, but that participation was not mandatory. In addition, three senior referees from the local area were recruited to serve on a "panel of experts" to meet during the project.

Recruitment of scorers. Scorers were recruited from a soccer referee association and a community college. By design, individuals with different soccer experience were selected to assess the need for scorers to have soccer experience. An experienced NISOA referee, a young player with limited referee experience, and an office clerk with no soccer experience were used.

The volunteers were presented a description of the project and explained their role in scoring the performance of 10 soccer officials. The volunteers were told after a one-hour orientation and training session that they would score two tests (i.e., a game simulation and a BEI). Scorers were told that NISOA management supported the project, but that scorer participation was voluntary and compensation would not be provided.

Recruitment of participants. At Site 1, volunteers were solicited as individuals arrived at the camp. Using a prepared script, I described the research and the tests that would be involved. Camp attendees were told that NISOA management supported the research, but that participation was not mandatory. They were asked not to discuss their decision about volunteering for testing with other camp attendees to help ensure the randomness of the participant sample. Thirty-eight of the 42 camp attendees and 1 staff member volunteered to participate. Because of time constraints, only 30 completed the tests at Site 1. Two volunteers completed the tests at Site 2.

Site 2 participants were recruited by telephone from a list of 50 NFIOA, NISOA, or USSF referees arranged in alphabetical order. Potential participants were informed of the purpose of the research and provided an outline of the procedures to be followed. Each individual was told that NISOA management supported the project, but that participation was not mandatory. Of the 50 referees, 38 volunteered. Of these, 31 were selected based on their availability when testing was done. Three other volunteers were selected to pilot test each component. Participants from both sites underwent medical examinations and completed a release of liability and an informed consent form as part of the research.

Research Design

The original design for this research was a 2 x 2 factorial design with two conditions of physical demand (i.e., low, high) and two levels of past referee performance (i.e., average, high). The "average" sample of referees was to be drawn from the referees now registered with NISOA excluding those selected for the National Referee Program. The "high" performance group was to be a sample of the NISOA National Referees. A program to identify National Referees within NISOA began in 1985 and was expected to have been completed by 1988. Circumstances resulted in a two-year delay in the program. This resulted in modification in the measurement of past referee performance.

Measurement of experience. As a substitute measure of past referee performance, it was hypothesized that, other things being equal, referees with more experience would perform at higher levels. Five measures of experience were identified. The first two were measures of *referee* experience: total years of referee experience and years of college level referee experience. The panel of experts formed for this research felt that college referee experience would be the most relevant measure of experience to use in the grouping of college referees, but they noted that differences in other refereeing experience, especially at the senior level, could reduce the meaningfulness of the college referee experience measure.

The senior level referee experience of participants ranged from 0 to 33 years ($\underline{M}$ = 9.64, $\underline{SD}$ = 7.42). College referee experience accounted for 0 to 11 years of that experience ($\underline{M}$ = 3.43, $\underline{SD}$ = 2.65). The correlation between senior level referee experience outside the college system and experience in the college system was .55 ($\underline{t}$ = 5.20, $\underline{p}$ < .01) for the total sample; however, for those with less than three years of college referee experience, the correlation was .22 ($\underline{t}$ = 1.12, $\underline{p}$ > .05). Thus, the appropriateness of college referee experience as a measure of past performance was questionable.

Senior amateur soccer is often played by present and past college players and is very similar in style of play to college soccer. The expert panel considered the combined years of college, senior amateur, and professional refereeing as a more appropriate measure of experience. In discussing experience, the panel suggested that experience as a player or coach impacts game performance as a referee.

Although playing or coaching experience may not be a substitute for refereeing experience, these experiences offer at least three potential benefits to referees. First, coaching or playing experience could significantly decrease learning time for the rules and referee duties. Second, playing or coaching experience could train an individual to attend to cues relevant to refereeing that a non-player (or coach) may not see. Finally, playing or coaching requires game-related decisions, the type and speed of which could be similar to those required of referees. This final point was particularly important because, if senior level soccer requires decisions of a type and at a speed different from youth soccer, then it was important to identify experience at the senior level and not just total soccer experience.

Three additional experience variables were created. One variable was a measure of total soccer experience in years without regard to the level of that experience. The other two were measures of senior level experience: senior level referee experience and total soccer experience at the senior level. Both total soccer experience measures were a combination of playing, coaching, and refereeing experience. No research was found to suggest the appropriate weights for the three experience sources. Therefore, equal weighting was used. Tables 3 through 5 illustrate the appropriateness of experience measures.

Assessment scores. For 33 participants from Site 1, expert assessors provided assessment scores from observations of performance as the referee or linesman in one or more games as a part of the camp. Assessment scores were to be used to order participants at the end of camp to identify the outstanding participants. The significance of the assessments was explained at the beginning of camp.

Four National Assessors who had just completed a two-day assessor training course did the assessments. The assessors were selected because of their soccer experience and past performance as assessors. Because there were over 40 camp participants, each assessor was assigned at least 10 games at a single field. One assessor was assigned to each field. Participants were assigned to games randomly as camp schedules permitted. Where possible, participants were assigned to a field and rotated among the linesman and referee position during the game so that each served as the referee for the same amount of time. All games involved youth players. Games were very different in terms of player age, player and team skills, game conditions (e.g., weather), and game intensity.

Assessors viewed each assigned game and made written notes of participant performance. Immediately after each game, participants were provided with oral performance feedback. At the end of the two days of games, assessors met to rank participants. The assessors discussed the performance of each participant, pooling their ratings where referees had been observed on more than one field. A single assessment score was assigned to each participant on an assessment worksheet. Worksheets were placed in numerical order. Adjacent worksheets were compared to account for the level of games worked, the number of observations, and differences in assessor observations (i.e., it was possible for a referee to score very high on an easy game while a "better" referee may have scored lower because the game or the game conditions were more difficult).

Table 3 shows the correlation of assessment scores with the five experience measures. Coefficients varied from .03 to .36. The correlations of ratings with total senior level soccer experience ($r = .36$) and college referee experience ($r = .36$) were significant ($p < .05$), suggesting that these meaures of experience could be used as substitute for past referee performance in the research design.

Inspection of Table 24 of Appendix A shows a significant between-assessor effect ($p < .01$) that was the result of two significant Tukey (HSD) pairwise comparisons. Inspection of assessor mean scores shows that the mean assessment from Assessor 2 was low ($M = 65.00$, $SD = 12.10$) while the mean score by Assessor 4 was high ($M = 83.00$, $SD = 6.80$). Anecdotal discussions with assessors and camp administrators revealed that games involving the older players (i.e., the difficult games) were played at the field where Assessor 2 was assigned. Games involving the youngest players (i.e., the easiest

Table 3

Intercorrelations Between Assessment Scores and Measures of Experience

	ASMNT	ET	ETOT	ER	EYR	EYRC	EYC	EYP
ASMNT	1.00							
ET	.36	1.00						
ETOT	.26	.85	1.00					
ER	.30	.77	.74	1.00				
EYR	.17	.47	.75	.77	1.00			
EYRC	.36	.36	.40	.60	.54	1.00		
EYC	.03	.71	.69	.51	.26	.09	1.00	
EYP	.25	.65	.52	.11	- .11	.03	.45	1.00

Note. Abbreviations: ASMNT, Assessment scores; ET, Total senior level soccer experience; ETOT, Total soccer experience; ER, Senior level referee experience; EYR, Total referee experience; EYRC, College referee experience; EYP, Soccer playing experience; and EYC, Soccer coaching experience.

$n = 33$.

All correlations greater than .34 were significant ($p < .05$).

games) were played at the field where Assessor 4 was assigned. These findings suggest that assessment scores were accurate reflections of referee performance and that between-assessor mean differences were more due to game differences than rater error.

Peer ratings. For 36 participants, peer ratings were obtained from 5 volunteers who were either qualified assessors or senior referees from the state where Site 2 was located. Raters were selected because of their experience in the college referee system and the level of expertise that they had shown as either a referee or an assessor. Volunteer raters were contacted by telephone and read the names of 50 referees in the state, among whom were 37 referees who had participated in the present research. Then, raters were told to:

> " . . .. rate the overall performance of each individual as a linesman in a game at or equivalent to the college level. Use a scale similar to the normal game assessment scale as follows:
>
> | | | |
> |---|---|---|
> | Well above average. | One of the top 5% | 85 |
> | Above average. | One of the top 15% | 75 |
> | Average | One of the middle 40% | 70 |
> | Below average | One of the bottom 15% | 65 |
> | Well below Average | One of the bottom 5% | 55 |
>
> You may select these values or interpolate between values if appropriate. Use all information you have to arrive at your ratings. Try to rate all individuals. If you are unable to rate an individual, just indicate so. You may change a previous rating at any time."

Raters were told before and after the rating process that their participation was voluntary and that NISOA management supported the research. They were told that all ratings would remain confidential and that raters would not be identified with their ratings or with participation in the research. Ratings were entered into the computer as they were provided. The list was sorted and reviewed with the rater immediately after completion to provide a final opportunity for revision.

Because all raters rated the 36 participants, a two-way fixed effects model was appropriate for use in estimating rater consistency (Shrout & Fleiss, 1979). Because mean ratings across judges was used as the unit of analysis in this research, the appropriate measure of reliability of measures was the intraclass correlation coefficient (ICC) computed using Equation 3 and the mean squares reported in Table 4. The index of reliability of measures was .93. The 95% confidence interval for the population value of this ICC was $.78 < \rho < .99$. All correlations of ratings among the five peer raters were significant ($p < .01$). Tukey's (HSD) test was used to investigate the significant between-rater effect. No significant pairwise comparisons were found.

The correlations of the 36 average ratings with the five experience variables are shown in Table 5. Significant correlations were found for total referee experience (r = .45), player experience (r = .49), and total senior level soccer experience (r = .49). Of these, only total senior level soccer experience showed significant correlations with both assessment scores and peer ratings. Both assessment scores and expert ratings were

Table 4

Analysis of Variance Of Peer Ratings

Source	df	MS	F-ratio
Between Participants	34	141.83	13.43 **
Within Participants			
Between Raters	4	36.02	3.67 *
Residual	136	9.81	

* $p < .05$. ** $p < .01$.

available only for twelve participants, too small a sample to permit comparison of ratings and assessments.

The evidence from the analyses of assessment scores and peer ratings was strong that experience could be used as a measure of past referee performance in the research design. The evidence was also strong that total senior level soccer experience was the most suitable measure of past performance. Correlations of senior level soccer experience as a trichotomous variable with ratings ($r = .46$, $p < .01$) and assessment scores ($r = .34$, $p < .05$) were significantly greater than zero and highly similar to the values obtained as a continuous variable.

Assignment to conditions. Before testing, participants were assigned randomly to receive one of the two physical demands during the game simulation. Participants in the low physical demand group stood and viewed the monitor during testing. Participants in the high demand group ran between viewing game segments on the monitor to simulate the physical demand of a soccer game. After testing, participants were assigned to one of three experience groups (i.e., low, medium, or high experience). Table 6 shows the distribution of participants in treatment groups.

Analysis of demographic variables. Analyses were done to identify significant differences in treatment groups for participant sex and age. Because data collection time

Table 5

Intercorrelations Between Peer Ratings and Measures of Experience

	RATING	ET	ETOT	ER	EYR	EYRC	EYC	EYP
RATING	1.00							
ET	.49	1.00						
ETOT	.45	.78	1.00					
ER	.33	.80	.52	1.00				
EYR	.20	.54	.70	.67	1.00			
EYRC	.14	.69	.42	.87	.66	1.00		
EYC	.27	.31	.70	-.02	.21	-.13	1.00	
EYP	.49	.74	.77	.27	.14	.17	.57	1.00

Note. Abbreviations: Rating,Peer rating; ET, Total senior level soccer experience; ETOT, Total soccer experience; ER, Senior level referee experience; EYR, Total referee experience; EYRC, College referee experience; EYP, Playing experience; and EYC, Coaching experience.

n = 36.

Correlations greater than .33 were significant (p < .05). Correlations greater than .42 were significant (p < .01).

Table 6

Numbers of Participants By Physical Demand and Experience Group

	Total Senior Level Soccer Experience		
Demand	Low	Medium	High
Low	12	10	8
High	11	9	11

requirements and participant availability forced the use of two testing sites, site was included as an independent variable along with demand, experience, and their interactions. Participant age and sex were used as dependent variables. Inspection of Tables 7 and 8 shows that there were no significant differences in sex or age by participant Demand, Experience or Demand x Experience groups. Two significant interactions with site found. In the analysis of participant sex, the Experience x Site interaction was significant. The interaction was not considered practically significant because the total number of women included in the sample was small (i.e., $\underline{n} = 5$). The 3-way interaction in the analysis of age was also significant. It too was considered not practically significant.

The age results showed that there were no significant relationships between age and the independent variables of this research (i.e., demand, experience, or site). Inspection of Table 26 in Appendix B showed the presence of a significant correlation between age and field level game simulation score ($\underline{r} = -.31$, $\underline{p} < .05$). A repeated measures analysis of variance was done. The dichotomous variable age served as the between variable and game simulation scores for the two camera angles as the repeated measure (i.e., field level and press box level). A significant effect for age was found, i.e., $\underline{F}$ (1, 59) = 6.54, $\underline{p} < .05$. Thus, although there were no significant relationships between age and the independent variables in the research design, the relationship between age and the dependent variable was significant. Therefore, age was retained as a covariate to control for its effects.

