This is to certify that the

dissertation entitled

An Investigation of One Alternative to the Group-process Format for Setting Performance Standards on a Medical Specialty Examination

presented by

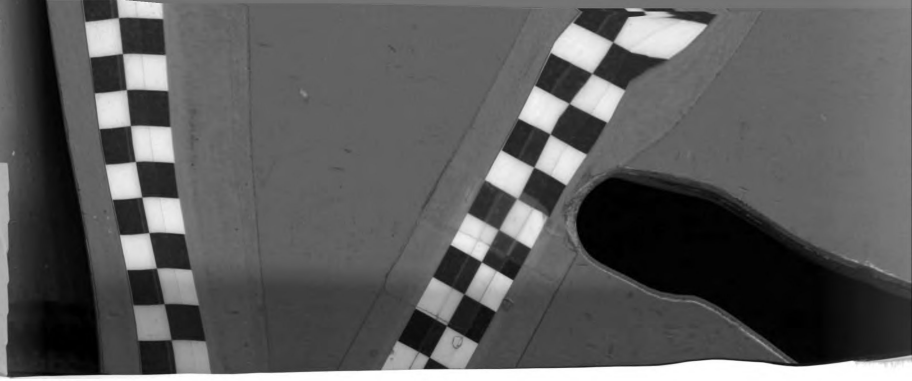
Gregory J. Cizek

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Measurement,
Evaluation, and Research Design

Major professor

Date 2/22/91



AN INVESTIGATION INTO ONE ALTERNATIVE TO THE GROUP-PROCESS PROCEDURE FOR SETTING PERFORMANCE STANDARDS ON A MEDICAL SPECIALTY EXAMINATION

By

Gregory J. Cizek

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology,
and Special Education

1991

AN INVESTIGATION INTO ONE ALTERNATIVE TO THE GROUP-PROCESS PROCEDURE FOR SETTING PERFORMANCE STANDARDS ON A MEDICAL SPECIALTY EXAMINATION

By

Gregory J. Olsen

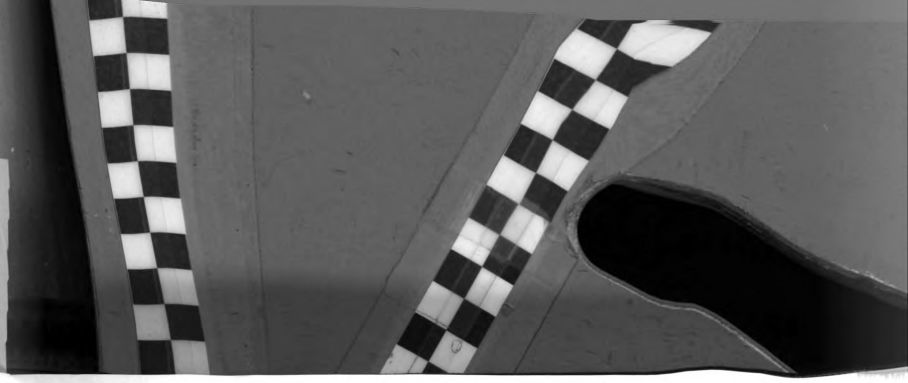
A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology,
and Special Education

1991



ABSTRACT

AN INVESTIGATION INTO ONE ALTERNATIVE TO THE GROUP-PROCESS PROCEDURE FOR SETTING PERFORMANCE STANDARDS ON A MEDICAL SPECIALTY EXAMINATION

By

Gregory J. Cizek

This study examined one variation of the traditional group-process procedure for establishing passing standards on a medical specialty examination using the Angoff methodology. The variation consisted of requiring subject-matter experts to provide Angoff ratings independently, without group interaction or other sources of information. The study also sought to isolate the effect of group interaction and information-sharing through comparison to a group-process condition and a condition in which independent item reviewers were provided with distributions of other the independent reviewers' ratings.

There were several major finding in the study. It was observed that the independent procedure produced a nonsignificantly higher passing standard than the group-process procedure did. The absence of statistical significance, however, did not exclude large practical consequences for the interested groups, such as the examinees and the standard setting board. These practical consequences are described and discussed. Also, it was observed that individual item reviewers' ratings were more variable in the independent condition compared to the group-process procedure. The independent condition was also less

ABSTRACT

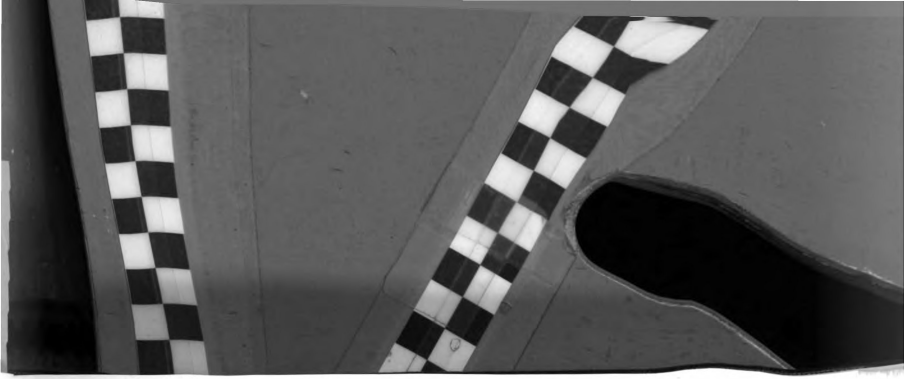
AN INVESTIGATION INTO ONE ALTERNATIVE TO THE GROUP-PROCESS
PROCEDURE FOR RESTORING EFFECTIVENESS DIMINISHED BY A
MEDICAL SPECIALTY EXAMINATION

By

Gregory J. Cizek

This study examined one variation of the traditional group-process procedure for rehabilitating passing standards on a medical specialty examination using the Angoff methodology. The variation correlated of requiring subject-matter experts to provide Angoff ratings independently, without group interaction or other sources of information. The study also sought to isolate the effect of group-process condition and a condition in which independent item reviewers were provided with distributions of other the independent reviewers' ratings.

There were several major finding in the study. It was observed that the independent procedure produced a substantially higher passing standard than the group-process procedure did. The absence of statistical significance, however, did not exclude large practical consequences for the interested groups, such as the examinee and the standard setting board. These practical consequences are described and discussed. Also, it was observed that individual item reviewers' ratings were more variable in the independent condition compared to the group-process procedure. The independent condition was also less



costly to implement. Item reviewers in both conditions produced ratings that exhibited less than desirable accuracy in terms of estimating the performance of the hypothetical minimally-competent group.

The provision of additional information to the independent group in the form of distributions of their own initial item ratings resulted in subsequent ratings that were significantly higher and less variable, but did not result in more precise estimates of performance for the minimally competent group. However, independent raters apparently utilized the additional information provided as distributions of ratings. It was found that knowledge of a reviewer's initial rating and the group's initial mean item rating was a moderately good predictor of a reviewer's subsequent ratings.

Implications for future design of standard setting procedures and policy considerations are discussed.

costly to implement. Item reviewers in both conditions produced ratings that exhibited less than desirable accuracy in terms of estimating the performance of the hypothetical minimally-competent group.

The provision of additional information to the independent group in the form of distributions of their own initial item ratings resulted in subsequent ratings that were significantly higher and less variable, but did not result in more precise estimates of performance for the minimally competent group. However, independent raters apparently utilized the additional information provided as distributions of ratings. It was found that knowledge of a reviewer's initial rating and the group's initial mean item rating was a moderately good predictor of a reviewer's subsequent ratings.

Implications for future design of standard setting procedures and policy considerations are discussed.

ACKNOWLEDGEMENTS

I appreciate the patience and encouragement of those who have helped me with this project: Stephen Raudenbush, Irvin Lehmann, Diana Pullin, William Mehrens, David Labaree, and Stephen Yelon of Michigan State University.

I am most grateful to my wife, Rita, and our children, Caroline, David, and Stephen for their unfailing love and support, and to my parents for their enduring confidence.

I thank God for his blessings, as surely evidenced to me through these people who have given so much.

TABLE OF CONTENTS

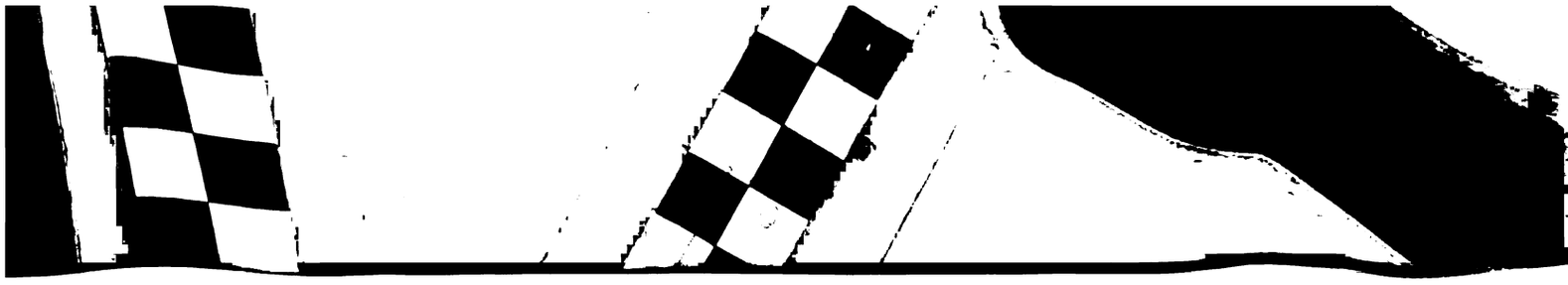
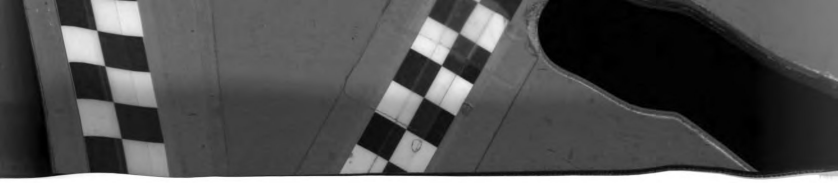


Table of Contents (cont'd)



LIST OF TABLES

LIST OF TABLES

List of Tables (cont'd)

LIST OF FIGURES

I. PROBLEM

Introduction

The licensure and certification processes represent the efforts of governmental and private entities to ascertain and recognize the competence of individuals in the practice of a profession or trade. Licensure, as commonly understood, is the granting, by a governmental entity, of the right to legally practice a profession or trade. The right, or license, is granted pursuant to the individual's demonstrated acquisition of the knowledge or skills required for safe practice. Licensure programs are conducted by governmental entities in their effort--and charge--to protect the public against unsafe practice. Certification is the process by which non-governmental entities, commonly professions or associations, confer a credential. The credential is also usually only conferred upon the individual after demonstration by the individual that a specified level of knowledge or skill has been acquired (Shimberg, 1981).

As reported by Nafziger and Hiscox (1976), over 2000 occupations employ some type of licensure or certification procedures. That number is surely increasing, even leading some to label Americans "the credential society" (Collins, 1979).

Additionally, many entities which once issued permanent licenses or certificates have now begun to reassess the concept of lifetime credential. Instead, time-limited certification or re-credentialing

concepts have begun to be seriously entertained and often implemented, especially in rapidly-changing technical fields such as the medical professions (American Board of Medical Specialties, 1987).

The competence required of a candidate for licensure or certification is usually stated in terms of requisite knowledge, skills, and abilities. Verification that the individual has acquired the knowledge, skills, and abilities is often linked to one or more of three components: a minimum educational attainment, a minimum practice or experience requirement, and a minimum level of performance on an objective test. The examinations used as part of the third component are increasingly criterion-referenced ones.[1] Hambleton, Swaminathan, Algina, and Coulson (1978) have defined such tests as ones that are "used to ascertain an individual's status (referred to as a domain score) with respect to a well-defined behavior domain" (p. 2). These tests consist of items that are a "representative set of items from a clearly-defined domain of behaviors measuring an objective" (p. 3).

The present research focusses on the last of the three components in licensure and certification testing programs--the criterion-referenced test. Specifically, this research examines one particular test score of unique interest--that score from which emanates inferences of mastery or competence--the passing score.

The passing score on an criterion-referenced examination represents the establishment of a standard of performance judged to

---

[1] It is recognized that terms such as "criterion-referenced," "domain-referenced," and "norm-referenced" precisely describe test score interpretations and inferences rather than the instruments themselves. However, imprecise use of these terms in referring to instruments is ubiquitous--even among measurement specialists (Cronbach, 1989). This relaxed usage, though imprecise, is followed throughout this manuscript for purposes of ease and clarity.

concepts have begun to be seriously embraced and often implemented, especially in rapidly-changing technical fields, such as the medical professions (American Board of Medical Specialties, 1997).

The competence required of a candidate for licensure or certification is usually stated in terms of requisite knowledge, skills, and abilities. Verification that the individual has acquired the knowledge, skills, and abilities is often tied to one or more of three assessment purposes: educational attainment, a minimum practice or performance, and a minimum level of performance on an objective [illegible] a competition used as part of the third component are criterion-referenced ones.[1] Hambleton, Swaminathan, Algina, and Coulson (1978) have defined such tests as ones that are "used to describe an individual's status (referred to as a domain score) with respect to a well-defined behavior domain" (p. 2). Those tests consist of items that are a "representative set of items from a clearly-defined domain of behaviors measuring an objective" (p. 3).

The present research focuses on the last of the three components in licensure and certification testing programs--the criterion-referenced test. Specifically, this research examines one particular test score of undue interest--that score from which ensuing inferences of mastery or competence--the passing score.

The passing score on an criterion-referenced examination represents the establishment of a standard of performance judged to

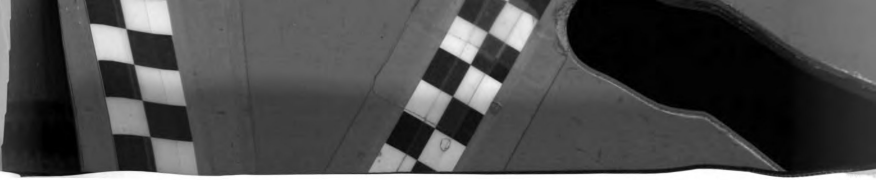
---

[1] It is recognized that terms such as "criterion-referenced," "domain-referenced," and "mast-referenced" precisely describe test score interpretations and inferences rather than the instruments themselves. However, imprecise use of these terms in referring to instruments is ubiquitous--even among measurement specialists (Cronbach, 1989). This relaxed usage, though imprecise, is followed throughout this manuscript for purposes of ease and clarity.

be acceptable. It is the lowest score that permits the examinee to receive the license or credential. Sometimes, though less and less so, the passing score is set in a norm-referenced manner. That is, the passing score is fixed relative to, or dependent upon, the performance of some group. For example, a norm-referenced or "relative" approach to standard-setting might result in requiring examinees to score at or above the 85th percentile, or at or above some number of standard deviations away from average performance on the examination.