Inspection of Table 9 showed that differences in total senior level soccer experience did occur across treatment groups. The main effect for experience groups was ignored, because it was by total experience that experience groups were formed. Main effects by site or demand were not significant; however, two site interactions were significant (i.e., Demand x Site and Demand x Experience x Site). Inspection of participant data showed that five of the six most experienced participants were tested at Site 2. All five were assigned to the low physical demand group. Because of the strong relationship found

Table 7

Analysis of Variance Of Participant Sex By Treatment Group

Source	df	MS	F-ratio
Demand (D)	1	0.04	0.50
Experience (E)	2	0.11	1.51
D x E	2	0.08	1.12
Site (S)	1	0.19	2.74
D x S	1	0.00	0.06
E x S	2	0.26	3.72 *
D x E x S	2	0.11	1.51
Error	49	0.07	

* p < .05.

Table 8

Analysis of Variance Of Participant Age By Treatment Group

Source	df	MS	F-ratio
Demand (D)	1	107.88	1.86
Experience (E)	2	10.18	0.18
D x E	2	18.23	0.31
Site (S)	1	13.84	0.24
D x S	1	100.88	1.74
E x S	2	59.90	1.03
D x E x S	2	260.70	4.50 *
Error	49	57.99	

* p < .05.

Table 9

Analysis of Variance of Total Senior Level Soccer Experience By Treatment Group

Source	df	MS	F-ratio
Demand (D)	1	0.59	0.02
Experience (E)	2	1922.06	60.77 **
D x E	2	3.36	0.11
Site (S)	1	0.46	1.60
D x S	1	153.92	4.87 *
E x S	2	30.19	0.95
D x E x S	2	102.38	3.24 *
Error	49	31.63	

*$p < .05$. **$p < .01$.

between performance (i.e., peer ratings and assessment scores) and experience, it was possible that site impacted game simulation performance. Therefore, site was included as an independent variable in the research design to control for its effects.

The resulting design was a 2 x 2 x 3 fixed effects factorial with two levels of physical demand, two testing sites, and three levels of experience. Age was used as a covariate. Subjects were nested in physical demand, site, and experience. Within treatment groups, participants were administered a battery of tests—a game simulation, a situational interview, a job knowledge test, and a physical performance test battery. All tests were developed by means of a content-oriented strategy (Alba & Dickinson, 1985; Schmitt & Ostroff, 1986). The game simulation was composed of segments from two different camera angles. Performance scores for the two camera angles were treated as separate dependent variables.

Test Development and Plan for Administration

Test development followed the procedures suggested by Alba and Dickinson (1985). A list of distinct, behaviorally defined job tasks was developed. Those job elements were rated by experts to determine their importance to the job and the referee selection process. A job element map was created based on the ratings. Tests were systematically constructed through the use of this map.

After pilot administration and revision, the tests were administered to the research participants. Because the major research focus was the game simulation, test sequence was to remain constant. The game simulation was to be administered first to minimize the contamination of test results from other tests. The tests were to be administered following a 10-minute pre-game briefing that provided participants with specific instructions to guide their performance throughout the testing. The order of the tests was to be as follows: (a) game simulation (30 minutes), (b) situational interview or BEI (30 minutes), and (c) multiple choice knowledge test (45 minutes), and (d) the physical performance test battery (45 minutes). The simulations and the knowledge test were to be administered in a two-hour block, while the physical performance test battery was to be administered as clinic schedules and participant availability permitted.

The procedures were modified at Site 1 because of imposed constraints by the sponsoring organization. The written test was administered to participants during the four-hour "arrival period" on the first day of the five day camp. Participants were provided an overview of the research, asked to complete the necessary waivers and consent forms, and given the written test to be taken in their rooms. Tests were completed and returned before the end of activities that night. The game simulation and the BEI were administered in one-hour blocks on the third and fourth camp days. At Site 2, the planned administration sequence was followed.

Task generation. In a soccer game, NISOA officials do one of two distinctly different jobs (i.e., referee and linesman). This research focused only on the job of linesman because the job is less complex and more likely to be performed by entry level officials. Extension to the job of referee was considered to be straightforward once the method was developed.

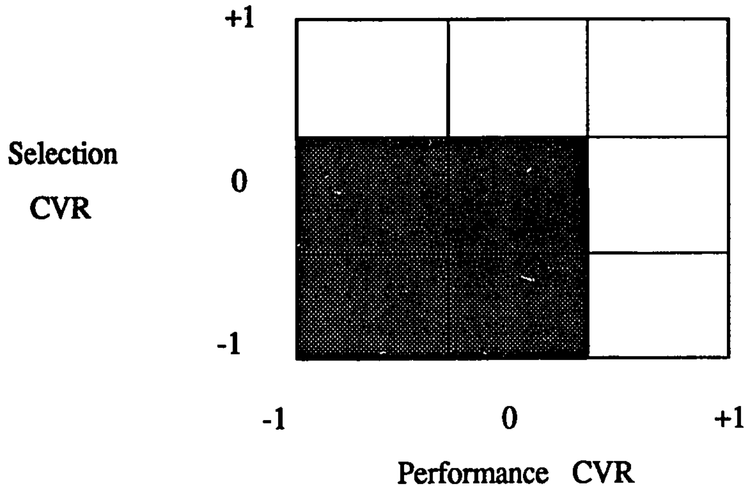
Linesman duties are stated in the NCAA Soccer Rules in terms of outcomes to be achieved, tasks to be accomplished, and "mechanics" to be used (e.g., positions to take and signals to give). An initial list of job elements was developed from the rule book. The list of job requirements was reviewed by 10 job experts. Because of the geographic locations of the experts, meetings were unrealistic. The delphi technique (Dalkey & Helmer, 1963) was used, because it takes advantage of the strengths of interactive groups and tends to provide better group decisions while avoiding their major weaknesses (Martino, 1983). A group size of 10 individuals has been shown to provide adequate reliability of results (Fusfeld & Foster, 1971). Anonymity was maintained throughout the research. An iterative process was used, but only two iterations were required because few changes were received from the second survey. The resulting list identified 36 job elements for linesmen in college soccer.

Job element rating and mapping. A questionnaire was constructed to gather data about the importance of job elements. Twenty different volunteer job experts were asked to rate job elements on a 7-point rating scale in terms of perceived importance to (a) job performance and (b) selection (job entry) screening. Ratings were trichotomized (i.e., important, unimportant, and neutral) as suggested by Ford & Wroton (1984) to permit CVR computation (Lawshe, 1975). The resulting CVR values were reviewed by the local panel of experts.

Job elements were mapped by CVR values as shown in Figure 1 to guide the process of choosing job elements for testing. The CVR map was used to choose the sample of tasks to be tested. Half of each test, including the work sample, was made by sampling job elements in the upper right area because these elements were consistently rated as important to both the job and the selection process.

Elements in the shaded area (i.e., the lower left where elements were consistently rated as unimportant and the three adjacent boxes where ratings tended to be unimportant or neutral) were reviewed by the local expert panel. Where reasons for inclusion (e.g., safety or game control) were identified, these elements were sampled for testing. The remainder of each test was made by sampling elements with one high positive CVR

Figure 1. Graphic mapping framework for content-validity-ratio values



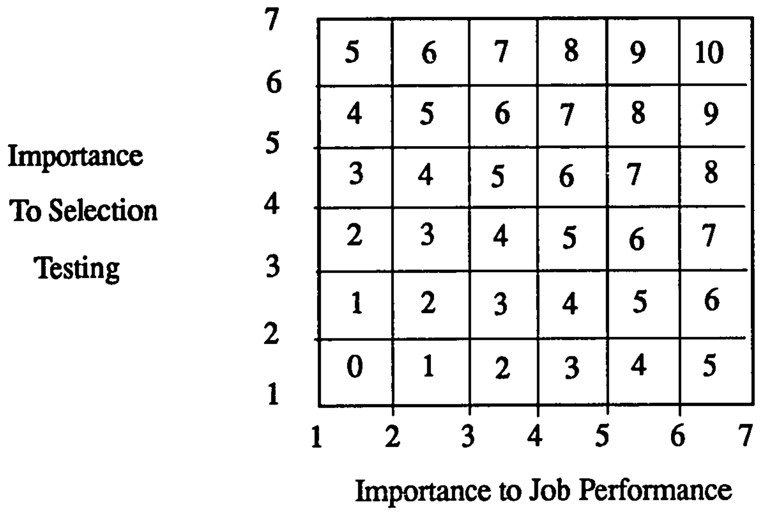
because they were consistently seen as important to either performance or selection.

Job expert input was also used to produce test item weights because some game situations were more critical than others. Weighting of test items was done to assess whether participant ranking by test score was significantly impacted by the method of scoring. Mean values of job element importance were plotted in graph form similar to Figure 2. Job element weights (0 to 10) were assigned to each unit block so that more important elements were weighted higher. For example, if an item received a mean importance to selection testing of 2.3 and a mean importance to job performance of 4.6, that item received a weight of 4. Test items (i.e., work sample segments, BEI questions, and written test items) were assigned the weight of the job element that they tested.

Three scoring schemes were used for the written test and the BEI. Total test scores were computed from unweighted and weighted item scores. Total test scores were also computed from job component percentage scores, creating three unit blocks which were averaged. The lowest correlation among scores was .91. Therefore, unweighted total scores were used throughout the research because they would be the easiest to compute in actual field settings.

Job components. The panel of experts reviewed the CVR map, the testing plan, and

Figure 2. Job element ratings of importance.



5	6	7	8	9	10
4	5	6	7	8	9
3	4	5	6	7	8
2	3	4	5	6	7
1	2	3	4	5	6
0	1	2	3	4	5

the rule book to identify appropriate Job Components for use in this research. Job Components were to be combinations of the job tasks sampled during testing. The experts agreed unanimously that pre-game activities (Job Component A) were important and different from tasks required during the game. Activities during the game tend to occur as a result of a combination of visual and oral input; however, the panel agreed that they should be divided into two categories. Some actions are routine and involve no judgment (e.g., if the referee observes cue X, then action Y, and only action Y, is appropriate). For these situations, actions tend to be correct unless the cue is not seen or the rules are not properly applied. These activities were grouped into Job Component B.

The remaining referee actions involve non-routine situations that require referee judgments which tend to be critical to referee success in the game. The panel felt that these situations were "game critical" and, from their experience, tended to differentiate among top referees at the senior level. An effective test for senior referees must differentiate performance along this dimension. Therefore, non-routine situations and situations where referee judgment was required were grouped as Job Component C.

Oral Instructions

Before a soccer game, the referee confers with the linesmen. The pre-game conference permits the referee to review the rules, the generally accepted referee and linesman mechanics, and the specific referee and linesman duties and responsibilities. It also permits the referee to specify linesmen actions that the game may dictate or that the particular referee may prefer where the actions are not mandated, but are left to the discretion of the referee.

The local expert panel constructed a script for a typical pre-game conference using the results of the job element questionnaire. The script was videotaped for presentation and pilot tested with local referees to identify ambiguities and avoidable difficulties. In the pilot test, two groups of two referees each received the pre-game instructions both videotaped and live, but in opposite sequence. One group saw the videotape first; the other received the instructions live first. The four referees unanimously reported that they preferred the "live" version to the videotape mainly because they could ask questions, even though questions were handled by rereading the section of the instructions related to the question. Based on these recommendations, the "live" version was used.

Game Simulation

<u>Description</u>. Because fidelity of a complete game simulation was impractical, videotapes of actual game segments were used. It was felt that advantages realized from the use of videotape segments (e.g., standardization of administration and scoring) outweighed the disadvantages (e.g., loss of actual game context).

Segments were selected from films of various games according to the CVR map. Segments were selected that would require specific linesman decisions and elicit behaviors clearly identified with those decisions. Segments requiring no action by the official were also selected and randomly placed in the simulation.

The tape was piloted in a group setting with 20 NISOA referees at a referee clinic. During the scoring of segments, the referees reported several problems with the tape, the most serious being that the tape was not representative of a "real" soccer game. The color of the player uniforms was not consistent throughout the tape. There was little or no consistency in field layout and markings. Camera location and angle varied. It was impossible to keep track of which direction the teams were attacking.

A second tape was prepared such that segments were drawn from one game. To determine which game to use, 10 games were previewed. Lists of actual events were compiled for each game. CVR values for game events were compiled. A measure of total game relevance was calculated as the total of all CVR values. The game which provided the highest composite CVR was selected.

Videotape segments were extracted for the game simulation. Segments were extracted in game sequence to preserve the "context" of the game itself. Events requiring no action were randomly interspersed in the tape. Blank space (3-6 sec.) was left between segments. A special-effects generator was used to fade into and out of segments. Referees reported the fade as distracting and not representative of actual game conditions. As an alternative, action was "frozen" on the screen during the space. Officials reported preferring this process, because it was more realistic, let them anticipate the play before the action began, and provided valuable visual cues before and after-the-play much like a real game.

One variable of additional concern in developing the tape was the effect of the camera angle from which events were filmed (i.e., field level or press box level). A videotape of a game involving two teams with similar uniform colors was found that was taped from field level. Twenty segments were extracted from this tape and placed at the beginning of the test tape. This tape was shown to the three local referees who were used for pilot testing. A much more positive response was received to the new tape. The first 20 segments were reported as more difficult to follow, but more representative of the job of linesman. The tape was reported to be "very realistic. "

Administration. The 68-segment videotape was presented to participants on a 19" color television (TV) placed at about their eye level. The test administrator sat behind the participants, controlling the video recorder with a remote control. Each administration required about 30 minutes (M = 28.50, SD = 3.62). Mean time to complete the game simulation was significantly different by site (t = 2.89, p < .01). Completion at Site 1 (M = 27.32, SD = 3.70) was significantly faster than at Site 2 (M = 29.84, SD = 3.12). Inspection of Tables 8 and 9 shows that this difference was not due to differences in age or experience. One possible explanation of the difference is the time pressure imposed at Site 1 where both the game simulation and the BEI were completed within one hour so that participants could return to the regular camp activities. Similar pressures were not present at Site 2.

A standardized checklist was used to document participant actions for each segment. The administrator's job during testing was limited to providing instructions to participants before the test began and recording participant responses. The instructions told participants to adjust their distance from the TV as needed to facilitate viewing. Members of the high demand group were instructed to run between segments, but only when instructed and at a normal jogging pace. Testing was done in rectangular rooms which permitted running to a point 30-40 feet from the TV and return. The next segment was not shown until the participant returned to the viewing position. Members of the low demand group were instructed to stand in front of the TV during the entire game simulation.

Scoring scheme: The local expert panel and I viewed each segment twice before preparing a list of possible referee actions. Individual lists of possible actions for each segment were pooled. Panel members rated the appropriateness of possible actions for each segment using the 5-point scale displayed in Table 10. This scheme was used to permit the identification of referees who made the most appropriate decisions in each situation as Vikhrov (1978) suggested. Ratings for the first five segments were compared and differences reconciled before continuing. In all, 397 possible actions were rated by the 4 experts.

An analysis of variance was done on the expert panel's ratings to assess interrater reliability. The summary of the analysis is shown in Table 11. The reliability for a typical expert was .73. The reliability for the average of the expert ratings was .91.

Table 10

Rating Scale For Alternative Actions in Video Segments

Rating	Meaning
5	Action is highly appropriate in this situation and/or is essential for game control.
4	Action is appropriate, but not essential, and/or will have positive impact on referee game control.
3	Action is routine and/or will have little or no impact on game control.
2	Action is slightly inappropriate and/or will have a negative impact on referee game control.
1	Action is highly inappropriate and/or will have significant negative impact on referee game control

Table 11

Analysis of Variance of Expert Judgments of Videotape Segment Weights

Source	df	MS	F-ratio
Experts	3	5.33	11.68 *
Measures	396	6.74	14.77 *
Error	1188	0.46	

* p < .01.

It had been anticipated that assessing the appropriateness of participant actions would be complicated by the need to consider what actions had been taken in previous situations in the game simulation. Soccer officials' actions vary between events, depending on the circumstances surrounding each event. Often, acceptable actions at one point in a game depend on actions taken in similar situations earlier in the game. Actions are situationally specific, and the assessment of the appropriateness of such actions is often very difficult. Thus, interrater reliability achieved was very encouraging.

A branching scoring network was anticipated where the scoring of some events would be dependent on responses in previous segments. The expert panel found only two instances where actions might be dependent. Alternative scoring procedures were provided for these segments.

Scorer training. The three volunteer scorers were trained to familiarize them with the procedures, the scoring guide, and the scoring forms before scoring the results of the game simulation. As part of the training session, 10 events from a practice data set were scored. Differences were reconciled before proceeding. Then, the results from 10 participants were scored. As shown in Table 12, score differences between participants were significant, and no differences occurred between scorers. Equation 3 was used to compute the appropriate index for reliability of the measures ($\underline{r} = .99$). The 95% confidence interval for the population value was $.97 < \underline{\rho} < 1.00$. These results provided support for Hypothesis 9, suggesting that scorers were interchangeable and that little measurement error was introduced by the soccer experience of the scorer.

Behavioral Event Interview

Description. Using questionnaire results and the test plan, a situational interview (BEI) was constructed. Questions focused on the critical job elements, especially those covered in the pre-game conference and those which could not be presented in the game simulation (e.g., interpersonal actions, unusual game situations, or emergencies such as light failure or a bench clearing brawl). At least 25% of the questions were written to be parallel with segments from the game simulation. The correlation between total scores for the parallel items was .53, providing support for Hypothesis 11.

Interview questions were constructed so that participants would have to describe the step-by-step actions to be taken in game situations. Questions were worded so that a novice referee candidate could provide a scoreable response (Ford & Wroten, 1984). Questions were placed in typical game sequence.

Table 12

Analysis of Variance of Game Simulation Total Score

Source	df	MS	F-ratio
Participants	9	993.11	254.64 *
Judges	2	0.70	0.16
Residual	18	4.26	

* $p < .01$.

The local expert panel reviewed the interview questions and developed a list of possible responses, instructions for asking clarifying questions, and procedures for recording participant responses. The interview was piloted with a group of three local referees. In a two-hour group session following the interviews, problems with the questions were identified. Final modifications were made to the interview based on these pilot-test data.

Administration. The local expert panel prepared a master checklist and guide to use in conducting the interview and recording participant responses. After completing the game simulation, participants were read the interview instructions. They were told (a) that the original oral (pre-game) instructions remained in effect, (b) that they would be presented with a series of game situations during the 30-minute interview, and (c) that they were to take the actions, if any, that they would as the linesman.

Scoring scheme. Using the interview questions, expert panel members developed a list of possible responses. Then, they rated the appropriateness of each response using a 5-point rating scale. Ratings were compared and differences reconciled before the master guide was prepared.

A pilot test of the scoring guide was done with the scorers from the game simulation. Each scored one interview set. In a one-half hour group session following the scoring, problems with the scoring guide were identified and final changes were made.

Testing of the scoring scheme was done with the three scorers. Each scorer used the master guide to score the interview responses of 10 participants. Total test scores and the three job component scores were compared to assess scoring consistency. Total test score reliability was computed using Equation 3 as .94. Measures of reliability for each job component, also computed using Equation 3, were consistent (i.e., .97, .97, and .95), providing evidence that the scoring procedures were equally effective for the job components. These results provide additional support for Hypothesis 9.

Job Knowledge Test

Description. The pencil-and-paper test was designed to assess participant technical knowledge. A job knowledge test was included in the battery because they have been found to be excellent predictors of training performance (Reilly & Chao, 1982). Pencil-and-paper tests in the form of general abilities tests have produced the highest average validity when work samples are used as criterion measures (Schmitt et al., 1984). Further, pencil-and-paper tests have displayed the least susceptibility to situational specificity of validity (Schmitt & Hunter, 1977; Hunter & Hunter, 1982). Finally, the validity of job knowledge tests was of interest to the sponsoring organization.

One hundred-thirty test items were written. One hundred items were written as matching pairs. The remaining 30 items were unique. All items were reviewed by the local expert panel for clarity and accuracy. Revisions were made in a two-hour session.

The technique suggested by Ford and Wroten (1985) was used to prepare two multiple-choice tests. Test items were grouped into six sub-sections to parallel the organization of the current rule book. These were (a) pre-game duties; (b) players, players equipment, and substitutions; (c) general referee duties; (d) timing, scoring, and normal play; (e) rule violations (e.g., fouls and misconduct); and (f) restarts. Each sub-section contained 10 to 15 questions arranged in random order. The test was piloted with a group of five local referees to identify problems with item and instruction wording. The expert panel reviewed the test and the comments from the pilot administration to finalize the test booklet, the answer key, and the administration guide.

Administration. Written tests were administered in a quiet location with no test administrator in the room. Assignment of test form (A or B) was done randomly. Participants were asked to complete the written test by marking the most correct response to each question on the answer sheet provided. At the end of testing, test booklets and answer sheets were returned to the administrator. The expert panel used the NCAA rules to prepare answer keys.

Test analysis and scoring scheme. In all, 85 referees from 3 clinics completed the written test. All answer sheets were used to compute preliminary item and total test statistics. Item statistics were compared for each item pair from Forms A and B. Where significant item difficulty differences were found, both items were eliminated from the test, unless review of the items by the expert panel found reason to retain them. Items (paired items and common items) were also eliminated if item difficulty values were less than .2 or greater than .9, unless a review found reason to retain an item. In all, 25 items were eliminated from each test form. Table 13 shows the mean differences between matching items for the resulting 55 item test.

Test items were grouped and scores computed for the three job components. The internal consistency of the three job component sections was assessed by means of coefficient alpha as shown in Table 14. Comparisons were also made of scores on items common to both tests. No significant differences in scores were found for the common items. Based on these results, it was concluded that test form could be excluded as a variable in this research.

Physical Performance Test

Description. The physical performance test battery suggested by Kuhnle and Yarbrough (1986) was administered to all participants. The battery consisted of four tests. In order, the tests were (a) an aerobic endurance run (i.e., a modified Cooper run), (b) a sprint (50 meters), (c) an agility test (8 x 10 meter shuttle), and (d) an anaerobic endurance run (5 x 60 meter run).

The battery was administered because the sponsoring organization was interested in

Table 13

Mean Differences of Matching Items From Test Forms A and B

Mean	Job Component				
Difference	A	B	C	N	Percentage
< .05	8	5	5	18	32.7
.05 < .10	3	7	4	14	25.5
.10 < .15	2	3	4	9	16.4
.15 < .20	2	0	1	3	5.4
.20 < .25	0	0	1	1	1.8
.25 < .30	2	1	2	5	4.5
> .30	3	0	2	5	4.5

Table 14

Coefficient Alpha For Sections of Written Tests

	Test Form	
Job Component	A	B
A	.64	.57
B	.43	.48
C	.61	.51

the predictive validity of the physical performance battery. Schmitt et al. (1984) reported high validity for three research investigations when work samples were used as the criterion. It was also hypothesized that physical performance (capability) impacted the level of accuracy achieved by officials in their decisions during a game (or a game simulation). Therefore, the correlation between game simulation scores and physical performance scores was investigated.

Scoring scheme. Results from each test were converted to standardized scores (i.e., M = 50, SD = 10) (Kuhnle & Yarbrough, 1986). A composite standardized score for the entire battery was also computed. The reliability of the physical performance battery had been established previously (Kuhnle & Yarbrough, 1986). Test-retest reliabilities ranged from a high of .91 for the Cooper Run to a low of .84 for the agility shuttle. The battery was administered using the procedures which produced those reliability estimates.

Post Questionnaire

The acceptability of each test was assessed by means of a brief attitudinal questionnaire following the test battery (Schmitt & Ostroff, 1986). Participants responded on a 7-point scale to identify the extent that the stated actions or outcomes were true. Verbal anchors were: almost never, to a very little extent, to a little extent, somewhat, to a great extent, to a very great extent, and almost always. All questions are contained in Appendix C.

For each component except the physical performance battery, participants were asked to judge how realistic the component was for testing college referees. For the game simulation, the BEI, and the written test, participants were asked to identify the extent that these tests measured their ability as a linesman.

The remaining four questions were directed at specific characteristics of the game simulation. They were asked because of the difficulties encountered in the preparation of the videotape. First, participants were asked to assess how well the videotape simulated game conditions. Second, they were asked to assess the flow of the videotape segments compared with the events in an actual game. Third, they were asked to assess how much the game simulation "felt" like a real game in terms of pressure placed on the referee. Finally, they were asked to identify the extent that the quality of the videotape interfered with their ability to make correct decisions.

Participant responses to the 11-item questionnaire are tabulated in Table 27 of Appendis D. Responses to items 2 and 7 were added to produce a composite Game Simulation measure. Responses to items 8 and 9 were added to produce a comparable

BEI measure. Responses to items 10 and 11 were added to produce a Written Test measure. Items 3, 4, 5, and 6 were combined to produce a measure of the acceptability of the format of the videotape. The scale of Item 6 was reversed when the scores were combined. Item 1 was analyzed alone as a measure of the acceptability of the oral instructions.

III. RESULTS

Analytic Approach

The focus of this research was on the game simulation and other more conventional tests as methods for assessing the performance of soccer referees. To that end, analyses were done in the following sequence. First, the criterion-relatedness of the game simulation was assessed. Second, the construct-relatedness of test components was assessed. Third, the validity of hypotheses relating to the influence of site, physical demand, experience, and camera angle on game simulation scores was assessed. Fourth, relationship between game simulation score and other tests was investigated. Finally, post questionnaire responses were analyzed for information about test acceptability to participants.

Criterion-relatedness. The relatedness of the game simulation to work performance was evaluated by comparing game simulation scores to peer ratings and game assessment scores.

Construct-relatedness. The primary focus of this research was to examine the ability of multiple methods to assess various components of linesman performance. Convergent validity, discriminant validity, and method bias were evaluated with multitrait-multimethod analyses of variance (Kavanaugh et al., 1971). In these analyses, the multi-methods were a game simulation, a BEI, and a written test. A main focus of this research was to assess differences between game simulation scores from two camera angles. Therefore, the two camera angles were analyzed as different methods. The multi-traits were the three job components. Analyses were done for the total sample and the three experience groups to assess differences between those groups.

Game simulation model. Analysis of variance techniques were used to assess the effects of physical demand, testing site, and experience on game simulation scores. Participant age was included as a covariate to control for its effects. Separate analyses were done to investigate the sources of variance in game simulation scores from the two camera angles. The suitability of the two camera angles was assessed by comparing analysis of variance results.

Relationship of game simulation and other tests. Linear regression analyses and analysis of variance techniques were used to investigate the relationships among the various tests and their components to determine the most appropriate test components for testing referees with different levels of experience.