However, because the focus of licensure and certification programs has increasingly become that of assessing examinees' competence with respect to a pre-judged standard of performance, norm-referenced standard-setting procedures have been called into question in terms of their propriety for the stated purpose. In their place, "absolute" or criterion-referenced methods of establishing passing standards have become more common. The absolute methodologies, while boasting of greater intuitive and political appeal, still face challenges with respect to the validity of inferences that are made as a result of their resulting standards (Jaeger, 1979). Specifically, the possibility of establishing a standard that results in the failure of a truly competent person (a "false negative") or results in the passing of a truly incompetent person (a "false positive"), is of particular concern.

Criterion-referenced standard-setting methodologies have clearly not yet accomplished technical perfection; much work remains to be done in this area (Hambleton, et al, 1978; Angoff, 1988). The present research addresses one aspect of the process by which standards are



set on a criterion-referenced certification examination in a medical specialty.

Background

Since at least 1954 when Nedelsky sought to derive "absolute grading standards for objective tests" (Nedelsky, 1954, p. 3), the problem of how to establish passing standards on criterion-referenced educational assessments has persisted. Nedelsky's early work prompted investigation of alternative standard setting procedures designed to establish passing standards that differed from the dominant norm-referenced approaches of the time. Nedelsky's objective, and that many of contemporary researchers in the field of standard setting, was straightforward:

> "The passing score [should] be based on the instructor's judgment of what constitutes an adequate achievement on the part of a student and not on the performance by the student relative to his class or to any other particular group of students" (Nedelsky, 1954, p. 3).

The past three and one-half decades have witnessed the introduction of many alternative methodologies that have shared the same objective—movement away from the dominant norm-referenced, or relative, approaches. Among the proposed "absolute" methods as they are sometimes called, the most well-known are those proposed by Nedelsky (1954), Angoff (1971), Ebel (1972), and Jaeger (1982).

Other methods have also been introduced that have tried to achieve a compromise between the absolute and relative approaches. Proposals by Beuk (1984), deGruijter (1980), and Hofstee (1983) represent attempts to synthesize absolute and relative methods.

Taken together, all of these methods represent efforts to

formalize a set of rules for establishing passing standards in a less arbitrary, or at least more justifiable, fashion than traditional, norm-referenced practice has offered. The methods rely primarily on the use of subject matter experts' (hereafter called "SMEs") judgments concerning one or both of two critical elements: a conceptualization of the "barely-passing," "minimally-competent," or "borderline" examinee; and, an expectation regarding the level of content knowledge and skill that such an examinee should possess (Livingston & Zieky, 1982).

After initial research efforts to derive absolute and, later, compromise methods of establishing passing standards, a second stream of research developed. This second line of inquiry focussed mainly on differences between methodologies (Mills & Melican, 1988). Investigations comparing two or more methods characterized this second phase of standard-setting inquiry. Appendix A lists some of these inter-methodological investigations.

Recently, however, a third phase of research need has emerged. Research in this phase is characterized by attempts to identify sources of variation within standard-setting methods.

Need

The proposed research is closely aligned with the third phase of research into standard setting methodologies and focusses on one method--the Angoff method. The Angoff method and its variations (sometimes called "Modified Angoff" procedures) are derived from the work of Angoff (1971) and others. The Angoff methods require SMEs to serve as item reviewers and to scrutinize each item in an

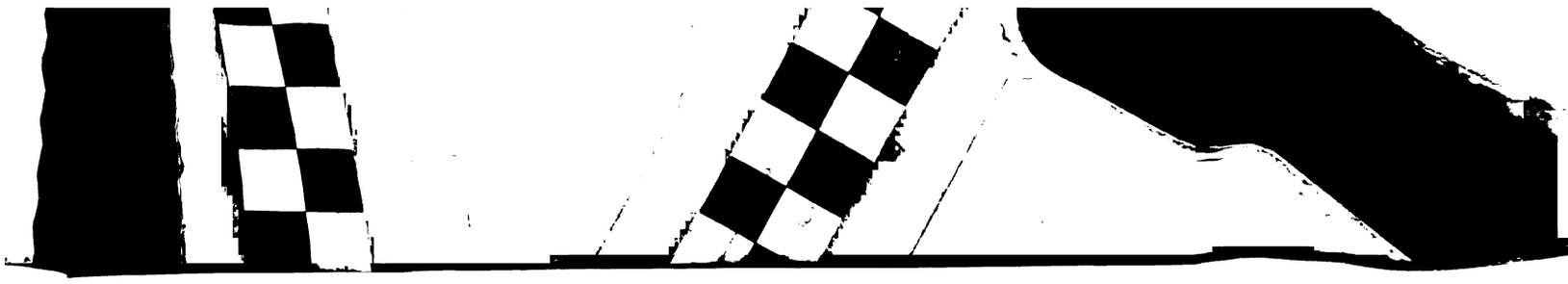
formalize a set of rules for establishing passing standards in a less arbitrary, or at least more justifiable, fashion than traditional, norm-referenced procedures have offered. The methods rely primarily on the use of modern mental capacity (hereafter called "CMCs") possessed commonly (one or both of two critical elements: a conceptual ability to "fairly-possess," "minimally-comprehend," or "mentally" construct and, an expectation regarding the level of content knowledge and skill that such an examinee should possess (Livingston & Kimy 1989).

After initial research efforts to derive absolute and, later, comparative methods of objectively passing standards, a second stream of research developed, that reflects line of inquiry focused mainly on difficulty judgments (Nedelsky (Mills & Melican, 1988). These judgments regarding two or more methods characterized this second type of standard-setting inquiry. Appendix A lists some of these inter-methodological investigations.

Recently, however, a third phase of research has emerged. Research in this phase is characterized by attempts to identify sources of variation within standard-setting methods.

Ebel

The proposed research is closely aligned with the third phase of research into standard setting methodologies and focusses on one method—the Angoff method. The Angoff method and its variations (sometimes called "Modified Angoff" procedures) are derived from the work of Angoff (1971) and others. The Angoff method requires SMEs to serve as item reviewers and to scrutinize each item in an



examination, usually prior to the administration of the examination. The item reviewers are then asked to judge, for each item, the proportion of minimally competent examinees who will answer the item correctly. The item reviewers' judgments, in the form of proportions, are commonly referred to as "Angoff ratings."

The now-preferred means of obtaining the item reviewers' judgments utilizes a group-process format. In this format, the panel of SMEs is convened in a single location, provided with training in the standard-setting methodology, and directed to provide their ratings for each item in a test. The group-process format is often preferred because, predictably, item reviewers do not produce identical ratings and the group-process format provides a means of resolving the differences in ratings. Most researchers agree that this reduction of variability is desirable (Jaeger, 1988; Meskauskas, 1986; Smith, Smith, Richards, & Barnhardt, 1989). However, it is common that an extensive portion of a group's meeting time is devoted to discussions about individual test items, debate, and, when applicable, to consensus-reaching regarding the ultimate rating for each test item.

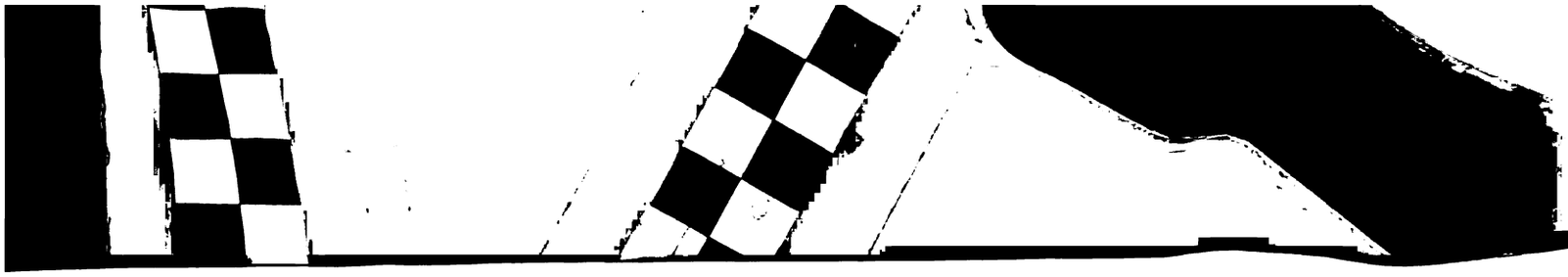
Several problems arising from this format necessitate the investigation of alternatives to the traditional group-process format. Norcini, Lipner, Langdon, & Strecken (1987) summarized two of the problems, including: the tediousness of the task of reviewing individual items and reaching consensus ratings (especially when a large number of items is involved); and, the expense of empaneling a sufficiently large group of SMEs in one location for, perhaps, several days. These problems are especially evident in the area of

professional licensure and certification where hundreds of credentialing programs employ criterion-referenced standard-setting methodologies, most of these relying on subject matter experts participation in a traditional group-process format to obtain item ratings.

Another frequently encountered problem is simply arriving at a single block of time that is available for each SME on the panel of item reviewers. This problem has been characterized by Lockwood, Halpin, and McLean (1986) as one of the "situational constraints" (p. 6) in the standard-setting process. Hambleton (1978, p. 282) specifically addresses the problem of time resource availability as one of the four primary considerations in selecting a standard-setting methodology.

In addition to the need for research to suggest alternatives for addressing the problems created through use of the group-process format in standard-setting studies, research is needed to examine the effect on resultant standards when such alternative strategies are tried. Many researchers have conducted comparative studies of standard-setting methodologies which employ a group-process format. Also, most have offered an opinion concerning the appropriateness of the group-process technique. For example, Brennan and Lockwood (1980) opine:

> "Sometimes...it is suggested that a cutting score be determined by a reconciliation process. For example, after the five raters in this study completed the Angoff procedure, they were instructed, as a group, to reconcile their differences on each item. One typical result of using a reconciliation process is that certain raters tend to dominate, or to influence unequally, the reconciled ratings... There is a certain logic to using a reconciliation process that appears to be compelling. It might be argued that the ideal of using either the Nedelsky



> or the Angoff procedure is for raters to agree on every item. Therefore, why not force them to concur? One argument against this logic is that forced consensus is not agreement, although forced consensus may effectively hide disagreement. Also, a reconciliation process does not guarantee that the same cutting score will result each time a study is replicated" (p. 235-236).

Although Brennan and Lockwood's remarks go beyond the effect of group-process and extend into the realm of requiring consensus of the expert group, their logic is equally applicable to the traditional group-process condition. That is, after appropriate training of item reviewers, the condition of group-process may not be necessary, desirable, or efficient for use in all standard-setting procedures.

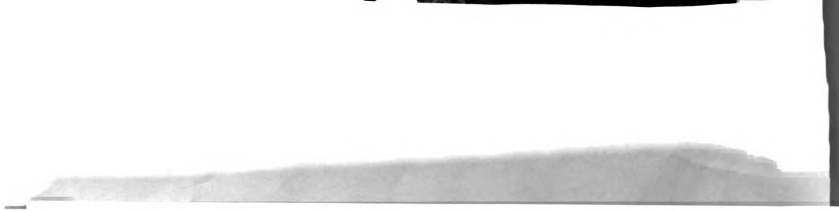
Jaeger (1988) offered his opinion on another aspect of achieving agreement among item reviewers:

> "Achieving consensus on an appropriate standard for a test is an admirable goal (certainly guaranteed through the use of a single judge), but it should not be pursued at the expense of fairly representing the population of judges whose recommendations are pertinent to the task of establishing a workable and equitable test standard" (p. 29).

Maslow (1983) has remarked that knowledge about "the optimal size and structure for the group of judges" is "basic to improving practice in standard setting (p.104), and that "the research literature gives only brief and unsteady guidance here" (p. 105). While some investigation of the issue of optimal group size has begun (Smith, Smith, et al, 1989), the issues surrounding optimal group structure remain largely unaddressed.

Meskauskus, (1986) has appropriately, and succinctly, noted that "[there] is a need to explore the determinants of intrajudge and interjudge variance in depth" (p. 200).

Mills and Barr (1983) reported that:



> "While general information concerning procedures for implementing the methods and calculating cut-off scores is available, specific guidelines are less well established. Issues of training, group interaction, independent ratings vs. discussion all affect the methods, but little is available in either discussion or guidelines concerning these and other implementation issues" (p. 2-3).

In 1984, Fitzpatrick perceived the need for research within standard-setting procedures in an integrative work applying research in the area of social psychology to the problems of standard setting. The need persists, as Fitzpatrick (1989) notes; specifically there is a need by those involved in standard-setting research to investigate the effects of group processes:

> "We must ask whether it is desirable that the decisions that [item reviewers] make be affected by interpersonal comparisons, by cognitive learning through the exchange of information, or by both types of processes" (p. 321).

Focussing in on the social aspects that affect group-based standard-setting methodologies, Fitzpatrick goes on to argue that:

> "standard-setting procedures should be designed to both minimize the effects of social comparison and maximize the effects of certain informational influences on the decisions to be made" (p. 322).

In summary, Fitzpatrick specifically urged that:

> "procedures proposed for reducing the impact of undesirable influences in the standard-setting context should be investigated. Whether or not the suggested procedures will be effective can only be decided by further research" (p. 325).

Unfortunately, scant attention has been paid to these, and similar, aspects of intra-methodological variation. Specifically, as Mills and Barr (1983) and Fitzpatrick (1984) have both remarked, little evidence has been brought to bear on the effect of the presence or absence of the group-process condition. Fewer still appealing alternatives to the group-process format have been

proposed. Curry (1987) has summarized the existing state of affairs aptly:

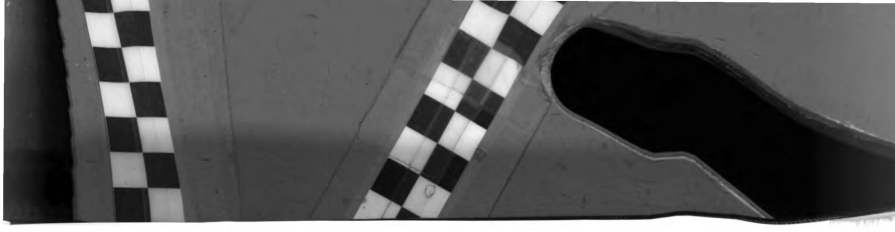
> "Almost all of these authors [on standard setting] acknowledge that the expert group process will have significant impact on the validity of the outcome, few have examined the dynamics involved" (p. 1).

Purpose

The present research attempts to identify an efficient variation of the traditional group-process method for use with the Angoff approach to establishing passing standards on a certification examination.