Criterion-Relatedness Results

General performance peer ratings. The correlations among average peer ratings and game simulation scores are shown in Table 15. Ratings correlated significantly with field level scores ($p < .05$), press box level scores ($p < .01$), and total game simulation scores ($p < .01$). These correlations between ratings and game simulation scores provide strong support for Hypothesis 1.

Inspection of Table 15 shows that the correlation of ratings with press box level scores was greater than its correlation with field level game simulation scores; however, the difference was not significant ($t = .81$, $p > .05$). Also, the correlation between field level and press box level scores was not significant, suggesting that there was a difference in the effectiveness of the two camera angles in soccer referee testing. The correlations among field level scores, press box scores and ratings suggest that the game simulation from the press box angle was better as a predictor of game simulation scores.

Game assessment scores. The correlations of assessment scores with game simulation scores are shown in Table 16. The correlation of assessment scores with field level scores was not significant ($p > .05$). However, the correlations of assessments with total game simulation scores and press box level scores were significant ($p < .01$), suggesting that the game simulation from the press box level was a better predictor of game performance than the game simulation from field level.

The correlations between assessment scores and game simulation scores shown in Table 16 provide additional support for Hypothesis 1. The presence of a significant correlation between assessment scores and game simulation scores from the press box camera angle and the absence of a significant correlation with scores from the field level

Table 15

Intercorrelations Between Game Simulation Scores and Peer Ratings

	Rating	GS Total	Fieldlvl	Pressbox
Rating	1.00			
GS Total	.65 **	1.00		
Fieldlvl	.36 *	.68**	1.00	
Pressbox	.54**	.90**	.31	1.00

Note: Abbreviations: Rating, Peer ratings; GS Total, Game simulation total score; Fieldlvl, Game simulation score from field level; and Pressbox, Game simulation score from press box level.

n = 36.

* p < .05. ** p < .01.

Table 16

Intercorrelations Between Game Simulation Scores and Assessment Scores

	Asmnt	GS Total	Fieldlvl	Pressbox
Asmnt	1.00			
GS Total	.50*	1.00		
Fieldlvl	.16	.65*	1.00	
Pressbox	.54*	.84*	.17	1.00

Note: Abbreviations: Asmnt, Assessment scores; GS Total, Game simulation total score; Fieldlvl, Game simulation score from field level; and Pressbox, Game simulation score from press box level.

n = 33.

*p < .01.

camera angle supported Hypothesis 7 and suggested the use of press box level camera angle for referee testing.

The combined evidence from Tables 15 and 16 is strong that the game simulation was an effective measure of both short term (game assessment) and long term (peer ratings) linesman performance at the senior level, thus supporting Hypothesis 1 that game simulation scores could be used as criterion measures against which to compare performance on other tests. Correlational evidence from ratings and assessment scores suggests that the game simulation from the press box camera angle was superior to the field level camera angle for ranking referees, thus supporting Hypothesis 7.

Construct-Relatedness Results

Multitrait-multimethod analyses were done to assess Hypotheses 2, 3, and 4 for the methods and traits (i.e., job components) of interest in this research. In addition to analyses of results from the total sample, analyses were done for each experience group. The low experience group was of interest because decisions about entry selection are more likely to occur for this group. The high experience group was of interest because it is from this group that choices would be made for significant games (e.g., tournaments, championships). Table 17 summarizes the 15 multitrait-multimethod results that are provided in Appendix E.

Support for Hypothesis 2 required moderate to high ICCs for the Participant source of variance (e.g., .20 or larger; Dickinson, Hassett & Tannenbaum, 1986). Inspection of Table 17 showed that 10 of 15 Participant ICCs were below .20. Thus, in general, convergent validity was low, and Hypothesis 2 was not supported. Evidence of moderate convergent validity was found in five analyses that compared game simulation and BEI scores. In three of the five analyses, the press box game simulation method was isolated with the BEI. In the fourth analysis, the field level game simulation score was added as a third method. In the fifth analysis, the total game simulation score (i.e., the sum scores from the field level, press box level and pre-game components) was used. Inspection of Table 17 shows that, in the total sample and in each age group, adding the field level game simulation (i.e., C1) as a third method to the press box (i.e., C2) and BEI methods decreased the Participant source of variance. This evidence also suggests that the press box game simulation was superior to the field level game simulation for the testing of soccer referees, thus supporting Hypothesis 7.

Support for Hypothesis 3 required moderate to high ICCs for the Participant x Trait source of variance. Inspection of Table 17 shows 13 of 15 ICC values ranging from .04 to .18. Further, discriminant validity declined with increasing experience for the methods tested. Thus, discriminant validity was low, and Hypothesis 3 was not supported. Moderate discriminant validity was found in the total sample when game simulation scores from the two camera angles were used as methods. This result suggests that the game simulation was able to differentiate between these job components across the sample.

Further investigation showed evidence of low discriminant validity in the medium (ICC = .10) and high (ICC = .04) experience groups. In contrast, high discriminant validity (ICC = .43) was found in the low experience group. This latter finding suggests that referees of low experience perform at different levels for Job Components B and C and that the game simulation is effective for detecting these differences.

Support for Hypothesis 4 required low ICCs for the Participant x Method source of variance. In the 15 multitrait-multimethod analyses, method bias ranged from very low (i.e., ICC = .04) to moderate (i.e., ICC = .27). Inspection of the Participant x Method ICCs in Table 17 shows that method bias was low (i.e., less than .20) except in the high experience group where it was consistently moderate (i.e., .20 to less than .30), suggesting that differences in the ordering of participants in the high experience group were due to differences in camera angle. Thus, in the high experience group, support for Hypothesis 4 was not shown.

In the low and medium experience groups, score differences were not due to testing methods. The evidence did not clearly show one game simulation method to be superior; however, the total multitrait-multimethod evidence suggested that camera angle was a

Table 17

Intraclass Correlation Coefficients From Multitrait-Multimethod Analyses

	Job	Experience	ICC		
Methods	Components	Level	P	P x T	P x M
GS, BEI	A, B, C	—	.28	.05	.06
GS, BEI, WT	A, B, C	—	.16	.04	.18
C1, C2	B, C	—	.17	.22	.14
C2, BEI	B, C	—	.24	.10	.13
C1, C2, BEI	B, C	—	.19	.13	.15
C1, C2, BEI, WT	B, C	—	.15	.13	.18
C1, C2	B, C	High	.13	.04	.26
C2, BEI	B, C	High	.28	.06	.23
C1, C2, BEI	B, C	High	.17	.05	.27
C1, C2	B, C	Medium	.17	.10	.13
C2, BEI	B, C	Medium	.24	.14	.18
C1, C2, BEI	B, C	Medium	.22	.14	.14
C1, C2	B, C	Low	.17	.43	.04
C2, BEI	B, C	Low	.14	.15	.01
C1, C2, BEI	B, C	Low	.14	.18	.07

Note: Abbreviations: ICC, Intraclass correlation coefficient; P, Participant source of variance; P x T, Participant x trait source of variance; P x M, Participant x method source of variance; GS, Game simulation total score; C1, Field level camera angle; C2, Press box camera angle; BEI, Behavioral event interview; WT, Written test; A, Pre-game job component; B, Routine game component; and C, Non-routine game component.

significant factor, especially in the high experience group. Further, the press box game simulation produced higher values of convergent and discriminant validity for the medium and high experience groups when analyzed with BEI scores.

Game Simulation Results

In the research design, site was included to control for its effects. Inspection of Tables 18 and 19 shows that the main effect for site and its interactions were not significant. Nonetheless, preliminary tests were not conducted for the purpose of pooling the site effects, because only fixed factors were involved (Winer, 1971).

Support for Hypothesis 5 required differences due to physical demand (D) or the Demand x Experience interaction. Inspection of Tables 18 and 19 shows that these effects were not significant ($\underline{p} > .05$), suggesting that physical demand did not impact game simulation performance.

Support for Hypothesis 6 required a significant main effect for experience (E) and an absence of interactions involving experience. Inspection of Table 18 shows that there were no significant experience effects in the analysis of field level scores. However, inspection of Table 19 shows that the main effect for experience was significant for press box level scores ($\underline{p} < .05$). This latter finding provides support for Hypothesis 6.

Tukey's HSD test was done to determine the source of the significant experience effect for press box level scores. Inspection of Table 20 shows that the mean scores of the low and high experience groups differed significantly ($\underline{p} < .01$). Differences in scores between the low and medium experience groups approached significance (i.e., $\underline{p} < .06$). These findings provide additional support for Hypothesis 6 and suggest that the game simulation from the press box level can be used to differentiate between referees with low experience and those with more experience.

Support for Hypothesis 7 required that the correlation between game simulation scores from the two camera angles be low and that additional evidence (i.e., external and internal validity) show one camera angle to be superior for testing soccer referees. Inspection of Table 26 of Appendix B indicates that the correlation between camera angle

Table 18

Analysis of Variance Of Game Simulation Scores From Field Level

Source	df	MS	F-ratio
Demand (D)	1	14.75	0.37
Experience (E)	2	69.13	1.72
D X E	2	67.51	1.68
Site (S)	1	23.50	0.59
D x S	1	153.60	3.82
E x S	2	44.80	1.12
D x E x S	2	89.27	2.22
Age	1	124.37	3.10
Error	48	40.18	

Table 19

Analysis of Variance Of Game Simulation Scores From The Press Box Level

Source	df	MS	F-ratio
Demand (D)	1	375.55	2.79
Experience (E)	2	540.67	4.02 *
D x E	2	32.11	0.24
Site (S)	1	397.61	2.96
D x S	1	60.12	0.45
E x S	2	120.31	0.90
D x E x S	2	193.60	1.44
Age	1	38.73	0.29
Error	48	134.43	

* $p < .05$.

Table 20

Tukey HSD of Experience Effect For Game Simulation at Press Box Level

	Experience Group		
Experience Groups	Low	Medium	High
Low	0.00		
Medium	7.61 *	0.00	
High	12.29 **	4.68	0.00

*p < .06. **p < .01.

scores was not significant (r = .19, p > .05). Inspection of Tables 18 through 20 and the multitrait-multimethod results of Table 17 provides strong evidence of the external and internal validity of press box camera angle scores. Thus, the field level camera angle was eliminated from further analysis.

Comparison of Game Simulation From The Press Box Level With Other Tests

The game simulation is expensive and time consuming to prepare and administer. If more conventional tests are highly correlated with game simulation scores, the more conventional tests could be used for referee selection. Stepwise linear regressions tested the ability of other tests and components of those tests to predict game simulation scores. The criterion for entry and removal of predictors was relaxed to $\alpha = .10$ so that predictors approaching significance would be identified. Because game simulation scores varied significantly with experience, separate regressions were also done for each experience group to assess whether different tests were appropriate for each group.

Support for Hypothesis 8 required evidence that performance on conventional tests was strongly correlated with game simulation scores. Table 21 displays the relationship of total test scores from the BEI, the written test, and the physical performance test with the press box game simulation score as a result of the stepwise regression analysis. The

Table 21

Stepwise Regression Analysis Of Behavioral Event Interview, Written Test, and Physical Performance Test Total Scores On Press Box Game Simulation Scores

Experience Group	n	Test	Corr
Total Sample	61	BEI	.36 **
Low	23	BEI	.42 *
Medium	19	-	
High	19	-	

Note: Abbreviations: Corr, Correlation coefficient; and BEI, Behavioral event interview.

*$p < .05$. **$p < .01$.

results show that the BEI total score was a significant predictor of game simulation score in the total sample. When analyses were done for each experience group, the BEI was a significant predictor ($r = .42$, $p < .05$) in the low experience group. No substitutes for the game simulation score were found in either the medium or high experience groups.

Stepwise regression analyses were also done with the components of all tests. Inspection of Table 22 shows that a written test for Rule 5 (i.e., Fouls and Misconduct), the physical performance test of agility (i.e., the 8 x 10 meter shuttle), and the BEI sub-test for job component B could be used to predict performance on the press box game simulation for the total sample. Investigation of the results of the analyses by experience group shows that the written test for Rule 5 was a significant predictor of game simulation score in the low and high experience groups ($p < .05$). Adding the 8 x 10 meter shuttle from the physical performance battery improved prediction for the low experience group ($r = .63$, $p < .01$). Prediction in the medium experience group was not significant ($p > .05$). Although support for Hypothesis 8 was found in the low and high experience group, the evidence about conventional tests did not provide strong support for Hypothesis 8.

Table 22

Stepwise Regression Analysis Of Test Component Scores On Press Box Game Simulation Scores

Experience	n	Test	R
Total Sample	61	R(5)	.37 ***
		FS(3)	.51 ***
		ID(2)	.58 ***
Low	23	R(5)	.49 **
		FS(3)	.63 ***
Medium	19	R(5)	.43 *
High	19	R(5)	.45 **

Note: Abbreviations: Corr, correlation coefficient; ID(3), Behavioral event interview for job component C; R(5), Written test for rule 5; and FS(3), Standardized score for 8 x 10 meter agility run.

Correlations are multiple correlation coefficients when more than one test entered the equation.

* p < .10. ** p < .05. ***p < .01.