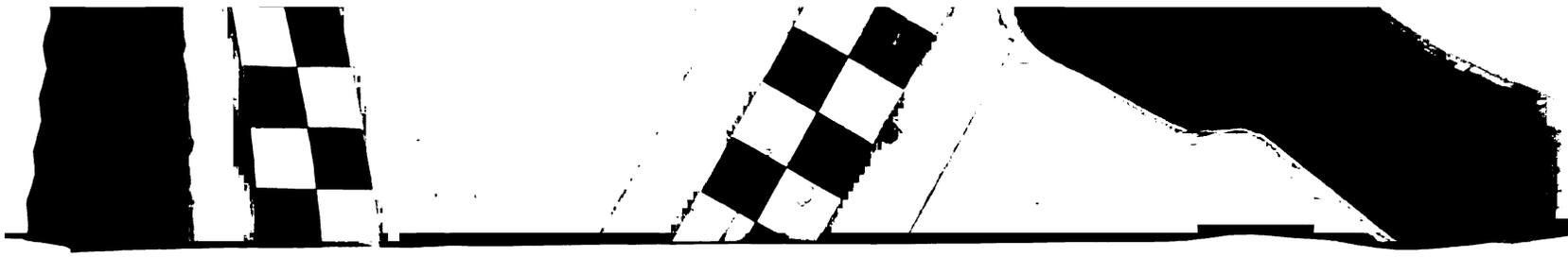
Using the Angoff (1971) method, the present research compares two procedures for establishing passing standards on a medical specialty certification examination. The two procedures used are: 1) the traditional group-process method; and, 2) an "independent" condition in which item reviewers provide their item ratings in isolation (i.e., without the effects of group-process). An attempt is made to determine whether, after both groups of item reviewers are provided with initial training in the Angoff method, results obtained from the group-process condition differ from those obtained in the isolation condition.

The primary focus of the Angoff standard-setting method is to identify a passing score for an examination. Accordingly, the primary focus of this research is to establish whether there is variation in the passing scores that result from exposure to the two conditions. It is hypothesized that variation will be observed between the two conditions, but that the magnitude of variation will be small. Additionally, it is hypothesized that the isolation



condition will provide a suitable, efficient alternative to the traditional group-process method of collecting SMEs' Angoff ratings for test items.

combine will provide a suitable, efficient alternative to the traditional hand-plucking method of collecting DATA. Ample testing for that issue.



II. REVIEW OF PREVIOUS RESEARCH

The setting of absolute performance standards on criterion-referenced educational assessments is a pervasive activity in the American educational system (Hambleton, 1978) and represents an ongoing line of inquiry in the field of educational measurement. Criterion-referenced standard-setting methods are currently utilized by groups responsible for industrial personnel selection, educational and training program evaluation, professional licensure or certification in medical, allied health, and business fields, and other national, state, and regional credentialing programs (AERA/APA/NCME, 1985; Meskauskas, 1986).

Adapting the conceptual view suggested by Mills and Melican (1988), research on criterion-referenced standard-setting can be viewed as having proceeded in three distinct phases: 1) Methodological Development; 2) Inter-Methodological Research; and, 3) Intra-Methodological Research. An overview of these three phases serves as an organizational framework for reviewing previous research and is presented in the following pages.

Methodological Development

As one author has noted, mentions of criterion-referenced passing standards are found in early historical accounts of testing

situations:

> "A very early minimal competency exam was when the Gilead Guards challenged the fugitives from Ephriam who tried to cross the Jordan river. 'Are you a member of the tribe of Ephriam?' they asked. If the man replied that he was not, then they demanded, 'Say Shibboleth.' But if he couldn't pronounce the 'sh' and said Sibboleth instead of Shibboleth he was dragged away and killed. So forty-two thousand people of Ephriam died there at that time" (Judges 12: 5-6, The Living Bible, quoted in Mehrens, 1981, p.1).

Since that time, so-called "high-stakes" tests, (though not that high), have remained prominent in the assessment of competence, and research efforts have been directed at refining the theoretical and applied aspects of setting passing scores on such tests. In a review of existing standard-setting methodologies, Berk (1986) reported that at least 38 methods of establishing or adjusting performance standards have been proposed. Berk (1980; 1986) and many others (Glass, 1978; Hambleton & Eignor, 1980; Hambleton, Swaminathan, Algina, & Coulson, 1978; Jaeger, 1989; Livingston & Zieky, 1982; Meskauskas, 1976; Meskauskas & Norcini, 1980; Millman, 1973; Mills and Melican, 1988; and, Shepard, 1980a) have also developed several similar catalogues and classification schemes to organize the various methodologies.

Again from a historical perspective, Nedelsky's (1954) work probably represents one of the first attempts to promote absolute, or criterion-referenced standards of performance on educational assessments. As late as the 1970s, norm-referenced methodologies dominated as the preferred standard setting approach. In a 1976 article, Andrew and Hecht reported that:

> "At present, the most widely used procedures for selecting ...pass-fail levels involves norm-referenced considerations in which the examination standard is set as a function of the performance of examinees in relation to one another" (Andrew & Hecht, 1976, p. 45).

A noticeable shift began to occur during the 1970s and 1980s, when considerable attention to establishing absolute passing standards resulted from an increasing popularization of criterion-referenced testing (Glaser, 1963; Popham & Husek, 1969), or—as some have termed it—a shift to a focus on educational "outputs" (Levin, 1978; Rothman, 1989). Since that time, many entities responsible for establishing passing standards have reevaluated their use of norm-referenced methodologies and have opted for implementation of absolute or compromise methods (Hambleton, 1978; Fabrey, 1988; Mills & Barr, 1983). In 1983, Francis and Holmes reported that "the more traditional norm-referenced approach is being seriously questioned" (p. 2). Meskauskas (1986) described the evident trend away from norm-referenced and toward absolute (or, "content-referenced") standard setting methodologies in the area of licensure and certification testing and offered this advice:

> "For those credentialing agencies still using normative standards, I recommend that plans to change over to content-referenced standards be initiated" (p. 198).

Nedelsky's work in search of an absolute standard-setting methodology thus represents a marked turning point in standard-setting technology and research. When using the Nedelsky method, subject matter experts carefully inspect the content and items in an examination and judge, for each item in the test, the option or options that a hypothetical minimally-competent examinee would rule out as incorrect. The reciprocal of the remaining number of options becomes each item's "Nedelsky rating"; the sum of the ratings—or some adjustment to the sum—is used as a passing score.

Further research and other now-popular methods of establishing

absolute passing standards on criterion-referenced examinations followed—though not quickly (Scriven, 1978)—after Nedelsky's 1954 publication. Angoff (1971) proposed a method that, like Nedelsky's, required SMEs to review test items and to provide estimations of the proportion of a subpopulation of examinees who would answer the items correctly:

> "A systematic procedure for deciding on the minimum raw scores for passing and honors might be developed as follows: keeping the hypothetical 'minimally acceptable person' in mind, one could go through the test item by item and decide whether such a person could answer correctly each item under consideration. If a score of one is given for each item answered correctly by the hypothetical person and a score of zero is given for each item answered incorrectly by that person, the sum of the item scores will equal the raw score earned by the 'minimally acceptable person'." (Angoff, 1971, pp. 514-515).

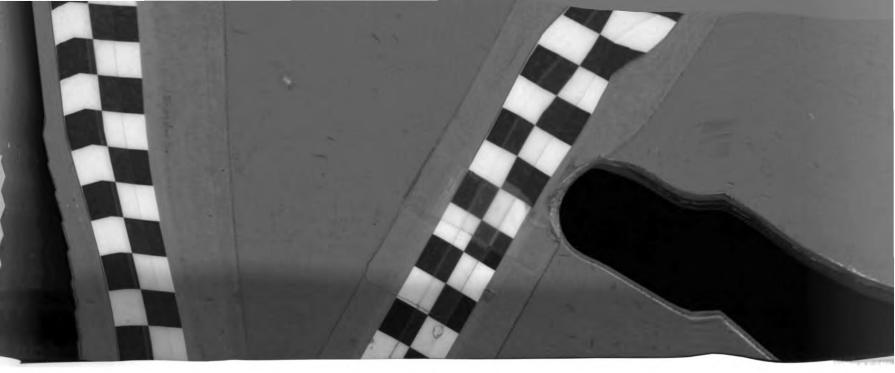
In practice, a footnoted variation to the procedure Angoff originally proposed has dominated applications of the Angoff method:

> "A slight variation of this procedure is to ask each judge to state the probability that the 'minimally acceptable person' would answer each item correctly. In effect, judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities would then represent the minimally acceptable score." (Angoff, 1971, p. 515).

A third absolute method was proposed by Ebel (1972), who also noted that norm-referenced methods had serious drawbacks:

> "The obvious drawback of this approach is that it allows the passing score to vary according to the general level of competence of the examinees at a specific testing." (Ebel, 1972, p. 494).

Ebel's methodology also involves the judgments of subject matter experts. The Ebel method requires SMEs to make decisions about the difficulty of individual test items and about the criticality of test

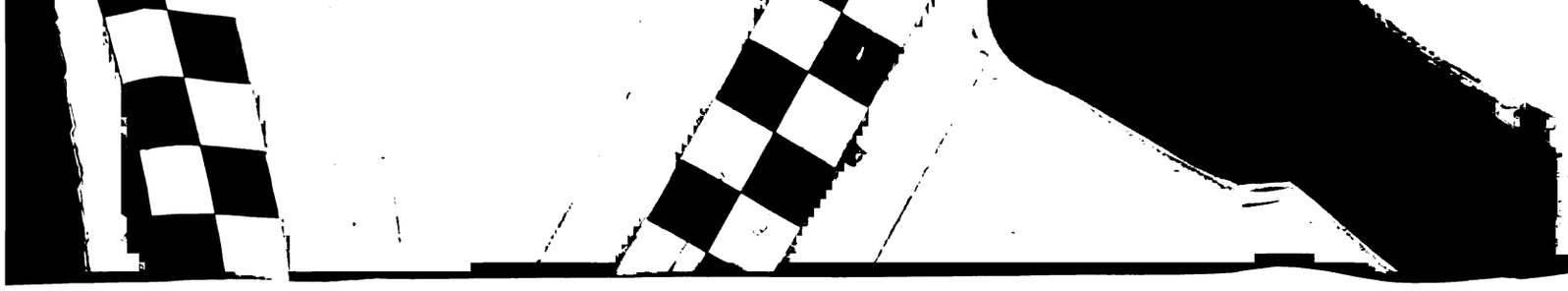


content areas.

Other absolute methodologies have also been proposed, some quite recently. One alternative based on rating test specifications was proposed by Cangelosi (1984). Lockwood, et al (1986) proposed a method of averaging the results of various standard-setting approaches in order to at a "true" standard, or precise estimate of some extant parameter. Another methodology has been proposed by Schoon, Rosen, and Jones (1988) in response to perceived weakness in the Angoff approach. Schoon, Rosen, and Jones also did some preliminary investigation into their "Direct Standard Setting Method" (Jones, Rosen, & Schoon, 1988), but it, like other alternatives to the Angoff, Ebel, and Nedelsky methodologies, has not received widespread acceptance or general use.

A second wave of proposed standard-setting methodologies followed early attempts at determining absolute passing standards. Predictably, the second wave aspired to identify a middle ground through the development of methodologies that would strike a compromise between purely norm-referenced (relative) approaches and absolute methods. Illustrative of these compromise efforts are methodologies suggested by Beuk (1984), Grosse and Wright (1986), Hofstee (1983), and deGruijter (1980). Overviews of these methodologies are provided in deGruijter (1985) and Mills and Melican (1986).

The compromise methodologies have failed to overtake the earlier absolute proposals, however. Currently, in the area of licensure and certification testing, the Angoff, Ebel, and Nedelsky approaches are still the most prevalent methodologies for establishing passing



standards, particularly the Angoff and Ebel approaches (Hambleton, 1978; Berk, 1986).

Albeit a ubiquitous task, the establishment of passing standards is not necessarily an easy one. Referring specifically to licensure and certification testing programs, the Standards for Educational and Psychological Testing remark that:

> "Defining the level of competence required for licensing or certification is one of the most important and difficult tasks facing those responsible for such programs" (AERA/APA/NCME, 1985, p. 63).

In a discussion of absolute standard-setting however, it should also be noted that considerable disagreement exists concerning just how absolute the absolute standard-setting procedures are. Glass (1978) calls decisionmaking within the absolute standard-setting process "judgmental, capricious, and essentially unexamined" (p. 253), and further notes that "to my knowledge, every attempt to derive a criterion score is either blatantly arbitrary or derives from a set of arbitrary premises" (p. 258). Similarly, Beuk (1984) has noted that "setting standards...is only partly a psychometric problem (p. 147). Hofstee offers support for the idea that:

> "a [standard-setting] solution satisfactory to all persons involved does not exist and...the choice between alternatives is ultimately a political, not a scientific, matter" (1983, p. 109).

Jaeger claims, flatly:

> "All standard-setting is judgmental. No amount of data collection, data analysis, and model building can replace the ultimate judgmental act of deciding which levels of performance are meritorious or acceptable and which are unacceptable or inadequate" (1979, p. 48).

Shepard identified the essence of the problem of arbitrariness in the so-called absolute methods:

> "[N]one of the [standard-setting] models provides a scientific means for discovering the 'true' standard. This is not only a deficiency of the current methods but is a permanent and insolvable problem because the underlying competencies being measured are continuous and not dichotomous" (1980b, p. 67; cf, Shepard, 1978, p. 62).

Even Ebel, whose standard-setting method has remained popular, resigned himself to the fact that a certain amount of subjectivity remains in "absolute" standard-setting methods:

> "A second popular belief is that when a test is used to pass or fail someone, the distinction between the two outcomes is clear-cut and unequivocal. This is almost never true. Determination of a minimum acceptable performance always involves some rather arbitrary and not wholly satisfactory decisions" (Ebel, 1972, p. 492).

Hambleton summarized the overwhelming consensus of opinion:

> What is clear is that all of the methods are arbitrary and this point has been made or implied by everyone whose work I have had an opportunity to read. The point is not disputed by anyone I am aware of." (1978, p. 281).

However arbitrary and problematic (deGruijter & Hambleton, 1984; Shepard, 1980b), standards are still essential for making certain inferences and, accordingly, credentialing decisions. The need for valid standard-setting is especially apparent in the areas of certification and licensure, where ensuring the public's protection against unsafe practice is the real and necessary charge of the responsible entities (Lerner, 1979; Maslow, 1983; Shepard, 1983). As Levin has remarked:

> "Unless all forms of certification are eliminated, however, a standard is still needed whether the performance is sufficient to receive the certification" (1978, pp. 306-307).