Post Questionnaire

The oral instructions item, the three composite test items (i.e., the game simulation, BEI, and written test), and the videotape format item were tested by means of analysis of variance techniques for differences by physical demand, experience, site, and age groups. Inspection of Tables 43 - 47 in Appendix F showed that no significant effects were found. Inspection of Table 43 showed that the site effect for the oral instructions item approached significance (i.e., p < .06) with the mean response from Site 1 (M = 4.73) less than the response from Site 2 (M = 5.39). This finding was expected, because pre-game instructions vary geographically and the pre-game instructions were prepared by a panel of referees from the locale of Site 2. Only 3 of 61 participants (5%) viewed the oral instructions as unrealistic for college soccer. Thus, evidence was found to support Hypothesis 10(a).

Support for Hypothesis 10(b) required significant differences between mean ratings of the game simulation, the BEI, and the written test with the game simulation being higher than either the BEI or the written test. Inspection of Table 27 in Appendix D shows that the BEI composite item had the greatest mean and the game simulation item the smallest. Inspection of the item mean scores indicates that only item five was less than the scale midpoint (i.e., 4.0), suggesting that all three tests were judged as acceptable by participants. The difference in means between the composite BEI item and the composite game simulation item was significant ($\underline{t} = 2.57$, $\underline{p} < .05$), suggesting that the BEI was viewed as more acceptable than the game simulation. Thus, evidence to support Hypothesis 10(b) was not found.

Support for Hypothesis 10(c) required favorable reactions to questions about the videotape format of the game simulation. The composite videotape format item was not significantly greater than the scale midpoint, suggesting that the videotape format was not consistently viewed as acceptable.

The mean for item 3 (i.e., "did the videotape simulate actual game conditions") was significantly greater than 4.0 ($\underline{p} < .01$). Of the 61 participants, 39 (64%) rated the videotape as simulating actual game conditions. Of the 15 (19.7%) who said that the videotape test did not simulate actual game conditions, only four participants rated it less than 3.0 (i.e., "to a very little extent" or lower).

The mean for item 4 (i.e., "did events flow as in an actual game") was not significantly greater than 4.0 ($\underline{p} > .05$). Nineteen participants rated the flow as being other than that expected in a college game. Further inspection showed that 16 of the 19 participants who rated the flow as being other than that of a college soccer game (84%) were from the low or medium experience group.

The mean for item 5 (i.e., "did the videotape capture the emotion and pressure of an actual game") was significantly less than 4.0 ($\underline{p} < .01$). Forty-two of the 61 ratings were below 4.0, suggesting that the game simulation as presented did not capture the "flavor" of an actual game.

The mean for item 6 (i.e., "did the videotape interfere with decision-making") was significantly less than 4.0 ($p < .01$), suggesting that the quality of the videotape was not perceived as interfering with ability to make correct decisions. Of the 11 participants (18%) who rated the videotape quality as interfering with their ability to make correct decisions, 7 were from the low experience group. Further investigation showed that mean game simulation score for these 11 participants was not significantly different from the mean score of the remaining 50 participants ($p > .05$), suggesting that the quality of the videotape did not interfere with performance. Post-hoc Tukey (HSD) tests of the total sample showed that the mean response in the low experience group ($M = 3.83$) was significantly greater ($p < .01$) than that of the high experience group ($M = 2.63$), suggesting that the acceptability of the game simulation increased as participant experience increased.

Inspection of Table 23 shows that the correlation between the videotape format item and game simulation score was not significant ($p > .05$). The correlation between game simulation score at the press box level and the game simulation item was significant ($p < .05$), but the game simulation item did not account for significant variance in the research design. These findings provide additional support for Hypothesis 10(c) that the quality of the videotape did not interfere with the decision-making ability of the participants.

In summary, questionnaire results showed that (a) the oral instructions were viewed as realistic, (b) the game simulation was not rated as better for assessing linesman performance compared to the BEI and the written test, and (c) the videotape format did not detract from game simulation performance. Reactions to videotape format did not account for significant differences in game simulation scores. Positive reactions to the quality of the videotape increased with participant experience.

Table 23

Intercorrelations Between Reaction Measures and Game Simulation Scores

	Pressbox	Oralinst	Gamesim	BEI	Wrtntest	Format
Pressbox	1.00					
Oralinst	.03	1.00				
Gamesim	.25*	.13	1.00			
BEI	.25*	.50**	.47**	1.00		
Wrtntest	.00	.44**	.37**	.52**	1.00	
Format	.16	.24	.52**	.33**	.33 **	1.00

Note: Abbreviations: Pressbox, game simulation score from press box level; Oralinst, Oral instructions post-measure; Gamesim; Game simulation post-measure; BEI, Behavioral event interview post-measure; Wrtntest, Written test post-measure; and Format, Videotape format post-measure

n = 61.

*p < .05. **p < .01.

IV. DISCUSSION

Prior to this research, very little was known about the effectiveness of videotape game simulations as work samples for testing sports officials. Similarly, no research had tested the effects of physical demand, experience, or camera angle on game simulation scores. Four major goals and 11 hypotheses were proposed to investigate the game simulation and other tests for use in testing of soccer officials as linesmen. This discussion focuses on each hypothesis, providing explanations for the results, integrating other research findings, and providing suggestions for future research.

The Relatedness of The Work Sample

Criterion-relatedness. The present research hypothesized that game simulation scores would be highly correlated with peer ratings and scores by expert assessors from observations of performance in actual games. Support for this hypothesis was found for the game simulation total score (r = .65) and the score of segments from the press box level camera angle (r = .54). Compared with the average coefficient for 18 performance measures (r = .37) and the average coefficient over all types of criteria (r = .41) reported by Reilly and Chao (1982), these correlation coefficients provided strong evidence of criterion-relatedness for peer ratings. Similarly, the correlation of assessment scores and game simulation scores from the press box level was impressive when compared with the average coefficient for supervisor ratings of job behaviors (r = .27) reported by Heneman (1986).

Wallace (1974) and Reilly and Chao (1982) raised the question as to what peer ratings actually measure. The question is relevant for this research since rater consistency was high (r = .93) and there was only one instance where a peer rater did not (i.e., elected not to) rate all 50 individuals. Thus, whatever was measured was done consistently across raters. The local expert panel suggested that peer ratings reflect long-term rater perceptions of ratee performance (Williams & Leavitt, 1947) based on firsthand (e.g., direct observation) and secondhand (e.g., stories and rumors) information. Since the raters used in this research were dispersed throughout the state, observations of performance occurred infrequently, and raters relied heavily on secondhand information. Davis (1973) reported that the flow of secondhand information, the "organizational grapevine", tends to be 75% to 95% accurate and travels very quickly, especially when the information is work-related, newsworthy, and consistent with perceptions and expectations. Because of this information, consistent perceptions of performance were to be expected.

Assessment scores, on the other hand, resulted from direct observations of game performance during a brief time period (i.e., less than 30 minutes) where the assessor had little or no past knowledge of the individual being assessed. Like ratings, assessments were composite scores; but unlike the peer raters used in this research, the assessors had been trained to arrive at their scores objectively by combining measures from multiple performance dimensions. Compared to peer ratings, assessment scores were "snapshots" of performance under very specific game conditions and were subject to considerable variation depending on the particular game conditions encountered.

The method used to assign assessors to fields prevented the investigation of the between-assessor effect; however, the evidence suggested that score differences were due to differences between games (i.e., assessment conditions) rather than differences between assessors. Future research in this area should clearly identify the sources of the between-assessor differences.

Similarly, additional research is needed to compare assessment scores and peer ratings. Such a comparison was not possible in this research. At the location where assessments were done (i.e., Site 1), participants did not know each other well enough to make peer ratings; and at the location were peer ratings were obtained (i.e., Site 2), assessment scores were not available. The high correlations of game simulation scores with rating and assessment scores suggest that future research should assess whether peer ratings, which were much easier to obtain than assessment scores or game simulation scores, are acceptable measures of past referee performance as suggested by Fiske and Cox (1960).

<u>**Construct-relatedness**</u>. Construct-relatedness involved the assessment of convergent validity, discriminant validity, and method bias. It was hypothesized that the intraclass correlation coefficients for convergent validity (i.e., Participant effect) and discriminant validity (i.e., Participant x Trait interaction) would be moderate to high and that the coefficient for method bias (i.e., Participant x Method interaction) would be low. The consistently low intraclass correlation coefficients for convergent and discriminant validity suggested that (1) the methods did not consistently order participants and (2) the ordering of participants was not significantly different for the job components of interest.

Game simulation total score and the game simulation score from the press box level exhibited moderate convergent validity when paired with the BEI. This result is encouraging and suggests that the game simulation from the press box level and the BEI could be used to provide consistent ordering of participants. Unfortunately, these same methods (i.e., those that displayed encouraging levels of convergent validity) also exhibited low discriminant validity, suggesting that their ordering of participants on Job Components B and C was redundant. In addition, inspection of Table 17 shows that including the BEI or the written test or both in the test battery consistently resulted in a finding of low discriminant validity, suggesting that neither the BEI nor the written test differentiated between Job Components B and C.

The finding of decreasing discriminant validity with experience for Job Components B and C was unexpected, especially the level of discriminant validity found in the high experience group (i.e., .04 to .06). One possible explanation for this finding is that referee performance increases (e.g., as shown in Tables 19 and 27) with experience reaching a ceiling of effectiveness, making the differences among referees negligible. Under these circumstances, low discriminant validity would be found.

High discriminant validity occurred in the low experience group when the two camera angles were used as methods. The finding of high discriminant validity here provides evidence that participant ordering for Job Components B and C was different in this group. As one possible explanation for this finding, the expert panel suggested that referees with low soccer experience tend to have little difficulty making decisions where judgment is not involved (i.e., Job Component B), but they tend to be far less competent at making consistent decisions when judgment is involved (i.e., Job Component C).

For the BEI and the written test, participants respond to verbal cues, and certain key words (e.g., deliberate, intentional, violent, or persistent) have specific meaning in the rules. When such words are recognized correctly, the correct response no longer involves judgment. For example, the word "violent" has a specific penalty associated with it. Once the judgment has been made that an act was violent, the referee's actions are specified in the rules. Since referees with low experience are knowledgeable about the written rules, they would be sensitive to verbal cues (i.e., key words) and would know the appropriate actions to take.

In comparison to making judgments that involve verbal cues, it is quite another task to view a videotape or an actual game and use visual cues to make judgments (Bandura, 1977). This research showed that experienced referees tend to be more accurate and more consistent in their responses based on visual cues (i.e., the press box game simulation). If referees with low experience have different perceptions about key words than do the more experienced referees, then, less experienced referees should make more errors and be less consistent where judgments are concerned. Thus, the finding of high discriminant validity for referees with low experience by the game simulation methods is encouraging and suggests that the game simulation could be used to identify the point at which referees begin to make consistent and correct judgments based on visual (i.e., job-relevant) cues.

For method bias, it was hypothesized that the methods would exhibit low method bias. Mixed results were found. As expected, method bias was low in the low and medium experience groups. Unexpectedly, method bias was moderate in the high experience group, suggesting that, for this group, the ordering of participants was significantly affected by the method used and that verbal and visual cues did not produce similar responses. The expert panel suggested that referees at this level know the rules well and could provide the "correct" response to either verbal or visual cues, but, in reality, the action required in the rules is not always used. Anecdotal evidence gathered during the BEI showed that participants with high experience frequently asked, "Do you want to know what the rules say I should do or do you want to know what I would really do?" The test administrator did not respond to that question; instead, the question was reread. Responses to the repeated question were mixed. Some participants gave the rule book answer (e.g., "The rule book says that I should . . ."); some said "I would . . ." Future research should investigate reasons for this difference in responses.

The method bias found in the high experience group when the two game simulation scores were used as methods was not expected. Low method bias was anticipated. Two possible explanations were offered by the expert panel. First, experienced officials tend to do fewer games as a linesman and more as the referee. Thus, they are less accustomed to viewing the game from the vantage point used in the field level game simulation. Instead, they tend to see the game from a broader perspective, more like that of the press box level. They also tend to have been in the college system longer and have viewed more games from the stands (i.e., the press box level) and more game films, nearly all of which have been made from the press box level. The field level, especially from the touch-line, is not their normal vantage point.

As another possible explanation for this finding, relationships among game simulation scores, age, and experience in Table 26 were examined. The relationship between age and experience group was not significant ($p > .05$), suggesting that the combination of playing, coaching and refereeing experience at the senior levels was not a function of age. This finding was explained by noting that the older participants had less playing and coaching experience at the senior levels. Their experience with senior level soccer was primarily referee experience. On the other hand, many young participants had considerably more playing and coaching experience at the senior levels. Thus, within each experience group, a similar age range was found.

In the low and medium experience groups, more than one-half of the participants were younger than age 40. In the high experience group, over one-half of the participants were older than age 40. The testing conditions (i.e., a 19" television set where participants stood about 4 to 6 feet from the screen) were such that differences in color vision and visual acuity, specifically the farsightedness normally associated with ages above 40, could have contributed to the method bias. A vision test was not included as part of the testing process and should be included in future research to control for acuity and color vision effects.

Game Simulation Variability

Physical demand. To assess whether participants must be tested under a condition of physical demand similar to game conditions, it was hypothesized that mean game simulation score for the high physical demand group would be significantly different than the mean score for the low demand group and that the differences could not be attributed to differences in physical ability. For the two demand groups, mean simulation scores were not significantly different ($p > .05$). Inspection of Table 26 shows that the physical performance scores were also not significantly different ($p > .05$).

One explanation for the absence of differences in the mean game simulation scores is that testing under high physical demand does not impact participant decision-making ability. Another explanation of this finding is that the high physical demand condition in this research was not sufficiently different from the low physical demand condition to produce score differences. This possibility is confounded by the lack of a clear definition of the high demand condition and the absence of a method to ensure that a high demand condition was achieved and maintained during testing. Additional research is needed to test the physical demand hypothesis. In future research, the high physical demand condition should involve a greater physical demand that is monitored during testing to control for its effects. For example, portable equipment could be used to monitor pulse rate during testing as has been done with soccer players (Astrand & Rodahl, 1986).

Performance level. The absence of a method to classify referees by grade or performance level prior to testing provided an opportunity in this research to investigate other measures of past referee performance. It was hypothesized that past referee experience should, with other things being equal, be a suitable substitute measure of referee performance. Analyses of variance of peer ratings and assessment scores showed that self-reported biographical data about past soccer experience was a significant source of variance and could be used as a substitute measure of past referee performance. Correlational data suggested that total senior level soccer experience was the experience measure most highly correlated with peer ratings and assessment scores. Senior level soccer experience was defined as the unweighted sum of the years as a player, coach, and referee at the senior level as reported on the biographical data blank. Additional research is needed to determine the most appropriate weighting among these three aspects of soccer experience. The same research could be used to identify the combinations of experience that tend to produce the "best" and "worst" referees. Such information would be invaluable in the selection and development of referees.

It was also hypothesized that game simulation scores could be predicted by a job-relevant measure of experience. In the analysis of variance, total senior level soccer experience accounted for a significant portion of the variance of the game simulation score from the press box level ($\underline{p} < .05$). Since the game simulation was prepared from videotapes of senior level games, the game simulation was able to identify referees with game experience at the level of the game from which the videotape was made and to assess performance at that game level.

This represents one of the most important opportunities for future research. Videotapes from different levels of youth and senior games could be used to test referees with varying amounts of youth and senior level experience. Referees should be expected to achieve high scores until they are placed in situations that requires skills and cue sensitivity above their past experience and present ability level (Bandua, 1977). The game simulation from the highest level of play before the referees' score declines could determine (1) the grade at which the referee should be assigned and (2) the level of training at which the referee should be placed.

<u>Camera angle</u>. A major goal of this research was to assess the importance of camera angle in the making of videotapes for the game simulation. It was hypothesized that game simulation scores from the field level camera angle would result in a significantly different ordering of participants than the press box camera angle. The expert panel felt that the field level camera angle would demonstrate higher external and internal validity, because it was considerably more like actual job conditions than the press box camera angle.

The findings of this research failed to support the camera angle hypothesis. As was expected, the correlation between the two game simulation scores was not significant ($p > .05$); moreover, the press box camera angle demonstrated greater external and internal validity than the field level camera angle. The expert panel suggested possible explanations for this finding related to the participants and the videotape. They felt that differences in the ability of participants to see the television screen clearly (i.e., visual acuity) and to distinguish between players and team colors (i.e., color blindness) could have been significant moderator variables of test scores. Additional research is needed to investigate the effects of participant vision and television screen size on game simulation score.

The expert panel suggested that the tendency to view games from the press box level (e.g., at a stadium or on the television) may increase with age and experience. If so, then the more experienced referees could be more comfortable with the press box camera angle. The expert panel agreed unanimously that the perspective of game events from the press box is different than that from field level and that most violations of the rules are easier to see at the press box level. Additional research is needed to permit a clearer understanding of the reasons for the validity of the press box game simulation.

Test Appropriateness and Acceptability

<u>Comparison of game simulation with other tests</u>. Once the relatedness of the game simulation was determined, a goal of this research was to identify the best possible combination of methods to use in the testing of soccer referees as linesmen. It was

hypothesized that more conventional tests could be substituted for the game simulation to reach the same decisions about participant ability to perform the job of linesman. Inspection of Tables 21, 22, and 26 shows that the relationships among testing methods and components of those methods were not strong and suggests that a combination of methods is appropriate for testing referees.

This research showed that a written test of Rule 5 (i.e., Fouls and Misconduct) was an effective tool for discriminating among referees in the total sample and at each experience level. Except for Rule 5, written test scores tended to have very low or negative correlations with game simulation scores. The expert panel suggested that performance in a game or in the game simulation does not depend on knowledge of all rules (e.g., the rules about the field, the ball, players and substitutes). Instead, the panel felt that the critical performance area was the recognition of and dealing with fouls and misconduct as the results suggested. The panel also suggested that knowledge of the "rule book" is most critical for new referees and that more experienced referees tend to study the rules less often. Therefore, a negative correlation between game simulation score and written test scores should have been expected, especially in the high experience group.

The potential usefulness of the BEI and game simulation measures (i.e., total and camera angle scores) was shown in Table 17 where moderate convergent validity was achieved in the total sample and in the high and medium experience groups. However, the low discriminant validity found when the BEI was included with the game simulation measures suggests that the BEI does not adequately distinguish between Job Components B and C. In contrast, the high discriminant validity found in the low experience group when only the two camera angles were used as methods suggests that the game simulation does adequately distinguish between Job Components B and C.

One possible explanation for these discriminant validity findings was provided by comparing responses to similar situations in different tests. Participant behavior during the game simulation was not always consistent with responses to BEI or written questions (e.g., participants consistently responded to BEI or written questions with a list of the correct procedures to be used following a “violent” act, but participant response to the visual cues of the game simulation was far less consistent, because of differences in the interpretation of what constituted a “violent” act.

This explanation was consistent with anecdotal reports from referee examiners of the Football Association in England that the BEI was best suited for use where the work situation could not be simulated (e.g., emergency situations). Additional research is needed to identify the best methods for testing each job task.

The usefulness of physical performance tests to referee testing was not clearly shown in this research except where the agility test entered the stepwise linear regression for the low experience group. One possible reason that the physical performance tests were not significant predictors of game simulation scores could be the testing method, because the testing method did not require participants to achieve and maintain a similar work rate during testing. Additional research is needed to assess game simulation performance under greater physical demand to determine the importance of the physical skills to the job of linesman.

Scoring scheme. Alba and Dickinson (1985) noted that a detailed scoring guide, including item weights, is important if consistency is to be achieved in the scoring of game simulation and BEI results. Considerable time and effort were required to gather job analysis data and determine the appropriate weights for test items. Although job analysis data suggested that aspects of the referee’s job are viewed to vary widely in importance, weighting of test items did not produce significant differences in participant ordering. Thus, until evidence suggests otherwise, future research should use unit item weights.

No previous research addressed the qualifications of scorers. The present research provided an opportunity to assess whether extensive soccer experience was needed to score participant responses. The results showed clearly that consistent game simulation and BEI scores were obtained by scorers who did not have extensive soccer experience.

Both the item weighting and scorer qualification findings have implications for the cost of soccer referee testing in the future.

Participant acceptance. Reactions to pilot testing of the pre-game oral instructions were unanimous that the live version should be used rather than the videotaped version. This finding was initially attributed to the fact that nearly all pre-game conferences are done live (i.e., without video or tape recordings). Preference of the live version was also consistent with findings that performance measurement systems are seen as more favorable (Dipboye & Pontbraind, 1981) and fairer and more accurate (Landy et al., 1978) when there is participation in the setting of goals and duties before measurement (Mount, 1983). Future research should investigate whether game simulation performance and satisfaction with the live pre-game instructions varies with the amount of two-way conversation during that meeting.

For the composite items for the three test methods, it was hypothesized that the game simulation would be most favorably received by participants. Although participant reaction to the game simulation item was favorable and significant ($p < .05$), the finding that the reactions to the BEI and the written test were both more favorable was not expected. The very favorable reaction to the BEI is encouraging and suggests that a BEI component could be added to the selection test battery with little or no resistance from referees.

The reasons for the less favorable reaction to the game simulation were not clear. Possible explanations include the fact that at Site 2 the test administrator could have been viewed as a threat (i.e., a peer who knew the "right" answers). If so, stress, anxiety and rigidity of response were likely (Staw, Sandelands, & Dutton, 1981). Less favorable reactions following testing are likely to be found when such a threat is perceived to exist.

The testing process was such that when errors were made they were readily seen by both the participant and the test administrator. Unlike the written test where poor performance can easily be blamed on a host of other external variables (e.g., poor wording, misread question, or incorrectly marked answer sheet), there were few legitimate reasons for "incorrect" actions in the game simulation. Thus, performance in a game simulation can be embarrassing to the participant when errors are made. Future research should focus on the determinants of satisfaction with performance on the game simulation.

It was hypothesized that the game simulation would be viewed favorably in terms of the way it simulates actual game conditions. It was also hypothesized that the videotape quality would not be perceived as detracting from performance on the game simulation. The overall reaction to the videotape format and the response to one item (i.e., item 4) were not significant ($p > .05$) in either direction, favorable or unfavorable. Responses to two of the four items that were used to form the composite videotape format item were significant ($p < .01$) and favorable (i.e., items 3 and 6). Responses to one of the other two format items were unfavorable and significant (i.e., item 5). These findings suggest that efforts should be expended in future research using videotapes to ensure the "realism" of the testing situation. For example, normal game sound could be used with the video.

Content-Related Strategy For Test Construction.

A major premise of this research was that tests of soccer referee performance could be developed using a content-related strategy. The procedures suggested by Alba and Dickinson (1985) were used in the construction of the game simulation and the BEI. Job analysis, scoring, and pilot testing were greatly enhanced by the availability and use of group input (e.g., the delphi technique). The evidence of criterion-related and construct-related validity suggested that the content-oriented strategy was used effectively for the construction of tests for soccer referees.

Summary

The combined evidence from this research suggests that a content-oriented strategy can be used to develop valid, reliable, and acceptable tests of linesman performance. Further, this research found that a videotape game simulation demonstrated high criterion-related validity with other measures of referee performance (i.e., ratings and assessment scores). This research also provided evidence that the testing of soccer officials for the job of linesman should contain elements from a combination of methods, including the game simulation, the BEI, a written test, and physical performance tests. Finally, this research provided evidence about four variables associated with the game simulation. Physical demand and testing site did not impact game simulation performance. Game simulation scores were impacted by total senior level soccer experience (i.e., the sum of playing, coaching, and refereeing experience). Performance increased with increasing total experience. Game simulation performance was also affected by the camera angle suggesting that the press box camera angle should be used for developing game simulation videotapes.

V. REFERENCES

AERA/APA/NCME Joint Committee. (1985). Standards for educational and psychologial testing. Washington, DC: American Psychological Association.

Alba, P. A., & Dickinson, T. L. (1985). Walk through performance testing documentation for jet engine mechanic (AFS426X2) (Job No. 120007000, Contract No. F33615-83-C-0030). San Antonio, TX: The Texas MAXIMA Corporation. [Submitted to Air Force Human Resources Laboratory]

Alba, P. A., & Wilcox, T. (1985). Walk through performance testing procedural guidelines manual (Task 0030-07, Contract No. F33615-83-C-0030-07). San Antonio, TX: The Texas MAXIMA Corporation. [Submitted to Air Force Human Resources Laboratory]

Alker, H. A., Straub, W. A., & Leary, J. (1973). Achieving consistency: A study of basketball officials. Journal of Vocational Behavior, 3, 335-343.

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monerey, CA: Brooks/Cole.

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work-sample tests: A review. Personnel Psychology, 27, 519-523.

Astrand, P.-O., & Rodahl, K. (1986). Textbook of work physiology (3rd ed.). New York: McGraw-Hill.

Astrand, P.-O., & Christensen, E. H. (1964). Aerobic work capacity. In F. Dickens, E. Neil, & W. F. Widdans (Eds.), Oxygen in the animal organism (p. 295). New York: Pergamon Press.

Bandura, A. (1977). Social learning theory. Englewood Cliffs, NJ: Prentice-Hall.

Borman, W. C. (1982). Behavior-based rating scales. In R. A. Beck (Ed.), Performance assessment: Methods & applications (pp. 100-120). Baltimore, MD: The Johns Hopkins University Press.

Boruch, R. F., Larkin, J. D., Wolins, L., & MacKinney, A. C. (1970). Alternate methods of analysis: Multitrait-multimethod data. Educational and Psychological Measurement, 30, 833-853.

Brodie, D. A. (1981, February). Work analysis of football league referees. Leeds, UK: Carnegie School, Leeds Polytechnic [A Consultancy Commissioned by The Football Association and The Football League].

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

Cascio, W. F. (1987). Applied psychology in personnel management (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Cronbach, L. J. (1960). Essentials of psychological testing (2nd ed.). New York: Harper and Row.

Cox, R. A. (1985). Sport psychology: Concepts and applications. Dubuque, IA: Wm. C. Brown.

Dalkey, N. C., & Helmer, O. (1963). An experimental application of the delphi method to the use of experts. Managerial Science, 9, 458-467.

Davis, K. (1973, July). The care and cultivation of the corporate grapevine. Dun's Review, pp. 44-47.

Dickinson, T. L. (1987). Designs for evaluating the validity and accuracy of performance ratings. Organizational Behavior and Human Performance, 40, 1-21.

Dickinson, T. L., Hassett, C. E., & Tannenbaum, S. I. (1986). Work performance ratings: A meta-analysis of multitrait-multimethod studies (AFHRL-TP-86-32, AD-A174 759). Brooks AFB, TX: Training Systems Division, Brooks Air Force Human Resources Laboratory.

Dipboye, R. L., & Pontbraind, R. (1981). Correlates of employee reactions to performance appraisals and appraisal systems. Journal of Applied Psychology, 66, 248-251.