In summary, while ambivalence remains over the degree of arbitrariness inherent in absolute standard-setting methods, their intuitive appeal, ease of implementation, and perceived advantages in

terms of both psychometric properties and defensibility over the previously popular norm-referenced approaches have been documented by numerous researchers (Berk, 1986; Cross, Impara, Frary, & Jaeger, 1984; Klein, 1984; Meskauskas, 1986). The use of absolute standard-setting methods continues to become increasingly widespread. Research into development of new methodologies, particularly compromise approaches, and empirically-based methods of adjusting standards (Hambleton, 1978) and into assessing the validity of the resultant standards (Jaeger, 1979; Kane, 1985) continues.

Inter-Methodological Research

Having gained increasing acceptance by the measurement profession generally, absolute methods of establishing passing standards began to realize widespread use in the determination of cut-off scores on educational, licensure, and certification tests (Gross, 1985). A logical second phase of research developed: investigation of the psychometric properties of the various standard-setting procedures. This second phase of research is characterized largely by attempts to compare two or more standard-setting methodologies in terms of their reliability and ability to identify an "acceptable" standard. As late as 1988, Smith and Smith reported that:

> "Much of the work in the area of standard setting has been concerned with comparisons of different methods for establishing a criterion." (p. 259).

In testament to the proliferation of inter-methodological research, Berk (1986) reports that in the five-year period, 1981-1986, 22 studies were conducted to compare standards resulting from the application of different standard-setting methodologies. Extensive

descriptions of the various inter-methodological comparison studies are provided elsewhere (Berk, 1986; Jaeger, 1989). A partial listing of inter-methodological is also provided in this work as Appendix A. (Because the present research is limited to applications of one absolute standard-setting approach, Appendix A lists only those studies reporting comparisons involving one or more absolute standard-setting methodologies.)

One result of the wealth of inter-methodological research appears certain: Different standard-setting methodologies yield different standards (Andrew & Hecht, 1976; Brennan & Lockwood, 1980; Koffler, 1980; and Skakun & Kling, 1980). Different methods even produce different performance standards when applied to the same tests by the same group of experts (Mills, 1983; Mills & Barr, 1983). More tentative and method-specific conclusions apply to studies when different groups of experts, apply the same methodology to the same test (Cross, et al, 1984; Fabrey & Raymond, 1987; Jaeger, 1988, 1989; Rock, Davis & Werts, 1980).

A second result of the inter-methodological research effort is also compelling: The Angoff approach seems to be the preferred absolute standard-setting methodology by several criteria. Mills and Melican (1988) report that,

> "the Angoff method appears to be the most widely used. The method is not difficult to explain and data collection and analysis are simpler than for other methods in this category" (p. 272).

Similarly, Klein (1984) noted that the Angoff method is preferable "because it can be explained and implemented relatively easily" (p. 2). Rock, Davis and Werts (1980) concluded that "the Angoff cutting score seems to be somewhat closer to the 'mark'" (p. 15).

Colton and Hecht (1981), in their comparison of the Angoff, Ebel, and Nedelsky methodologies, report that "the Angoff technique and the Angoff consensus techniques are superior to the others" (p. 15). Cross, et al (1984) found that the Angoff method "yielded the most defensible standards" (p. 113). Berk (1986) concluded that "the Angoff method appears to offer the best balance between technical adequacy and practicability" (p. 147). Meskauskas (1986) states that, "the present method of choice for standard-setting is the Angoff method (p. 199). Finally, in their study comparing the Angoff and Nedelsky methods, Smith and Smith (1988) report "an urge to say, 'Yes, the Angoff approach is more valid'" (p. 272).

Intra-Methodological Research

The line of inquiry joined by the present research is a newly emerging one (Mills & Melican, 1988) that seeks to identify sources of variation within and efficient refinements of existing standard-setting methodologies. Few systematic research efforts have been directed at this critical facet within the field of standard-setting research. As Smith and Smith reported bluntly, "little work has been done to explain why differences in standards occur" (1988, p. 259). Smith and Smith (1990) proceeded to pursue one aspect of why differences in standards might occur in an investigation where item reviewers using the Angoff method were asked to attend to only specified characteristics of reading comprehension items. Unfortunately, the authors reported somewhat discouraging results, and asked:

> "Where does this leave us? First of all, perplexed, as usual. Second, reluctant to recommend giving judges

> information about what characteristics to use or ignore" (p. 22).

Historically, the necessity of identifying sources of intra-methodological variation has never been totally overlooked by research efforts in the area. Nedelsky at once recognized the need to identify and reduce sources of variation in the method he proposed. One hypothesized source of variation—and, possibly invalidity—was the training of item reviewers. Early on in the search for absolute standards, Nedelsky warned:

> "To make a proper judgment of this kind, requires time and considerable pedagogical and test-wise sophistication; with responses more heterogenous than in the example cited a reliable judgment may be impossible." (Nedelsky, 1954, p. 7).

Indeed, the proper training of qualified item reviewers has been repeatedly emphasized by those involved in standard setting research as crucial to the validity of the process (Francis & Holmes, 1983; Jaeger, 1979, 1989; Klein, 1984; Scriven, 1978). For example, in their procedural guide to several popular standard-setting methodologies, Livingston and Zieky (1982) restate the necessity of reducing variation and invalidity of judgments made by SMEs, devoting extensive portions of their manual to describing the proper training of judges. Smith, et al. (1989) state succinctly: "Variability in the judgmental process needs to be reduced" (p. 7).

Aside from admonitions concerning the training of item reviewers, attention to other intra-methodological considerations has been slight, but growing. A beginning, though sophisticated attempt to identify other sources of intra-methodological variation was put forth by Smith and Smith (1988) who compared sources of information used by

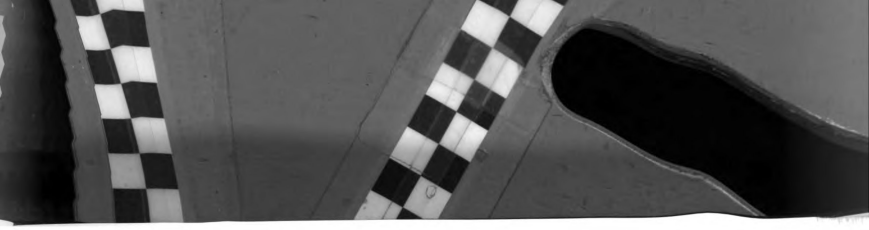
Angoff and Nedelsky item reviewers in the item rating process. The primary objective of their work was to pinpoint differences between the two methodologies; thus, it would still be categorized in the second-phase of research efforts. However, it also clearly represents one of the first attempts at identifying intra-methodological variation because of its unique investigation into which item characteristics are salient to item reviewers within the Angoff and Nedelsky methods.

Saunders, Ryan, and Huynh (1981) also investigated two variations of the Nedelsky approach, differing only in the extent to which item reviewers were permitted to respond "undecided" when considering whether minimally-competent examinees would rule out an item's option as incorrect. They found that the two conditions "produce[d] essentially equivalent results" (p. 209).

Another investigation into the Nedelsky procedure by Gross (1984) led the author to suggest a refinement in the test construction process that would maximize the consistency of the Nedelsky methodology.

Plake and Melican (1986) found that, with the Nedelsky method, item reviewers for a mathematics test made fairly consistent item ratings, regardless of test length or difficulty. Dillon (1990) found no strong relationship between the position of an item in an examination and the Angoff rating reviewers assigned to the item.

Saunders, et al, (1981) and Halpin, Sigmon, and Halpin (1983) found significant within-method differences in item reviewers' ratings due to differences in the reviewers' own levels of achievement in the subject areas, although Behuniak, Archambault, and Gable (1982)



reported finding no such differences. Mills and Melican (1990) reported that little or no differences in passing standards were observed for randomly equivalent panels of item reviewers.

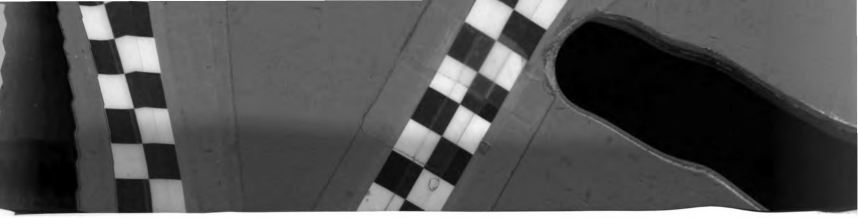
Norcini, Shea, and Kanya (1988) reported fairly high consistency in experts' estimates of borderline group performance when using the Angoff method on a medical specialty examination. Melican and Mills (1987) reported increased p-values and higher intercorrelations among item reviewers' item ratings when reviewers were provided with knowledge about the other reviewers' ratings.

Garrido and Payne (1987) studied two variations of the Angoff method under two conditions—with and without item performance information provided to the item reviewers. In this experiment, untrained item reviewers were asked to independently provide ratings for 20 items. The provision of item performance information (p-values) resulted in higher average passing standards and resulted in reduced interjudge variability. However, the authors note that the high correlation between "With-Data" judges' ratings and empirical p-values (r = .98) called into question "the creditability of the judges in their performance of the judging task" (p. 7). The authors further wondered:

> "Did the presentation of such information influence the judges to the extent that they disregarded their own judgments and relied soley on the item difficulty index in determining their probabilities?" (p. 8).

(Interesting, Skakun (1990) also found that the provision of item performance data—even purposefully incorrect item performance data—has the effect of reducing variability in item ratings.)

In another recent study, Friedman and Ho (1990) investigated the



relationship between interjudge variation (consensus) and intrajudge variation (consistency) and found that procedures aimed at improving consensus (such as the provision of item performance information) "did not have an adverse affect on intrajudge consistency; in fact...techniques designed to improve consensus also improved consistency" (p. 10). The authors also utilized several procedures designed to evaluate the effect of eliminating judges with poor internal consistency and judges with poor agreement with the group; however, none of the methods appreared to appreciably alter the overall passing standard.

Another study of intra-methodological variation is reported by Curry (1987) who investigated a standard-setting procedure for a certification examination. Using the Nedelsky standard-setting method, and a group-process format, Curry found significant variation in reviewers' ratings of items resulting in a large percentage of items requiring extensive group interaction (i.e., iterations of the rating process) to achieve consensus on item ratings. While noting a strong "group press towards a norm" and a critical need "to reduce the normative press involved in the use of an expert group" (Curry, 1987, p. 2), Curry does not provide a strong rationale for why the variation in ratings must be reduced, or information about what effect, if any, the initial variation in ratings would have upon a resultant passing score.

Fitzpatrick (1989) reviewed several standard-setting research efforts touching specifically on the group-process format and reported:

> "Discussion among group members ... is thought to elicit informational influences through the exchange of arguments

> as well as social comparison processes through the inferences that group members make about others' positions. Hence, the role of discussion in the standard-setting process is an important topic for future research" (p.322).

Fitzpatrick did not report on research to test the effects of the group-process format in standard-setting. However, she did proceed to strongly suggest that "further studies of standard-setting that involve structured discussion or other methods of controlling biased argumentation clearly are warranted" (p. 323).

Lastly, a study of intra-methodological variation was conducted by Norcini, Lipner, Langdon, and Strecker (1987). Using the Angoff method, Norcini, et al. explicitly tested "the common notion that a group setting is most appropriate for implementation [of the Angoff method]." (Norcini, et al., 1987, p. 56). The research of Norcini, Lipner, et al. was an attempt "to determine whether more efficient variations of the process will provide consistent and accurate results" (p. 56).

Norcini, et al. found relatively small differences between the passing scores obtained using three variations of the Angoff method, each variation differing only in the extent to which item reviewers were exposed to a group-process format. (Each of the modifications of the Angoff method used by Norcini, et al. involved the use of normative feedback with the item reviewers; that is, reviewers were provided with the correct responses to the items under review, as well as with empirically-obtained difficulty indices (p-values) for each item.) Norcini, et al. argued that tentative support had been provided for the notion that two of the three variations of the Angoff method resulted in acceptable passing standards and less interrater

variation.

The basic conclusion of the research by Norcini, et al. is straightforward:

> "In conclusion, this work implies that judgments gathered after an initial traditional group-process session can provide a mechanism for setting cutting scores using a modified Angoff method and make more efficient use of meeting time." (Norcini, et al., 1987, p. 63).

One troubling aspect of the research reported by Norcini, et al. is the failure to control for possible training or "practice" effects in the item reviewers. In their study, SMEs were asked to review test items in each of three conditions. In the first condition, the group reviewed materials sent through the mail describing the Angoff method to be used. Next, the reviewers attended a group meeting where the method was further described, definitions of a "minimally competent examinee," etc., were discussed, and ten practice items were reviewed. Following this training, the item reviewers then received a booklet containing the actual test items, answer key, normative information consisting of item performance statistics, and further review of the Angoff procedures necessary for completing their item ratings. These features are characterized by Norcini, et al., as the "Before-Meeting" condition.

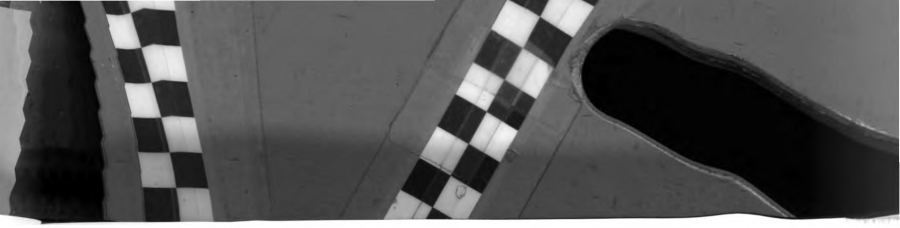
The second condition (called "During-Meeting") was characterized by the same group of item reviewers participating in another meeting to review the Angoff procedure and definitions. Following this review, a traditional group-process Angoff procedure was conducted, with normative information again provided.

The third, and final, condition (called "After-Meeting") was conducted approximately one month following the "During-Meeting"

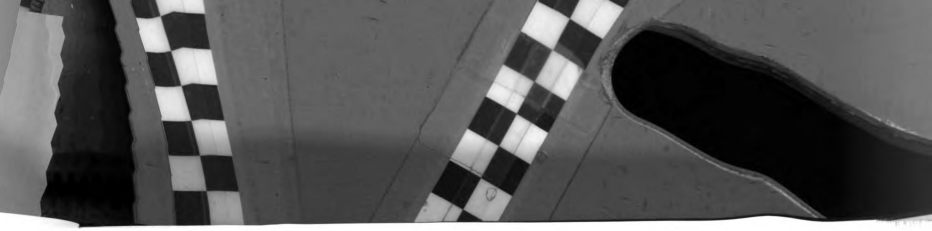
condition. In the "After-Meeting" condition, the same group of item reviewers were again sent a packet of instructional materials, a set of items, answer key, and normative information, and were asked to provide item ratings.