Fiske, D. W., & Cox, J. A. Jr. (1960). The consistency of ratings by peers. Journal of Applied Psychology, 44, 11-17.

Ford, J. K., & Wroten, S. P. (1984). Introducing new methods for conducting training evaluation to program redesign. Personnel Psychology, 37, 651-666.

Fratzke, M. R. (1975). Personality and biographical traits of superior and average college basketball officials. Research Quarterly, 46, 484-488.

Fusfeld, A. R., & Foster, R. J. (1971). The delphi technique: Survey and comment. Business Horizons, 14, 63-74.

Grimby, G., & Saltin, B. (1983). The ageing muscle. Clinical Physiology, 3, 209.

Guion, R. M. (1978). "Content validity" in moderation. Personnel Psychology, 31, 205-213.

Heneman, R. L. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. Personnel Psychology, 39, 811-826.

Hanin, Y. L. (1980). Applying sport psychology: Past, present and future. In C. H. Nadean, W. R. Halliwell, K. M. Newell, & G. C. Roberts (Eds.), Psychology of motor behavior in sport (pp 37-48). Urbana, IL: Human Kinetics.

Hunter, J. E., & Hunter, R. F. (1984). The validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72-98.

Kane, J. S., & Lawler, E. E., III. (1978). Methods of peer assessment. Psychological Bulletin, 85, 555-586.

Kavanaugh, M. J., Borman, W. C., Hedge, J. W., & Gould, R. B. (1986). Job performance classification scheme for validation research in the military (AFHRL-TP-85-51, AD-A164 837). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Kavanaugh, M. J., McKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analysis of ratings. Psychological Bulletin, 75, 34-49.

Klimoski, R. J., & London, M. (1974). Role of the rater in performance appraisal. Journal of Applied Psychology, 59, 445-451.

Kuhnle, R. L., & Yarbrough, R. E. (1986). Physical performance testing of college soccer referees. Unpublished manuscript.

Landy, F. J., Barnes, J. L., & Murphy, K. R. (1978). Correlates of perceived fairness and accuracy of performance evaluation. Journal of Applied Psychology, 63, 751-754.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. Journal of Applied Psychology, 65, 422-427.

Lawler, E. E., III (1967). The multitrait-multirater approach to measuring managerial job performance. Journal of Applied Psychology, 51, 369-381.

Lawshe, C. H. (1975). A quantitative approach to content validity. Personnel Psychology, 28, 563-565.

Levine, E. L. (1983). Everything you always wanted to know about job analysis and more! . . . A job analysis primer). Tampa, FL: Mariner Publishing Company.

Lewin, A. Y., & Zwany, A. (1976). Peer nominations: A model, literature critique, and a paradigm for research. Personnel Psychology, 29, 423-447.

Londeree, B. R., & Moeschberger, M. L. (1982). Effect of age and other factors on maximal heart rate. Research Quarterly, 53, 297-304.

Martino, J. P. (1983). Technological forecasting for decision making (2nd ed.). New York: North-Holland.

McElvoy, G. M., & Buller, P. F. (1987). User acceptance of peer appraisals in an industrial setting. Personnel Psychology, 40, 785-797.

Mount, M. (1983). Comparison of managerial and employee satisfaction with a performance appraisal system. Personnel Psychology, 36, 99-110.

Morgan, W. P. (1980). The trait psychology controversy. Research Quarterly, 51, 324-339.

Morris, D. (1981). The soccer tribe. London: Johathan Cape Ltd.

Nathan, B. R., & Alexander, R. A. (1988). A comparison of criteria for test validation: A meta-analytic investigation. Personnel Psychology, 41, 517-536.

Ronan, W. W., & Prien, E. P. (1966). Toward a criterion theory: A review and analysis of research and opinion. Greensboro, NC: The Richardson Foundation.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative selection procedures. Personnel Psychology, 35, 1-62.

Robinson, S. (1938). Experimental studies of physical fitness in relation to age, Arbeitsphysiology, 10, 251.

Sackett, P. R. (1987). Assessment centers and content validity: Some neglected issues. Personnel Psychology, 40, 13-26.

Schlenker, B. R. (1975). Self-presentation: Managing the impression of consistency when reality interferes with self-enhancement. Journal of Personality and Social Psychology, 32, 1030-1037.

Schmitt, F. L., Gooding, R. Z., Noe, R. A., & Kirsh, M. (1984). Metaanalysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 69, 407-422.

Schmitt, F. L., & Hunter, J. F. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Schmitt, F. L., & Hunter, J. F. (1984). A within setting empirical test of the situational specificity hypothesis in personnel selection. Personnel Psychology, 37, 317-326.

Schmitt, N., & Ostroff, C. (1986). Operationalizing the 'behavioral consistency' approach: Selection test development based on a content-oriented strategy. Personnel Psychology, 39, 91-109.

Schurr, E. L., & Phillips, J. A. (1971). Women sports officials. Journal of Health Physical Education and Recreation, 42, 71-72.

Sherwood, J. J. (1966). Self-report and projective measures of achievement and affiliation. Journal of Consulting Psychology, 30, 329-334.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

Sorenson, A. G. (1956). A note of the “fakability” of the Minnesota Teacher Attitude Inventory. Journal of Applied Psychology, 40, 192-194.

Staw, B. M., Sandelands, L. E., & Dutton, J. E. (1981). Threat-rigidity effects in organizational behavior: A multilevel analysis. Administrative Science Quarterly, 26, 501-524.

Vaughn, G. M., & Corballis, M. C. (1969). Beyond tests of significance: Estimating strength of effects in selecting ANOVA designs. Psychological Bulletin, 72, 204-213.

Vikhrov, K. (1978). Factors and pedagogical means for measuring the quality of officiating of soccer contests. Unpublished doctoral dissertation. Soviet Sports Center, Kiev. (Abstract translated 1987).

Wallace, S. R. (1974). How high the validity? Personnel Psychology, 27, 397-407.

Williams, S. B., & Leavitt, H. J. (1947). Group opinion as a predictor of military leadership. Journal of Applied Psychology, 11, 283-291.

Winer, B. J. (1971). Statistical principles in experiemental design (2nd ed.). New York: McGraw-Hill.

Zammuto, R. F., London, M., & Rowland, K. M. (1982). Organization and rater differences in performance appraisals. Personnel Psychology, 35, 643-658.

VI. APPENDIX A

TUKEY (HSD) RESULTS

FOR ASSESSOR SCORES

AND PEER RATINGS

Table 24

Tukey (HSD) Test Results For Assessor Scores

	Assessments		
Assessor	n	M	SD
1	10	74.80	10.27
2	8	65.19	12.10
3	7	77.43	5.13
4	8	83.13	6.83

Analysis of Variance Summary

Source	df	MS	F-ratio
Between Assessors	3	446.68	5.27**
Within Assessors	29	84.76	

Matrix of Pairwise Mean Differences

	Assessor			
Assessor	1	2	3	4
1	.00			
2	9.61	.00		
3	2.63	12.24*	.00	
4	8.33	17.94**	5.70	.00

*p < .05. **p < .01.

Table 25

Tukey (HSD) Test Results For Peer Ratings

	Ratings		
Experience	n	M	SD
Low	14	66.00	6.18
Med	10	68.60	3.53
High	12	72.17	5.04

Analysis of Variance Summary

Source	df	MS	F-ratio
Between Groups	2	123.08	4.57 *
Within Groups	33	26.91	

Matrix of Pairwise Mean Differences

	Experience Group		
	Low	Medium	High
Low	.00		
Medium	2.60	.00	
High	6.17**	3.57	.00

*p < .05. **p < .01.

VII. APPENDIX B

PEARSON PRODUCT MOMENT

CORRELATION MATRIX

Table 26

Intercorrelations of Tests, Test Components, and Study Variables For Total Sample

	Fieldlvl	Pressbox	GS Total	GSim(A)	GSim(B)	GSim(C)	BEI Total
Fieldlvl	1.00						
Pressbox	.19	1.00					
GS Total	.61	.88	1.00				
GSim(A)	.35	.44	.62	1.00			
GSim(B)	.49	.44	.58	.25	1.00		
GSim(C)	.48	.85	.90	.49	.19	1.00	
BEI Total	.32	.36	.44	.28	.37	.34	1.00
BEI(A)	.24	.26	.33	.23	.25	.26	.59
BEI(B)	.18	.26	.31	.22	.34	.19	.60
BEI(C)	.28	.30	.37	.21	.29	.30	.94
WT Total	.04	- .05	- .04	- .13	.02	- .04	.21
WT(A)	.10	- .24	- .18	- .25	- .03	- .17	- .33
WT(B)	.04	- .01	.02	.00	.13	- .05	.30
WT(C)	.02	.12	.10	.02	- .08	.18	.21
Rule(1)	.07	- .16	- .12	- .16	.05	- .15	.04
Rule(2)	.04	- .19	- .14	- .14	- .16	- .08	- .01
Rule(3)	.08	- .16	- .12	- .22	.09	- .16	- .09
Rule(4)	- .03	- .01	- .04	- .15	.14	- .09	.36
Rule(5)	- .10	.37	.27	.24	.03	.31	.21
Rule(6)	.09	- .09	- .03	- .06	- .07	.01	.15
Fit Total	.16	.14	.17	- .01	.14	.15	.13
Fit(1)	.06	.11	.11	- .01	.05	.11	.06
Fit(2)	.15	.09	.13	- .06	.17	.04	.06

Table 26 (Continued)

	Field Lvl	Press Box	Total GS	GSim(A)	GSim(B)	GSim(C)	BEI Total
Fit(3)	.19	.36	.36	.16	.26	- .23	.16
Fit(4)	.17	.08	.12	- .01	.10	.11	.19
SrLvlExp	.08	.51	.46	.29	.19	- .19	.16
Tri-SLExp	.16	.41	.41	.23	.22	- .34	.19
Demand	- .08	.12	.07	.11	- .13	.16	.05
Site	.09	.30	.33	.44	.23	- .51	- .05
Age	- .31	.03	- .09	.14	- .28	- .08	- .18
Sex	.15	- .17	- .04	.17	.05	- .05	.16

	BEI(A)	BEI(B)	BEI(C)	WT Total	WT(A)	WT(B)	WT(C)
BEI(A)	1.00						
BEI(B)	.08	1.00					
BEI(C)	.36	.45	1.00				
WT Total	.19	- .09	.24	1.00			
WT(A)	- .02	- .18	.03	.77	1.00		
WT(B)	.18	.09	.31	.67	.32	1.00	
WT(C)	.26	- .11	.22	.67	.21	.23	1.00
Rule(1)	.04	- .035	.05	.55	.78	.16	.12
Rule(2)	.06	- .20	.02	.75	.81	.36	.38
Rule(3)	- .23	- .09	.01	.40	.50	.26	.04
Rule(4)	.20	.18	.36	.79	.47	.76	.48
Rule(5)	.37	- .12	.18	.36	- .08	.09	.74
Rule(6)	.14	- .09	.19	.66	.24	.70	.54

Table 26 (Continued)

	BEI(A)	BEI(B)	BEI(C)	WT Total	WT(A)	WT(B)	WT(C)
Fit Total	.14	- .03	.13	.05	.14	- .07	- .03
Fit(1)	.12	- .15	.08	.11	.18	- .02	- .01
Fit(2)	.10	- .05	.06	.04	.09	- .06	- .00
Fit(3)	.14	.09	.14	- .23	- .11	- .22	- .16
Fit(4)	.13	.05	.19	.11	.22	- .04	- .01
SrLvlExp	.01	.18	.15	- .19	- .16	- .10	- .15
Tri-SLExp	.06	.25	.15	- .34	- .34	- .12	- .23
Demand	.12	- .13	.06	.16	.14	- .07	.26
Site	.03	.12	- .13	- .51	- .51	- .28	- .27
Age	- .27	- .03	- .13	- .08	- .06	- .01	- .08
Sex	- .09	.36	.14	- .05	- .12	- .08	.05

	Rule(1)	Rule(2)	Rule(3)	Rule(4)	Rule(5)	Rule(6)	Fit Total
Rule(1)	1.00						
Rule(2)	.41	1.00					
Rule(3)	.03	.36	1.00				
Rule(4)	.28	.46	.34	1.00			
Rule(5)	- .03	.12	- .27	.18	1.00		
Rule(6)	.10	.34	.18	.50	.16	1.00	
Fit Total	.26	- .08	- .02	.05	.01	- .11	1.00
Fit(1)	.23	- .01	.06	.06	.05	- .05	.89
Fit(2)	.16	- .10	- .02	.07	.02	- .05	.87
Fit(3)	.16	- .32	- .22	- .19	.00	- .31	.74

Table 26 (Continued)

	Rule(1)	Rule(2)	Rule(3)	Rule(4)	Rule(5)	Rule(6)	Fit Total
Fit(4)	.32	.01	.02	.12	- .05	- .08	.95
SrLvlExp	- .06	- .21	- .10	- .12	- .05	- .14	.15
Tri-SLExp	- .23	- .40	- .10	- .14	- .11	- .21	- .10
Demand	.00	.22	.07	.04	.30	- .07	.02
Site	- .29	- .48	- .33	- .50	.02	- .23	- .11
Age	- .20	.06	.15	- .12	.06	- .14	- .46
Sex	- .09	.00	- .15	- .02	- .04	.11	- .54