Norcini, et al., report that the resulting passing scores obtained in each of the three conditions did vary, though not significantly [$F(2,10) = 2.04$, $p = .181$]. Also reported is an unsurprising reduction in the variation of item ratings from the "Before-Meeting" condition to the "After-Meeting" condition. Standard deviations of the item reviewers' ratings were 5.8, 2.4, and 1.7 for the Before-, During-, and After-Meeting conditions, respectively.

These results might imply, as Norcini, et al., suggest, that Angoff item ratings collected from item reviewers performing independent item reviews are as reliable as those collected using a traditional group-process format. However, a weaker conclusion also seems tenable: A single group of item reviewers using the Angoff method tends to become less variable in their item ratings when afforded repeated exposure to the method and permitted greater opportunities for practice. Additionally, Norcini, et al., reported that, for the ratings generated in the Before-Meeting condition, all of the item reviewers failed to take guessing into account when providing their ratings. The reviewers were, however, instructed to account for examinee guessing for ratings they subsequently provided in the During- and After-Meeting conditions (presumably using $p = .20$ or $p = .25$ as the lowest rating possibility). This factor could well have contributed substantially to the reduction in variation observed across conditions.



In summary, another test of the propositions put forth by Norcini, et al. seems warranted and is offered in the present study.



III. STUDY DESIGN

The present research has two purposes: 1) to determine whether item reviewers, using the Angoff (1971) method of assigning probabilities to examination items, produce different ratings as a result of exposure to a traditional group-process condition and an isolation condition, and 2) to investigate the effect of knowledge of other item reviewers' initial Angoff ratings on a subsequent rating of the same items. Two experiments to address these questions are presented.

Experiment 1

The design for the first experiment is one which: 1) randomly assigned item reviewers to each of the two conditions; 2) obtained the reviewers' ratings on a common set of items; and, 3) compared the resultant ratings.

The design for Experiment 1 is analogous to the "Posttest-Only Control Group Design" presented by Campbell and Stanley (1963, p. 25). In the notation suggested by Campbell and Stanley, this true experimental design can be symbolized as follows:

GROUP 1: R O_1 [control group - (group-process condition)]

GROUP 2: R X O_2 [treatment group - (independent condition)],

where:

R indicates random assignment to a condition,

X indicates the administration of a treatment, and

O indicates an observation or data collection.

In the present research, item reviewers were randomly assigned to one of the two conditions--isolation or group-process. The traditional group-process condition is analogous to a "no treatment" or control group, and the isolation condition represents a new treatment. The above design, called "greatly underused in educational and psychological research" (Campbell and Stanley, 1963, p. 26), has the advantage over other design choices of offering strong resistance to factors that would weaken the internal validity of the research. That is, the experimental design--primarily due to the initial random assignment to the two conditions--offers a strong potential for discovering true differences between the two groups' ratings after the treatment has been administered, if such differences exist.

Although, "'knowing for sure' that the...groups were 'equal'" (Campbell & Stanley, 1963, p. 25) before the experimental treatment is administered is impossible due to the lack of pre-random assignment comparisons, many of the factors that could weaken the study's internal validity (particularly, selection) are effectively controlled for through randomization.

Empirical Treatments

Subjects in the present research were divided into two groups and were exposed to two differing conditions. For purposes of clarity, Group 1--the group that was exposed to the traditional group-process condition--will be referred to as the control group; that is, the

group-process condition can be conceived of as a "no treatment" condition. Group 2—the group that was exposed to the independent condition—will be referred to as the treatment group; the independent condition represents the application of a new treatment. Precise descriptions of the characteristics of the control and treatment groups are important and are presented below.

Control Group

Each subject in the control group was mailed a description of the Angoff (1971) methodology for establishing passing scores approximately one month prior to a meeting at which the actual item ratings were collected. A copy of these instructional materials is included as Appendix B. Approximately two weeks prior to the passing score meeting, each subject in the control group was telephoned by the investigator and questioned concerning his[1] understanding of the mailed materials and feelings of preparedness to undertake application of the Angoff methodology.

A whole-group meeting, including subjects in both the treatment and control groups, was conducted by the investigator on the day of the passing score meeting. At this meeting, the packet of informational materials which was mailed to subjects prior to the meeting served as a foundation for review of important concepts and definitions. Together, both the treatment and control groups then participated in performing practice ratings for 10 non-operational test items. The practice items were drawn from a recently

---

[1] All subjects (treatment and control groups) in the present study were male.

administered test form from the medical specialty program under study and were chosen to be representative of items found in the upcoming, operational test form. Practice items covered a representative range of difficulty, discrimination, and format. Table 1 provides a description of the 10 practice items.

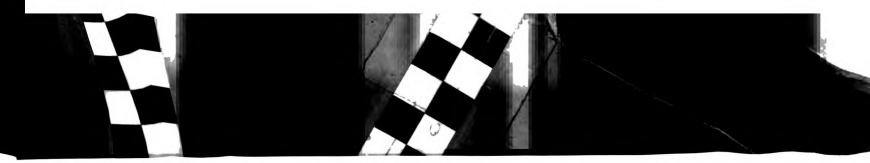


Table 1

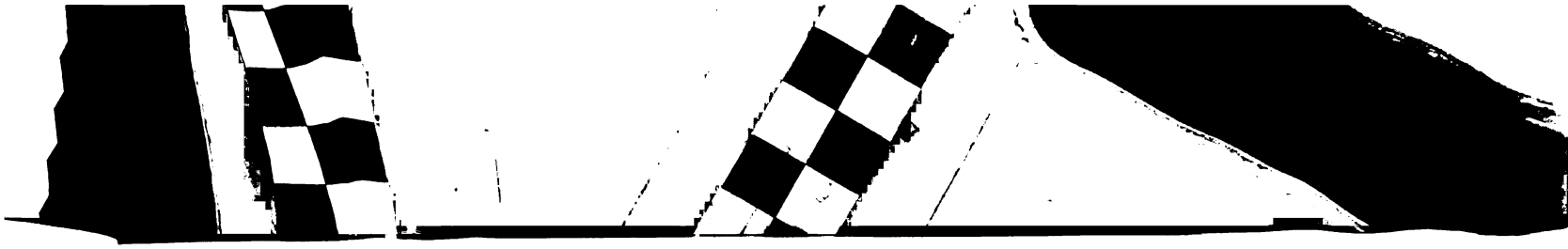
Description of Practice Items Used in Passing Score Study Group Training Session

Item No.	Difficulty	Discrimination*	Wording**	Item Type***
1	.75	.36	P	A
2	.21	.16	P	A
3	.84	.31	P	A
4	.40	.15	P	K
5	.74	.27	P	K
6	.92	.27	N	A
7	.22	.01	P	A
8	.16	.13	P	A
9	.94	.34	P	A
10	.74	.22	P	A

Notes: * - Discrimination indices reported are point biserial correlations.

** - Key to wording: P = positively worded item; N = negatively worded.

*** - Notation for item types is consistent with those suggested in Hubbard (1978).



Each of the practice items was accompanied by item analysis information corresponding to the item's recent use. Provision of item analysis information is consistent with the suggestions of many researchers in standard-setting (Berk, 1986; Conaway, 1979; Jaeger, 1982; Livingston & Zieky, 1982; and Shepard, 1976; 1979; 1980a; 1980b; 1983) for inclusion of normative information to item reviewers so that more reasonable item probabilities (ratings) are obtained.

After both the treatment and control groups completed rating the 10 practice items, all members of both groups were polled to determine their perceived familiarity and comfort with proceeding in the application of the Angoff methodology to the operational test form. Questions and answers and a brief discussion moderated by the investigator followed.

After questions and clarifications, subjects assigned to the control group (group-process condition) remained in the group setting for the remainder of the meeting time. A booklet containing the operational test items was distributed to each subject in the control group. No additional information except an indicator of each item's key was provided to the control group. The group was, however, encouraged to utilize each other and their packets of mailed informational materials on the Angoff method as needed. The investigator remained with the group-process condition group to monitor the discussion of items in that group, and to observe the frequency of discussion, the content of discussion, and the extent to which discussion was dominated by one or more group members.

Subjects in the control group were then asked to record their ratings for each test item on a rating sheet that was provided.

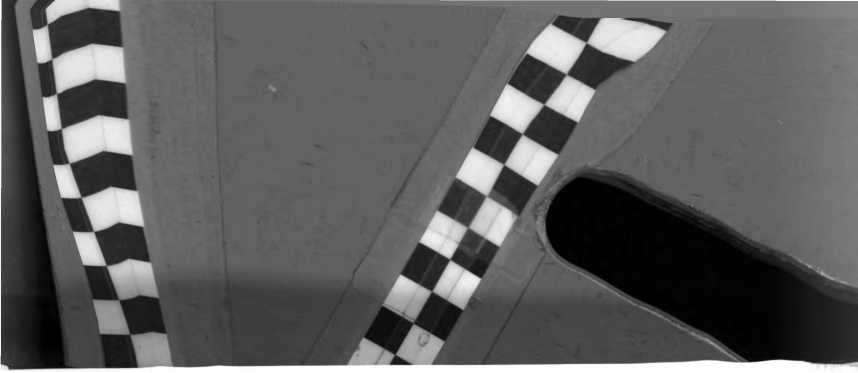
Subjects in the control group proceeded through the test items as a group, pausing frequently to discuss difficult item wordings, review their conceptualization of the minimally-competent examinee, and to compare item ratings for questionable items. However, no forced consensus for item ratings was required, nor was any item reviewer encouraged to change his item rating.

In the present research, item difficulty information (p-values) was not provided to subjects in either the treatment or control groups. Although some researchers have argued that item difficulty information (p-values) should be provided to item reviewers when rating test items in order to increase the consistency of ratings (Cross, Impara, et al, 1984; Norcini, Shea, & Kanya, 1988; and Subkoviak & Huff, 1986), such information was not presented to item reviewers in this study because all items in the to-be-administered test form being reviewed were new (previously untested) items for which performance data were not available.

Rating sheets and all materials were collected from each subject in the control group when the group had completed their ratings for each item. Finally, subjects in the control group responded to a brief questionnaire to obtain descriptive information on the subjects and indicators of their perceptions concerning the passing score study methodology.

Treatment Group

Each subject in the treatment group (isolation condition) was exposed to experiences identical to those encountered by subjects in the control group until the time the treatment was administered.



Specifically, subjects in the treatment group received the same packet of informational materials mailed approximately one month prior to the passing score study meeting, received a follow-up telephone call approximately two weeks prior to the meeting, and participated in the whole-group practice session and discussions on the day of the meeting.

At the conclusion of the practice session, subjects in the treatment group were each provided with the same booklet of test items as subjects in the control group. Subjects in the treatment group were asked to use the booklets and previously mailed informational materials to provide ratings on an accompanying rating sheet for each item in the test form. However, subjects in the treatment group were asked not to discuss their ratings with other treatment group members, members of the control group, or other professional colleagues. Rather, subjects in the treatment group were asked to consider and provide their ratings independently and to return their completed rating forms to the investigator. Like the subjects in the control group, subjects in the treatment group completed and returned, along with their ratings, the post-meeting follow-up questionnaire. All materials were returned by the treatment group to the investigator within two days of the whole-group meeting.

Subjects

Subjects for the present research were 10 members of the Written Examination Committee of a national medical specialty certification Board. Members of the Written Examination Committee are charged with establishing performance standards for the Board's examinations.

Subjects were recognized content experts in the medical specialty area and represented various areas of subspecialty with the profession; each was also a member of the profession's academy. None of the subjects possessed expertise in criterion-referenced standard setting methodologies. Also, each subject indicated that he had not participated in a previous standard-setting study.

Consent

Each member of the Written Examination Committee agreed to participate in the present research. The Board's permission to conduct the study was granted through execution of a contract with the American College Testing Program, Inc., to perform various assessment services. The contract specifically covered the conduct of a passing score study for the Board. Permission to use data obtained in the conduct of the passing score study for research purposes was obtained by the American College Testing Program, Inc., and by the investigator in correspondence with the Executive Director of the medical specialty Board. Also, individual subjects were contacted by mail to request their participation in the study and each subject provided his consent.

Validity Concerns

For the medical specialty board under study, length of service on the Board is long, and changes in composition of its Written Examination Committee are slight from year-to-year. Also, all members of the standard setting body (n = 10) were included in the study. Thus, external validity within the medical specialty group is

substantial. External validity is weaker when viewed across medical specialty licensure and certification groups. However, similarities in the composition of, experiences, and roles played by other medical specialty groups suggests that results of the proposed research may be generalizable to other medical specialty groups as well.

A second validity concern also relates to the composition of the groups. In the present study, all subjects were male; the question of whether female item reviewers would respond differentially to the treatment (ie, the isolation condition) is not answered by the proposed research.

Internal validity concerns (previously discussed) are somewhat ameliorated due to the random assignment of five subjects to each of the two conditions (Treatment Group, n = 5; Control Group, n = 5). Two additional concerns exist, however. First, there is the possible effect of subjects' knowledge about the purposes of the study. Subjects in both the treatment and control groups were made aware of what condition the other group's members were exposed to. It is likely that the subjects were able, with that knowledge, to surmise the intent of the study. It is unknown, however, what effect such knowledge will have on the results of the present research, though no systematic bias in either group's item ratings is expected. Second, differences between treatment and control groups could be magnified (or depressed) as a result of the domination of discussion by one or more individual raters in the control group. For example, a dominant personality in the control group could influence the ratings of others such that ratings appear to be less variable that they would have been in the absence of the dominant personality. However, careful

attention was paid to this concern by the investigator, who moderated the group-process condition. Although discussions about the difficulty of individual items and the concept of the minimally-competent candidate were common in the control group, the discussions were participated in by all group members; no dominance of discussion or hegemony over ideas expressed was observed.

A final internal validity concerned is raised by the relatively small sample sizes involved in the study (n = 5 per group), Specifically, the power of the present study to detect true differences between the groups, should such differences exist, is only modest. Thus, if statistically significant differences between the groups are not observed, strong statements concerning the presence or absence of true differences cannot be made; that is, the hypotheses that true differences between the groups do not exist and that true differences were simply not detected (a type II error occurred) would remain equally tenable.

Instrumentation

Three instruments were used to record observations in the present research. First, a rating form to collect item reviewers' estimates of the proportion of minimally-competent examinees who will answer an item correctly was used. The same item rating collection form was used by both the treatment and control groups. A sample item rating collection form is reproduced in Appendix C.

The second instrument used was a questionnaire designed to elicit certain information from the item reviewers. Information on the following variables was desired:

- length of service on the Written Examination Committee
- type of professional practice setting (e.g., clinic, university, private practice).

Additionally, questions using Likert-type response choices (Likert, 1932) were asked, concerning:

- perceptions of the adequacy of training in the standard-setting methodology;
- perceptions of item reviewers' comprehension of the standard-setting methodology;
- perceptions of the ease of implementation of the standard-setting methodology; and,
- confidence that application of the standard-setting methodology would result in acceptable (accurate) separation of minimally-competent/not minimally-competent examinees.