	Fit(1)	Fit(2)	Fit(3)	Fit(4)	SrLvlExp	Tri-SLExp	
Fit(1)	1.00						
Fit(2)	.65	1.00					
Fit(3)	.58	.54	1.00				
Fit(4)	.82	.77	.68	1.00			
SrLvlExp	.07	.12	.30	.12	1.00		
Tri-SLExp	.01	.07	.26	.10	.83	1.00	
Demand	.06	- .05	- .00	.05	- .14	- .08	1.00
Site	- .03	- .14	.15	- .21	.29	.24	- .02
Age	- .36	- .50	- .27	- .42	.19	.12	.13
Sex	- .48	- .49	- .45	- .48	- .16	- .12	- .06

Table 26 (Concluded)

	Site	Age	Sex
Site	1.00		
Age	.09	1.00	
Sex	.17	-0.02	1.00

Note: Abbreviations: Fieldlvl, Field level game simulation score; Pressbox, Press box game simulation score; GS Total, Game simulation total score; GSim(A), Game simulation score for Job Component A; GSim(B), Game simululation score for job component B; GSim(C), Game simulation score for job component C; BEI Total, BEI total score; BEI(A), BEI score for job component A; BEI(B), BEI score for job component B; BEI(C), BEI score for job component C; WT Total, Written test total score; WT(A), Written test score for job component A; WT(B), Written test score for job component B; WT(C), Written test score for job component C; Rule(1), Written test score for rule 1; Rule(2), Written test score for rule 2; Rule(3), Written test score for rule 3; Rule(4), Written test score for rule 4; Rule(5), Written test score for rule 5; Rule(6), Written test score for rule 6; Fit Total, Standardized total score from physical performance test battery; Fit(1), Standardized score for test 1 (Aerobic endurance run); Fit(2), Standardized score for test 2 (Sprint); Fit(3), Standardized score for test 3 (Agility run); Fit(4), Standardized score for test 4 (Anaerobic endurance run); SrLvlExp, Senior level soccer experience in years; Tri-SLExp, Senior level experience as a trichotomous variable; Demand, Physical demand condition (Low or High); Site, Testing location (Site 1 or 2); Age, Participant age in years; and Sex, Participant sex.

$\underline{n} = 61$.

All correlations greater than 0.25 are significant at the $\underline{p} < .05$ level. All correlations greater than 0.33 are significant at the $\underline{p} < .01$ level.

VIII. APPENDIX C

POST QUESTIONNAIRE

PARTICIPANT QUESTIONNAIRE

Instructions: Read each statement. Then circle the number that MOST ACCURATELY captures your response.

Use the following scale:

1 - Almost NEVER

2 - To a VERY LITTLE extent

3 - To a LITTLE extent

4 - Somewhat

5 - To a GREAT extent

6 - To a VERY GREAT extent

7 - Almost ALWAYS

"TO WHAT EXTENT ....."

1. ... were the pre-game instructions realistic for college soccer.
2. ... did the videotape scenes represent realistic game situations for college soccer.
3. ... did the videotape simulate actual game conditions.
4. ... did the videotape events flow as might be expected in an actual game.
5. ... did the videotape capture the emotion and pressure of an actual game.
6. ... did the quality of the videotapes interfere with your ability to make correct decisions.
7. ... did the videotape test your ability as a linesman.
8. ... did the oral interview ask realistic questions about linesman duties.
9. ... did the oral interview test your ability as a linesman.
10. ... did the written examination test your knowledge of the rules as they apply to linesmen.
11. ... did the written test ask realistic questions about linesman duties.

IX. APPENDIX D

POST QUESTIONNAIRE RESPONSES

Table 27

Post-Questionnaire Responses

	Response								
Item	1	2	3	4	5	6	7	M	SD
1	0	1	2	16	23	11	8	5.76	1.12
2	1	1	2	8	15	17	17	5.35	1.35
3	2	2	11	9	14	12	11	4.80	1.62
4	4	6	9	19	11	8	4	4.10	1.57
5	11	19	12	9	7	3	0	2.82	1.45
6	7	11	12	20	9	2	0	3.31	1.34
7	0	2	7	16	16	15	5	4.82	1.26
8	0	0	1	5	19	23	13	5.69	0.96
9	0	0	3	11	16	24	7	5.33	1.06
10	0	1	0	8	24	19	9	5.43	1.01
11	0	0	4	12	21	14	10	5.23	1.16
GS								10.17	2.17
BEI								11.03	1.85
WT								10.66	1.99

Note: Abbreviations: GS, Game simulation post-measure; BEI, Behavioral event interview post-measure; and WT, Written test post-measure.

n = 61.

X. APPENDIX E

MULTITRAIT-MULTIMETHOD

ANALYSIS OF VARIANCE

SUMMARY TABLES

Table 28

Multitrait-Multimethod Analysis Of Game Simulation and BEI Methods For Job Components A, B, and C In the Total Sample

Source	df	MS	F-ratio	VC	ICC
Participant (P)	60	2.41	3.69*	.29	.27
P X Trait	120	.76	1.16	.05	.05
P X Method	60	.86	1.31*	.07	.06
Error	120	.65		.65	

n = 61.

*p < .01.

Table 29

Multitrait-Multimethod Analysis Of Game Simulation, BEI, and Written Test Methods For Job Components A, B, and C In the Total Sample

Source	df	MS	F-ratio	VC	ICC
Participant (P)	60	2.24	3.33*	.15	.16
P X Trait	120	.82	1.20	.05	.04
P X Method	120	1.28	1.89*	.20	.18
Error	240	.68		.68	

n = 61.

*p < .01.

Table 30

Multitrait-Multimethod Analysis Of Field Level Game Simulation and Press Box Game Simulation Methods For Job Components B and C In the Total Sample

Source	df	MS	F-ratio	VC	ICC
Participant (P)	60	1.41	2.42*	.21	.17
P X Trait	60	1.13	1.95*	.28	.22
P X Method	60	.93	1.60*	.17	.14
Error	60	.58		.58	

n = 61.

*p < .05.

Table 31

Multitrait-Multimethod Analysis Of Press Box Game Simulation and BEI Methods For Job Components B and C In the Total Sample

Source	df	MS	F-ratio	VC	ICC
Participant (P)	60	1.73	2.88	.28	.24
P X Trait	60	.84	1.40	.12	.17
P X Method	60	.90	1.51	.15	.13
Error	60	.60		.60	

n = 61.

Table 32

Multitrait-Multimethod Analysis of Field Level Game Simulation, Press Box Game Simulation, and BEI Methods For Job Components B and C In the Total Sample

Source	df	MS	F-ratio	VC	ICC
Participant (P)	60	1.93	3.15*	.22	.19
P X Trait	60	1.05	1.71*	.15	.13
P X Method	120	.94	1.54*	.17	.15
Error	120	.61		.61	

n = 61.

*p < .01.

Table 33

Multitrait-Multimethod Analysis of Field Level Game Simulation, Press Box Game Simulation, BEI, and Written Test Methods For Job Components B and C In the Total Sample

Source	df	MS	F-ratio	VC	ICC
Participant (P)	60	1.97	3.16*	.17	.15
P X Trait	60	1.19	1.91*	.14	.13
P X Method	180	1.03	1.66*	.20	.18
Error 180	.62		.62		

n = 61.

*p < .01.

Table 34

Multitrait-Multimethod Analysis of Field Level Game Simulation and Press Box Game Simulation Methods For Job Components B and C In the High Experience Group

Source	df	MS	F-Ratio	VC	ICC
Participant (P)	18	1.37	1.94	.17	.13
P X Trait	18	.81	1.15	.05	.04
P X Method	18	1.34	1.90	.32	.26
Error	18	.71		.71	

n = 19.

Table 35

Multitrait-Multimethod Analysis of Press Box Game Simulation and BEI Methods For Job Components B and C In the High Experience Group

Source	df	MS	F-ratio	VC	ICC
Participant (P)	18	1.90	3.55*	.34	.28
P X Trait	18	.67	1.26	.07	.06
P X Method	18	1.11	2.08*	.29	.23
Error	18	.54		.54	

n = 19.

*p < .01.

Table 36

Multitrait-Multimethod Analysis of Field Level Game Simulation, Press Box Game Simulation and BEI Methods ForWith Job Components B and C In the High Experience Group

Source	df	MS	F-Ratio	VC	ICC
Participant (P)	18	1.82	3.01*	.20	.17
P X Trait	18	.79	1.31	.06	.05
P X Method	36	1.25	2.06*	.35	.27
Error	36	.61		.61	

n = 19.

*p < .01.

Table 37

Multitrait-Multimethod Analysis Of Field Level Game Simulation and Press Box Game Simulation Methods For Job Components B and C In the Medium Experience Group

Source	df	MS	F-ratio	VC	ICC
Participant (P)	18	1.53	2.15	.21	.17
P X Method	18	1.02	1.44	.16	.13
P X Trait	18	.96	1.35	.12	.10
Error	18	.71		.71	

n = 19.

Table 38

Multitrait-Multimethod Analysis of Press Box Game Simulation and BEI Methods For Job Components B and C In the Medium Experience Group

Source	df	MS	F-ratio	VC	ICC
Participant (P)	18	1.77	3.27 *	.31	.24
P X Trait	18	.90	1.67	.18	.14
P X Method	18	1.00	1.85 *	.23	.18
Error	18	.54		.54	

n = 19.

*p < .01.

Table 39

Multitrait-Multimethod Analysis of Field Level Game Simulation, Press Box Game Simulation, and BEI Methods For Job Components B and C In the Medium Experience Group

Source	df	MS	F-ratio	VC	ICC
Participant (P)	18	2.19	3.68*	.27	.22
P X Trait	18	1.10	1.85	.17	.14
P X Method	36	.93	1.56*	.17	.14
Error	36	.60		.60	

n = 19.

*p < .01.

Table 40

Multitrait-Multimethod Analysis of Field Level Game Simulation and Press Box Game Simulation Methods For Job Components B and C In The Low Experience Group

Source	df	MS	F-ratio	VC	ICC
Participant (P)	22	1.41	2.84*	.23	.17
P X Trait	22	1.67	3.38*	.59	.43
P X Method	22	.61	1.24	.06	.04
Error	22	.49		.49	

n = 23.

*p < .01.

Table 41

Multitrait-Multimethod Analysis of Press Box Game Simulation and BEI Methods For Job Components B and C In The Low Experience Group

Source	df	MS	F-Ratio	VC	ICC
Participant (P)	22	1.43	1.80	.16	.14
P X Trait	22	1.13	1.42	.17	.15
P X Method	22	.82	1.03	.01	.01
Error	22	.80		.80	

n = 23.

Table 42

Multitrait-Multimethod Analysis of Field Level Game Simulation, Press Box Game Simulation, and BEI Methods For Job Component B and C In the Low Experience Group

Source	df	MS	F-Ratio	VC	ICC
Participant (P)	22	1.69	2.31**	.16	.14
P X Trait	22	1.35	1.85*	.21	.18
P X Method	44	.88	1.21**	.08	.07
Error	44	.73		.73	

n = 23.

*p < .05. **p < .01.

XI. APPENDIX F

ANALYSIS OF VARIANCE

SUMMARY TABLES FOR

POST QUESTIONNAIRE ITEMS

Table 43

Analysis of Variance Of Oral Instructions Item From Post Questionnaire

Source	df	MS	F-ratio
Demand (D)	1	.22	.18
Experience (E)	2	.09	.08
D x E	2	1.32	1.08
Site (S)	1	4.74	3.89 *
D x S	1	.87	.72
E x S	2	1.60	1.32
D x E x S	2	1.32	1.08
Age (A)	1	.23	.19
Error	48	1.22	

n = 61.

* p < .06.

Table 44

Analysis of Variance Of Game Simulation Item From Post Questionnaire

Source	df	MS	F-ratio
Demand (D)	1	2.02	.42
Experience (E)	2	4.47	.93
D x E	2	.98	.20
Site (S)	1	.01	.00
D x S	1	.83	.17
E x S	2	11.40	2.38
D x E x S	2	6.36	1.32
Age (A)	1	1.76	.37
Error	48	4.80	

n = 61.

Table 45

Analysis of Variance Of BEI Item From Post Questionnaire

Source	df	MS	F-ratio
Demand (D)	1	.93	.30
Experience (E)	2	4.77	1.52
D x E	2	.30	.10
Site (S)	1	2.45	0.78
D x S	1	.66	.21
E x S	2	8.44	2.69
D x E x S	2	7.12	2.27
Age (A)	1	.04	.01
Error	48	3.14	

n = 61.

Table 46

Analysis of Variance Of Written Test Item From Post Questionnaire

Source	df	MS	F-ratio
Demand (D)	1	.00	.01
Experience (E)	2	1.01	.25
D x E	2	3.32	.82
Site (S)	1	4.12	1.02
D x S	1	4.06	1.01
E x S	2	3.72	.92
D x E x S	2	5.71	1.42
Age (A)	1	2.65	.66
Error	48	4.03	

n = 61.

Table 47

Analysis of Variance Of Videotape Format Test Item From Post Questionnaire

Source	df	MS	F-ratio
Demand (D)	1	25.09	1.99
Experience (E)	2	4.30	.34
D x E	2	38.91	3.09 *
Site (S)	1	3.08	.25
D x S	1	.01	.00
E x S	2	24.68	1.96
D x E x S	2	35.23	2.80
Age (A)	1	1.42	.11
Error	48	12.60	

$\underline{n} = 61$.

* $\underline{p} < .06$.