Information from the questionnaire was gathered in order to obtain demographic characteristics of the content expert panel and to identify other variables that might be related to precision and variability in item ratings. The questionnaire was developed by the investigator following recommendations set forth in Babbie (1973) and Schaeffer, Mendenhall, and Ott (1979). The questionnaire is reproduced in Appendix D and was administered to both the treatment and control groups.

The third instrument used in the present research was the medical specialty examination itself. The examination is used by the medical specialty board as a component of its certification process. One form of the examination is administered annually to approximately 750 residency program graduates. The examination consists of 200

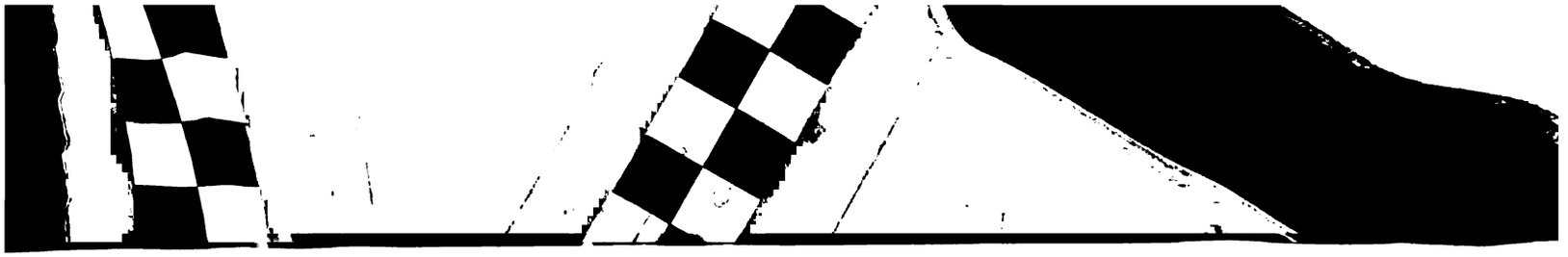
previously untested multiple-choice questions (types A and K) with five option choices. The examination is developed by the medical specialty board based on test specifications that include eleven subtest classifications. Previous analyses of the eleven subtest areas has revealed high subtest intercorrelations (some exceeding 1.00 when corrected for unreliability) suggesting a fairly unidimensional examination (Cizek, 1989). However, on this certification examination, examinees pass or fail the test based on their total test score only.

Previous administrations of examination forms have revealed the test to be quite reliable; KR-20 indices of internal consistency (Kuder & Richardson, 1937) for the past eight annual administrations of the test (1982-1989) have been .92, .93, .92, .92, .92, .92, .92, .92, respectively.

Statistical Analyses

The purpose of the statistical analyses employed in Experiment 1 was to identify any differences between the two groups that would be observable as a result of their exposure to the two conditions (group-process and isolation). Of primary interest is whether the conditions result in different passing scores. In each case, an individual item reviewer's passing score is defined as the sum of his ratings for each of the 200 items. The passing score for each condition is defined as the average of the passing scores for each of the reviewers in the condition. These definitions can be represented notationally as:

$$\bar{x}_{.jc} = \sum_{i=1}^{200} x_{ijc}$$



where:

$\bar{x}_{.jc}$ is the passing score for a reviewer j, in condition c;

x_{ijc} is the rating of item i by reviewer j in condition c;

i is the index for items (i = 1 ... 200);

j is the index for item reviewers (j = 1 ... 5); and

c is the index for conditions (c = 1, 2).

And:

$$\bar{x}_{..c} = \frac{\sum_{j=1}^{5} \bar{x}_{.jc}}{5}$$

where:

$\bar{x}_{..c}$ is the passing score for a condition, and

$\bar{x}_{.jc}$ is defined as above.

There also exists a mean rating for each item with each group, which is obtained by averaging the individual reviewers' ratings for the item. That is there exists, for an item, i, a mean rating across reviewers in a condition, represented by $\bar{x}_{i.c}$ such that:

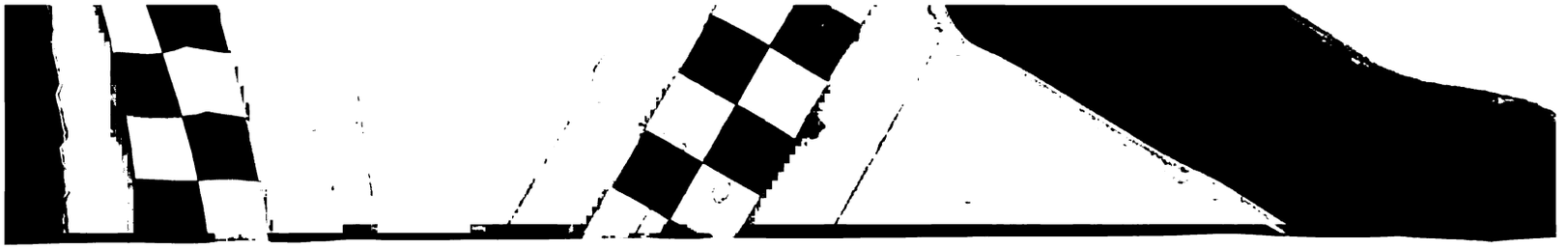
$$\bar{x}_{i.c} = \frac{\sum_{j=1}^{5} x_{ijc}}{5}$$

where: x_{ijc} again represents the rating of item i by reviewer j in condition c.

Investigation into possible effects on rating means and variance produced by exposure to the two conditions is of primary importance in the present research. Recall that the passing score for each condition is the sum of the averaged item ratings for each item from each reviewer assigned to that condition. A test for significant difference between the two condition means, $\bar{x}_{..1}$ and $\bar{x}_{..2}$ was performed using procedures for conducting a one-way analysis of variance (ANOVA) as outlined in Glass and Hopkins (1984). The test was conducted to determine if the treatment (isolation) condition resulted in a different passing score than that resulting from the group-process condition.

Although the primary practical interest of the research was in ascertaining whether there were between-group mean differences, the possible existence of within-group mean differences (ie, variation in passing scores assigned by individual reviewers) was also of interest. Specifically, do reviewers within a condition vary significantly in the individual passing scores they suggest? The mean passing score of each reviewer within conditions, that is, the five $\bar{x}_{.j1}$ and the five $\bar{x}_{.j2}$, were observed for within-condition mean differences using separate randomized block ANOVAs to test for such differences with reviewers' ($p = 5$) ratings blocked by items ($n = 200$).

The second question of primary interest was: Did assignment to the

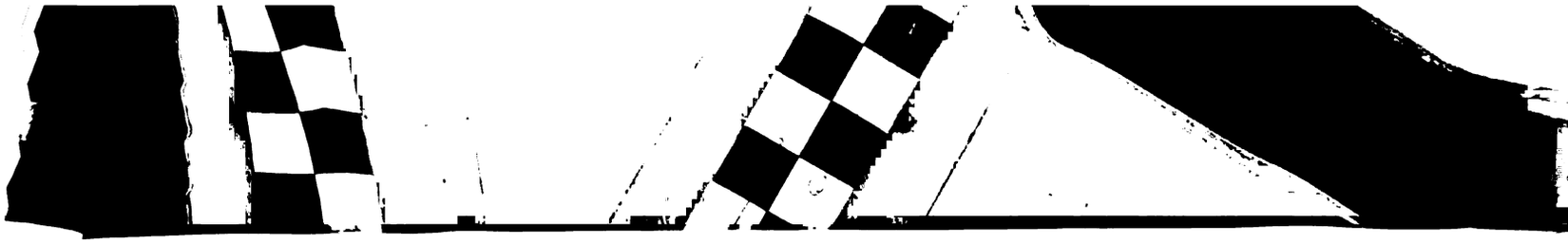


two conditions result in differential variability in reviewers' item ratings? A review of Appendix E shows that, across judges within conditions, variability in item ratings can be observed. This variability of ratings for an item, i, across raters in condition 1, can be represented notationally as $s^2_{i.1}$. Two columns of these item rating variances (one column for each condition) are shown in Table 1.

The results of the two randomized block ANOVAs (above) were combined for the next analysis. An F-test using the ratio of the two error variances was conducted to the likelihood of homogeneity of within-condition variances. The test also provided a means of answering the second question of primary interest: did assignment to the two conditions affect the within-block variability of reviewers' ratings?

Two correlation coefficients were also calculated on the condition mean item ratings (ie, on the $\bar{x}_{i.1}$ s and $\bar{x}_{i.2}$ s) to answer the question: Do the two methods of rating items (independent and group-process) produce similar orderings of item ratings? In this case, the Pearson product-moment correlation coefficient was calculated to assess the extent to which a linear relationship existed between the ratings of reviewers assigned to each condition. Also, the rank order correlation coefficient was calculated to obtain an indication of the extent to which the two conditions produce similar rankings of Angoff values.

An intercorrelation matrix of reviewers' ratings based on the 200-item set was also calculated. The intercorrelation matrix lends itself to 1) visual examination of the row entries, and 2) statistical



testing for differences between within-group mean correlations. For example, each row can be visually examined to verify the hypothesis that a reviewer's ratings should correlate more highly with other same-group reviewers' ratings than they should with the ratings of reviewers assigned to the other condition. More specifically, it was hypothesized that the mean of the group-process reviewers' intercorrelations should exceed that of the independent condition, primarily due to the sharing of information that occurs during the group-process condition. To address this hypothesis, a test for differences in the mean intra-group correlations was performed.

Two methods were used to evaluate the comparability of the two conditions using an additional source of data--the empirical item performance statistics from administration of the examination for which items were rated. The first evaluation was based on the extent to which the two conditions resulted in dependable classification (pass/fail) decisions. As the Standards for Educational and Psychological Testing (1985) state:

> "estimates of the reliability of licensure or certification decisions should be provided"... and "the reliability of the decision of whether or not to certify is of primary importance" (p. 65).

Two estimates of decision consistency were utilized, $\hat{p}_o$ and $\hat{k}$. These estimates of decision consistency, using randomly parallel tests, are elegantly defined by Millman (1979). Millman has characterized $\hat{p}_o$ as "the proportion of individuals classified the same way on each administration [of a test]" and he defines $\hat{k}$ as "the proportion of the total number of agreements [in classification] above the chance level of agreement" (p. 86). It is also possible to conceive of these two indices as an indicator of classification

consistency ($\hat{p}_o$), and an indicator of the relative contribution of the test to that level of classification consistency ($\hat{k}$). Procedures are available for obtaining estimates of decision consistency using only one form of a test, and these procedures were used in the present research. Detailed explications of the procedures have been provided by Huynh (1976) and Subkoviak (1976; 1984; 1988).

The second evaluation consisted of two ways of examining the relationship between reviewers' ratings and item statistics obtained from the actual administration of the examination. For one analysis, individual item reviewers' ratings for each item were compared with empirically-obtained difficulty indices (p-values) derived from the administration of the 200-item test. Modified p-values (symbolized p_i) were used for this comparison. The modification consisted of calculating the p-values based upon the performance of "minimally-competent" examinees only, rather than on the total group, following the suggestions of others (see, for example, Kane, 1984; 1986; DeMauro & Powers, 1990; Cramer, 1990). For this analysis, minimally-competent examinees were defined as those scoring within two standard errors of measurement of the operational passing score on the examination[2]. The analysis consisted of obtaining an indication of absolute error, or the extent to which reviewers' item ratings approximated the items' actual performance in the minimally-competent

---

[2] For p-values to be calculated based only upon the responses of the "minimally-competent" group, an external criterion was needed. That is, the minimally-competent group could not be established with reference to the passing standard based upon the Angoff ratings. For the examination under study, the actual operational passing standard was established using the Beuk (1984) methodology, thereby avoiding a circular definition of competence.

group. Following the conceptual framework suggested by others (van der Linden, 1982; Subkoviak & Huff, 1986; Friedman & Ho, 1990) the variable E was created to reflect error, or misspecification of item performance by the reviewers. Thus, the absolute root mean squared error (RMSE) of specification for a reviewer, j, in condition c, is represented by:

$$E_{.jc} = \sqrt{\sum_{i=1}^{200} (x_{ijc} - p_i)^2 / (n_i - 1)}$$

where x_{ijc} is the rating of item i by reviewer j in condition c

and p_i is the modified p-value for item i (described above).

A second analysis was conducted to obtain an indication of relative error, or the extent to which reviewers' ratings approximated group mean item ratings. Thus, the relative RMSE of specification for a reviewer j in condition c is given by:

$$E'_{.jc} = \sqrt{\sum_{i=1}^{200} (x_{ijc} - \bar{x}_{i.c})^2 / (n_i - 1)}$$

where the elements are defined as above. Overall, the absolute and relative error analyses were aimed at determining whether the two treatments differed in the extent to which they affected reviewers' approximations of mean group ratings or approximations of the actual performance of the minimally-competent examinee group.

Further analyses were conducted using E and E' as described above to determine if the accuracy[3] of item reviewers' ratings was

---

[3] The term "accuracy" is used somewhat inaccurately in this context. Strictly speaking, accuracy would apply more appropriately to a situation in which reviewers' estimates

significantly related to the variables identified in the Post-Meeting Questionnaire (i.e., with length of service on the examination committee, practice setting, and perceptions of the adequacy of training, perceptions of the adequacy of materials, and perceptions of the ease of implementing the Angoff methodology). For this analysis, an intercorrelation matrix was produced using each reviewer's responses to questionnaire items and average errors of specification (E and E') as input. The purpose of this analysis is straightforward, asking: Are any of the questionnaire variables significantly related to reviewer accuracy (as operationalized in E and E') or to each other?

Generalizability analyses were also conducted, following procedures set forth by Cronbach, Gleser, Nada, and Rajaratnam (1972) and Brennan (1983). Generalizability analyses were conceptually quite appropriate for the present research because of two primary characteristics. First, generalizability theory allows for differentiation between multiple sources of rating variation (error) in the two procedures studied. Further, it also allows for comparisons of the relative magnitude of the variances. This second characteristic is especially useful in designing subsequent measurement procedures of improved dependability under different combinations of items, raters, and procedures. For example, generalizability results might point to changing the number of item

---

could be compared to a known standard or to "truth." However, because the construct of minimal-competence is not a fixed, knowable parameter, it is not precisely correct to discuss accuracy in estimating the parameter. Thus, accuracy in this context refers not to the ability of item reviewers to approximate truth, but to their ability to approximate an admittedly weak proxy for it. For the sake of clarity, however, this relaxed usage of the term "accuracy" will be followed.

reviewers using one condition to achieve a level of dependability associated with the other condition. In that sense, the generalizability study (G-study) results using "real" data from the proposed research could complement the simulation studies done by Smith, et al (1989) regarding the optimal number of item reviewers for implementing a passing score methodology.

In the generalizability study (following notational conventions suggested by Brennan, 1983), the facets to be examined were those of items and raters. The generalizability study (G-study) design used is formally referred to as a completely crossed, random factor i x r design where:

i is the indicator for the item facet (random) and,

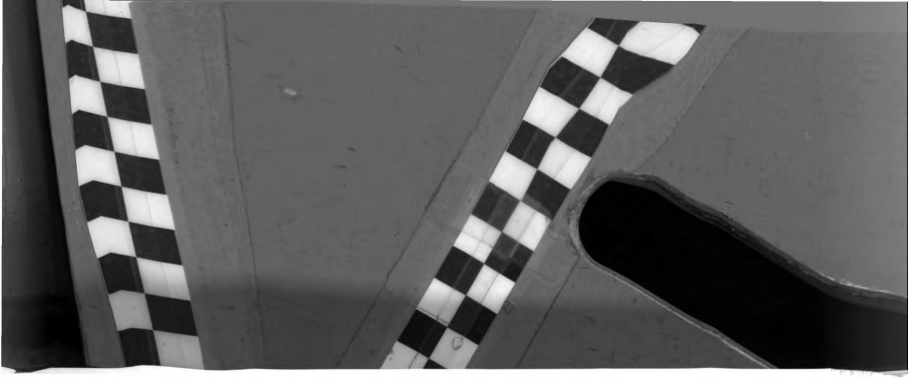
r is the indicator for the item rater facet (random).

In the i x r design, items are completely crossed with raters.

The G-study design employed yields three estimable effects (variance components):

- the effect associated with items, $\hat{\sigma}^2(i)$;
- the effect associated with raters, $\hat{\sigma}^2(r)$; and,
- the effect associated with an item by rater interaction, $\hat{\sigma}^2(ir)$.

The following generalizability analyses were of primary interest. First, examination of the relative magnitude of the variance components was performed to reveal the extent to which items, item reviewers, procedures, and interactions contribute to variability in



item ratings. The analysis of variance components associated with these facets sheds additional light on the primary question: Is the alternative methodology (i.e., the independent condition) a viable alternative to the traditional group-process format for gathering item ratings? For example, a comparatively large variance component associated with procedures would suggest that raters do assign different ratings to items depending on the procedure used (independent or group-process).

Secondly, two indices describing the magnitude of error associated with the item ratings were calculated. An index of absolute error variance is given by $\sigma^2(\Delta)$ which is the sum of all variance components except that for items, that is:

$$\sigma^2(\Delta) = \sigma^2(r) + \sigma^2(ir).$$

An index of relative error variance is given by $\sigma^2(\delta)$, which is the sum of all interaction variance components that contain the object of measurement index—in this case, the index for items—and is shown below. (Note that for the i x r design, only the ir interaction variance component contains the object of measurement index.)

$$\sigma^2(\delta) = \sigma^2(ir).$$

A general index of measurement dependability is given by $E\rho^2$, where:

$$E\rho^2 = \sigma^2(i) / (\sigma^2(i) + \sigma^2(\Delta)).$$

The generalizability analysis (G-study) also provided the estimates of variance components necessary for subsequent decision

studies (d-studies). One of the major goals of obtaining estimates of variance components in a G-study is so that "these estimates can be used to design efficient measurement procedures for operational use" (Brennan, 1983, p. 63). Thus, d-studies were designed to observe how reconfiguration of the measurement process could reduce error variance (and, accordingly, increase Ep^2) to some acceptable level. For example, a relatively large variance component associated with raters would support the recommendation that additional item reviewers be utilized in future uses of the independent condition to compensate for its comparatively reduced dependability and the number of additional reviewers necessary could be established. It would also be of substantial interest to learn, for example, that two additional reviewers in the independent condition would yield as dependable a standard setting procedure as in the group-process condition.

Finally, a cost analysis was performed. Because the independent condition is offered as a cost-effective alternative to the traditional group-process condition, an investigation of that hypothesis was warranted and of considerable interest. Specifically, the costs in terms of time, travel expenses, postal fees, and on-site expenses were examined for conducting the standard-setting methodology under each of the two conditions.

Experiment 2

In the first experiment, each item reviewer was exposed to one of two conditions--independent or traditional group-process. In the independent condition, item reviewers were able to draw on only internal sources of information--their own experiences, expertise, and

conceptions of a minimally competent examinee. On the other hand, those reviewers assigned to the group-process condition were influenced by both their own internal resources as well as the additional information obtained from the group interaction.

The information gleaned from the group interaction is, however, of two different types (Fitzpatrick, 1984; 1989). First, there are what can be called "pure" informational influences; that is, relevant information about the difficulty of the item, the appropriateness of the item, the conception of the minimally-competent examinee, etc., that is obtained from others in the group.

The second kind of information gained is less pure. Information of the social comparison type is less appropriate to the standard setting task. This kind of influence on the item ratings consists of individual's responses to pressure to conform to the opinions of others, for various social reasons. The second experiment attempted to assess the extent to which pure (or, relevant) informational influences affect the Angoff ratings of item reviewers.

Empirical Treatment

The design for the second experiment was analogous to the "One Group Pretest-Posttest Design" presented by Campbell and Stanley (1963, p. 7), and is symbolized as follows:

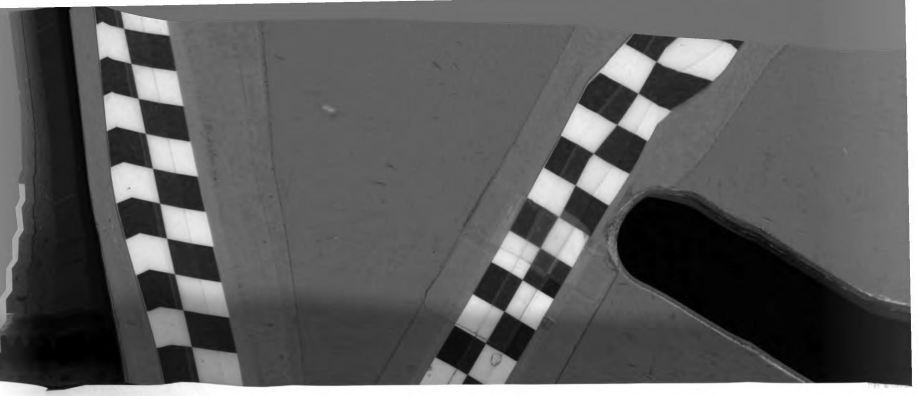
GROUP 1 - O_1 X O_2 [treatment group (independent condition)]

where:

O indicates and observation or data collection, and

X indicates administration of a treatment.

In the second study, subjects were the same (n = 5) item reviewers



who participated in the independent condition for the rating of items in Experiment 1. The first observation consisted of these subjects' ratings on the first 100 items from Experiment 1.

The treatment in Experiment 2 was defined as the presentation, to each item reviewer, of the distribution of item ratings from the initial rating of the 100 items. That is, each reviewer was provided with all five ratings (one from each reviewer, including himself) for each of the 100 items. Reviewers were then asked to review the distribution of ratings for each item and to provide a second rating for each of the 100 items.

Two precautions were taken. First, in the distributions of ratings presented reviewers, a reviewer's initial rating was not identified. Second, no other information, such as the mean rating for each item or item difficulty value, was given to the reviewers. These precautions were taken in order to (1) avoid the possibility that reviewers would automatically provide a second rating that was identical to their initial rating, and (2) to avoid the possibility that reviewers would simply select the group mean of the initial ratings for the second rating.

For the second experiment, all materials were mailed to the item reviewers. Each subject was given the same informational packet provided for the initial rating collection, a booklet containing the test items to be rated, and a form on which to record his ratings. The booklet contained the same first 100 items from the initial rating task. Subjects were again asked to provide Angoff ratings for the set of items, and were instructed not to consult with other group members about the ratings or to consult other informational sources,

such as textbooks, resident physicians, etc.

Validity Concerns

The most prominent validity concerns associated with conducting the second experiment center on aspects of internal validity. Although several threats to valid inferences of cause and effect are controlled for in the experimental design, two threats to valid inferences warrant attention.

First, testing is a concern. The initial ratings provided by the item reviewers represented the first attempt for each reviewer at application of the Angoff methodology. The second set of ratings collected as part of Experiment 2 represent the reviewers' second experience with the methodology. (No reviewer used the methodology between the data collection for Experiment 1 and the data collection for Experiment 2.) It is possible that experience with the methodology could account for some change in the second item ratings. It is believed, however, that the extensive training and practice with the methodology prior to collection of the initial ratings was sufficient and, therefore, little subsequent change would take place simply due to experience with the methodology.

Secondly, as with Experiment 1, statistical conclusion validity is of some concern. The modest sample size employed in the second experiment means that a failure to detect significant changes in the group's item ratings could be attributable either to the possibility that no true change occurred or to the possibility that a type II error was committed.

Instrumentation

The same instrumentation was used in Experiment 2 to collect the second set of ratings from item reviewers as was used to collect their ratings in Experiment 1. In Experiment 2, however, the rating collection forms were shortened to accommodate the ratings for 100 items and the original five ratings for each margin were printed on the form next to each item number. A sample rating form for Experiment 2 showing hypothetical distributions of item ratings is provided in Appendix F.

Statistical Analyses

Statistical analyses of the data collected for Experiment 2 were aimed at identifying any differences in reviewers' ratings resulting from exposure to the treatment. Of primary interest was whether the provision of information (the distributions of ratings) to item reviewers affected the reviewers' subsequent ratings.

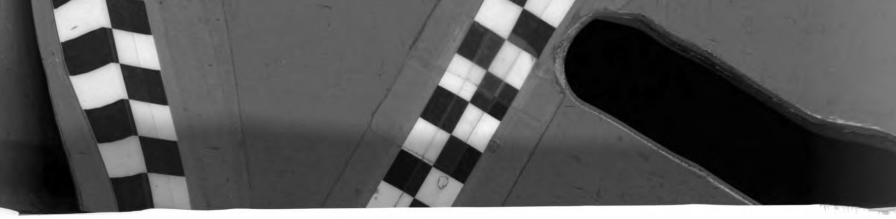
Analytic methods used on the data collected for Experiment 2 were similar to those use to analyze the data from Experiment 1. As in Experiment 1, primary interest centered on any observable differences in mean ratings and on the possibility of differential variability in ratings as a result of the additional information provided to item reviewers. An investigation into possible differences in variability was undertaken using a test for equality of variances in paired observations as described in Glass and Hopkins (1984, pp. 268-269). These results addressed the question of whether the provision of additional information to item reviewers has a variance reducing effect, a polarizing (variance increasing) effect, or no effect on

variation in the overall item ratings.

A repeated measures analysis of variance was also performed (two observations per subject). The repeated measures design treated items as the objects of measurement (n = 100), with raters (n = 5) being observed different conditions (n = 2). The analyses presented in the next chapter focus on the main effects of conditions and raters. A significant effect of conditions would point to the fairly unambiguous conclusion that exposure to the two conditions produces different passing scores. A significant effect of raters (across conditions) would similarly indicate that different raters produce different passing scores. A rater by condition interaction effect was also tested and a plot of the five raters' means under each of the two conditions was constructed to help interpret the interaction effect. In summary, these analyses should provide information identifying which factors contribute to differences in item ratings under the two conditions--independent/no-information and independent/with-information.

Further investigation into the effects of the treatments on resulting passing scores parallel the analyses proposed in Experiment 1. For example, the Pearson product-moment correlation and rank order correlation coefficients were computed to observe the relationship between item ratings under the "no-information" and "with-information" conditions.

Decision consistency analyses were also performed to help ascertain whether the provision of additional information actually results in any improvement in the categorical classifications (ie, pass/fail decisions) that are made on certification and licensure



examinations. As in the first experiment, the two procedures were evaluated using consistency of the classification decisions as the criterion. The classification consistency indices of $\hat{p}_o$ and $\hat{k}$ were calculated and compared for the two conditions.

As in Experiment 1, absolute error of specification (E) was used as the criterion to assess whether the information or no-information condition resulted in more accurate approximations by the reviewers of the empirical (modified) p-values. Relative error of specification (E') was also examined, comparing ratings under each condition to the condition mean ratings for each item.

Finally, to discern the extent to which item reviewers operate under an implicit "opinion revision" model, a regression analysis was conducted. The linear model for the regression equation was:

$$y_{ij} = a + B_1 x_{ij} + B_2 \bar{x}_{i.} + e$$

where: y_{ij} is the revised (2nd) rating for an item i by a reviewer j in the independent group;

x_{ij} is the original (1st) rating for item i by reviewer j;

$\bar{x}_{i.}$ is the original group mean rating (across reviewers for item i;

a is a constant; and,

e is an error term.

The opinion revision model postulates that the eventual (with information) item rating for a reviewer is predicted from knowledge of the reviewer's original rating (x_{ij}) and the effect of knowledge of

the group's original ratings ($\bar{x}_{i.}$). Such a model is useful because it further illuminates the effect of pure informational influences on reviewers' subsequent item ratings.

IV. RESULTS

Experiment 1

Between-group Mean Differences

Of primary interest in Experiment 1 was whether exposure to the two conditions (i.e., the independent rating of items or the use of the group-process method) resulted in differing overall passing standards. Table 2 provides descriptive statistics comparing the ratings produced under the two conditions and Figure 1 provides a plot of the group and independent reviewers' means.

Visual inspection of the individual reviewer means listed in Table 2 suggests some interesting observations. First, each conditions apparently contains one or more outliers. For example, while the reviewer means and standard deviations for the independent condition appear to be fairly similar (High to Low range of means equals 11.00) the variability of Reviewer 5's ratings is quite large compared to the rest of the reviewers in the independent condition. Similarly, in the group-process condition, Reviewer 10 produced an overall mean rating that was substantially lower that the other group-process condition reviewers. Of note also is that the variability of Reviewer 6's ratings is somewhat greater that the other reviewers in his group, although still not as large as the variability exhibited by Reviewer 5. Interestingly, Reviewer 10, who produced the lowest overall

Table 2

Descriptive Statistics for Independent and Group-Process Reviewers Across 200 Items

Independent Condition				Group-Process Condition			
Reviewer	Mean	Standard Deviation	Skew	Reviewer	Mean	Standard Deviation	Skew
1	60.23	17.67	-.255	6	45.57	21.91	.646
2	49.23	17.96	.252	7	60.68	16.56	-.561
3	57.23	16.66	-.201	8	64.13	15.24	-.771
4	51.18	17.26	.019	9	50.55	17.61	.140
5	58.79	26.55	-.217	10	34.13	14.41	1.373
Means	55.33				51.01		

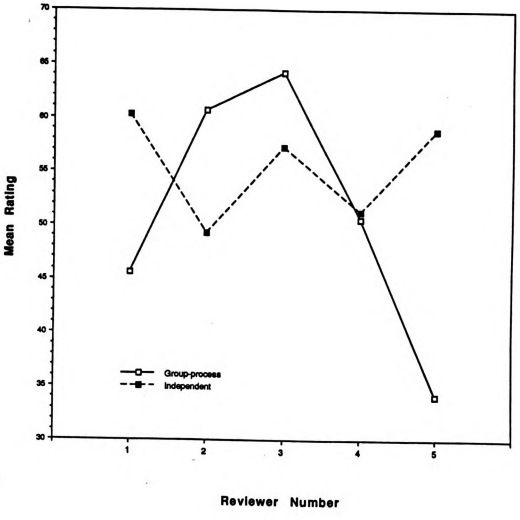


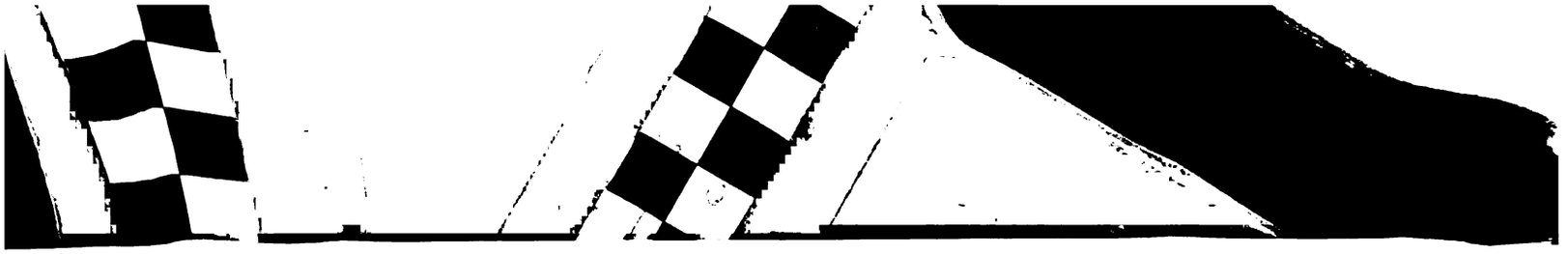
Figure 1

Plot of Independent and Group-Process Condition Reviewers' Means

rating, also produced some disproportionately low ratings within his own distribution of ratings, as evidenced by his relatively large positive value for skewness (1.373).

As a first step in exploring for possible differences between the two conditions, two variables were created (INDEPRATE and GROUPRATE) to represent the overall rating for each item within each condition. INDEPRATE represents the overall rating for each of the 200 items provided by reviewers in the independent condition. GROUPRATE represents the overall rating for each of the 200 items provided by reviewers in the group-process condition. In each case, INDEPRATE and GROUPRATE were obtained by calculating the mean rating for each item across reviewers within the respective conditions. This procedure resulted in 200 pairs of ratings (one pair for each item). A correlation between the 200 pairs of overall ratings provided under each condition (i.e., between INDEPRATE and GROUPRATE) was also calculated and found to be .712, which was significantly different from zero at $p < .001$. Substantively, the magnitude of the correlation seems to indicate that the reviewers in the two conditions tended to agree on the overall ratings for items.

The overall condition means calculated for the group-process condition and the independent conditions were 51.01 and 55.33, respectively. It should be noted that these values represent a proposed passing score for each condition, expressed as a percentage. That is, application of the standard proposed by reviewers in the group-process condition would result in a passing percentage of approximately 51% compared to the approximately 55% correct standard that would result from application of the standard based on the



independent ratings. This means, in terms of raw score units, that the independent condition mean of 55.33% correct would require examinees to respond correctly to 111 items in order to pass the examination, whereas the passing standard suggested by the group-process condition (51.01%) would be only 102 items correct. It is fairly obvious then, that regardless of whether the difference in the condition means proves to be statistically significant, the nine-point raw score difference in passing standards suggested by the two conditions is clearly of practical significance.

A test for a statistically significant difference between the two overall condition means was conducted using a one-way analysis of variance (ANOVA). The results of the significance test are presented in Table 3. Despite the practical significance of the difference between the two suggested passing standards, the results of the analysis of variance failed to reveal a statistically significant difference between the two passing scores. The F test for significant mean differences between the overall reviewer passing score means was nonsignificant, with $F(1,8) = 0.55$.

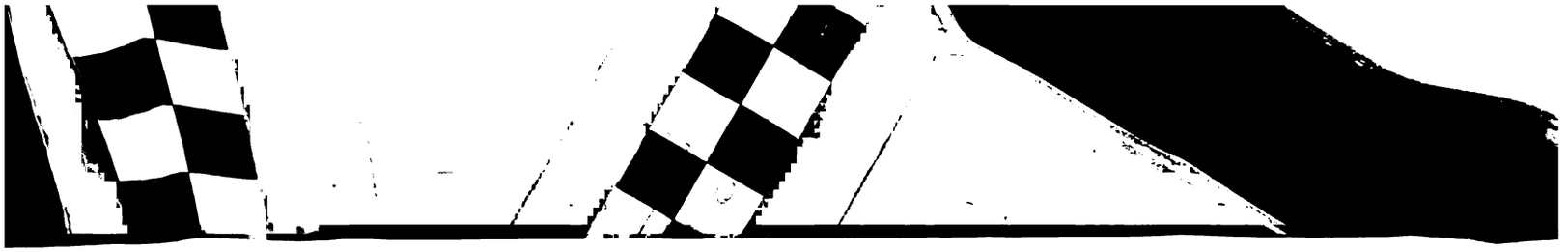
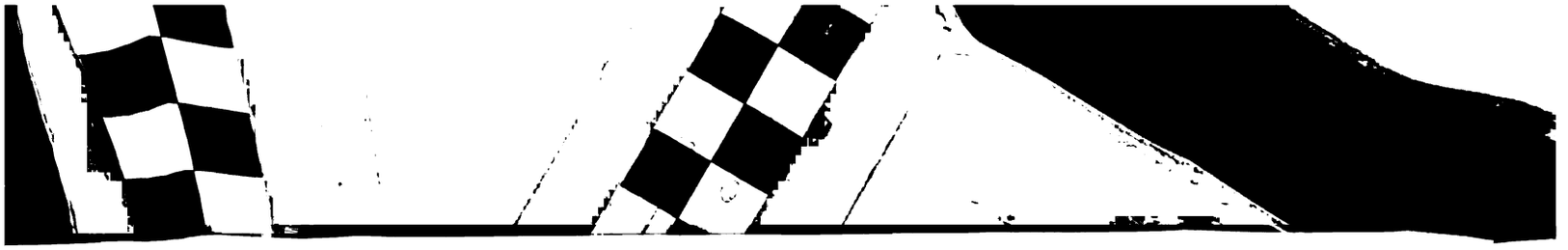


Table 3

Test for Significant Mean Differences between Independent and Group-Process Condition Passing Scores

	INDEPRATE	GROUPRATE
Mean	55.33	51.01
Standard Deviation	12.81	11.78
r INDEPRATE, GROUPRATE =	.712	$p < .001$
Mean Difference =	4.318	
Standard Deviation of Overall Reviewer Means	4.85	12.05

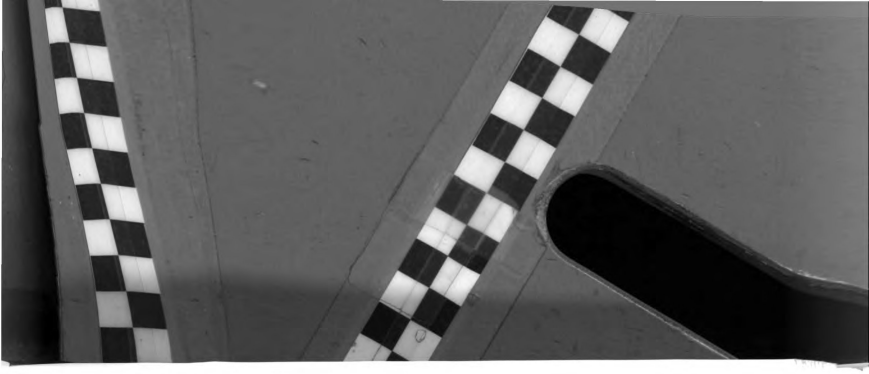
Source	Mean Square	df	F
Conditions	46.70	1	0.554 ns
Error	84.30	8	—
Total	80.12	9	—



It is interesting at this point to note that the presence of profound practical significance in the absence of statistically significant differences is a somewhat uncommon finding. It is regularly observed that statistically significant findings can be of little practical importance (see, for example, Glass and Hopkins, 1984, p. 215-216). In the present study, however, although the mean difference in condition passing standards was not statistically significant, a substantial effect on pass/fail classifications could result from even small differences in the suggested passing standards.

It should be noted that while the nonsignificant F-test result does not indicate that the treatment was ineffective, neither can the presence of practical significance be confidently attributed to the experimental treatment. The differences between the treatment and control groups could be due to the treatment, although because of the small sample sizes used, that claim cannot be substantiated based upon this experiment. In other words, the observed differences could be due solely to random error and different groups of item reviewers could produce different results.

The extent of classification changes that would be seen if the independent and group-process standards were applied to the actual distribution of scores observed for this examination was explored. Application of the group-process condition standard (approximately 102 items correct) would have resulted in a passing rate of 93.0% and a corresponding failure rate of 7.0%. On the other hand, had the independent condition standard been applied (requiring approximately 111 items correct), the passing rate would have been 85.8% and the failure rate (14.2%) would have nearly doubled compared to the



independent condition failure rate.

Finally, to ascertain whether overall item rating variances for the two groups were homogeneous, an F-test using the ratio of the variances of INDEPRATE and GROUPRATE was performed. In this case, the ratio of the variances was 1.18, which did not exceed the critical value of 1.32 for alpha = .05 with df = 199 numerator and denominator. This finding suggests that the overall ratings were not more variable under either of the two conditions.

Within-group Differences

Separate randomized block ANOVAs [with n = 200 items (random) and n = 5 raters (random)] were performed to learn if the overall ratings of individual reviewers within a condition differed significantly from each other. Additionally, the results of the two ANOVAs were used to address the question of whether exposure to either the independent or group-process condition affects the variability of reviewers' ratings. Plots of the group and independent raters' overall means are provided in Figure 1 and ANOVA results are presented in Table 4. Inspection of Figure 1 reveals that raters in both conditions were variable in their ratings. The results of the two randomized block ANOVAs presented in Table 4 present a similar picture. The ANOVAs reveal a significant effect for raters in both the group-process ($F_{4,796} = 143.12$, $p < .001$) and independent ($F_{4,796} = 17.17$, $p < .001$) conditions, indicating that raters within a condition do produce different ratings (passing standards). As would be expected, the effect of items was also significant in both the group and independent conditions.

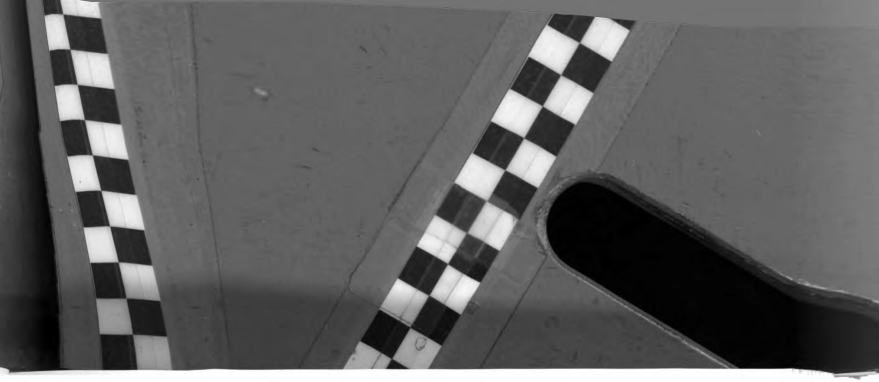
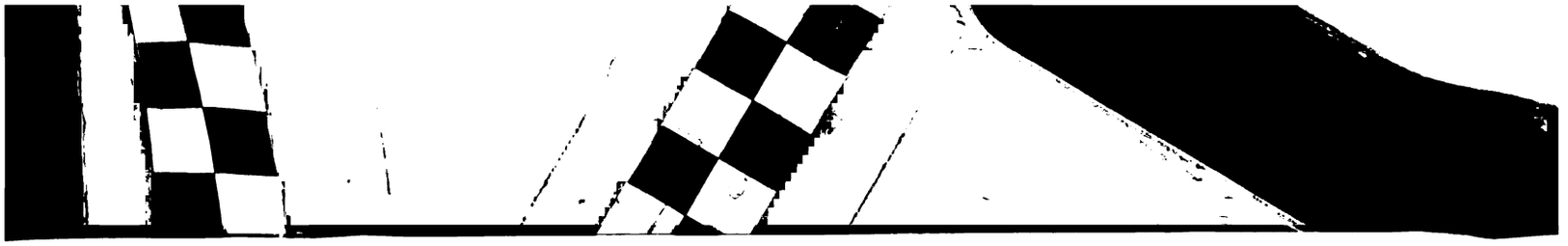


Table 4

Randomized Block ANOVA Results for Independent and Group-Process Conditions

	Independent Condition			Group-Process Condition		
Source	Mean Square	df	F	Mean Square	df	F
Items (Blocks)	820.03	199	2.99*	693.64	199	3.42*
Raters	4701.19	4	17.17*	29016.19	4	143.12*
Residual	273.79	796		202.74	796	
Total	5795.01	999		29912.57	999	

* = $p < .001$



A second variability issue was addressed using the results from the two randomized block ANOVAs. Specifically, it was of interest to learn whether exposure to the two conditions lead to more variable within block ratings. To assess this effect, the ratio of the residual variances from each ANOVA were compared. Although the residual variances actually contain two sources of error (interaction effect plus error), a non-significant finding would lead to the inference that a within-block effect was absent.

Because it was hypothesized that the independent condition might lead to greater within-block variability, the variance estimate from the independent condition was chosen as the numerator for the F-test. The test revealed a significant F ratio ($F_{796,796} = 1.35$, $p < .001$), indicating that the hypothesis of increased within-block variability for the independent condition remains tenable.

Relationship between Group and Independent Ratings

As reported earlier in Table 3, a significant Pearson product-moment correlation between group-process condition and independent condition overall item ratings was observed (r INDEPRATE, GROUPRATE = .712, $p < .001$). Calculation of the rank order correlation coefficient yielded similar results (r INDEPRATErank, GROUPRATErank = .702, $p < .001$). These results indicate that the group-process and independent conditions provided overall item ratings of moderately corresponding linear relationship and rank.

The intercorrelation matrix of all reviewers' item ratings was also produced and is presented in Table 5. Visual inspection of the correlations did not immediately lend support to the hypothesis that a

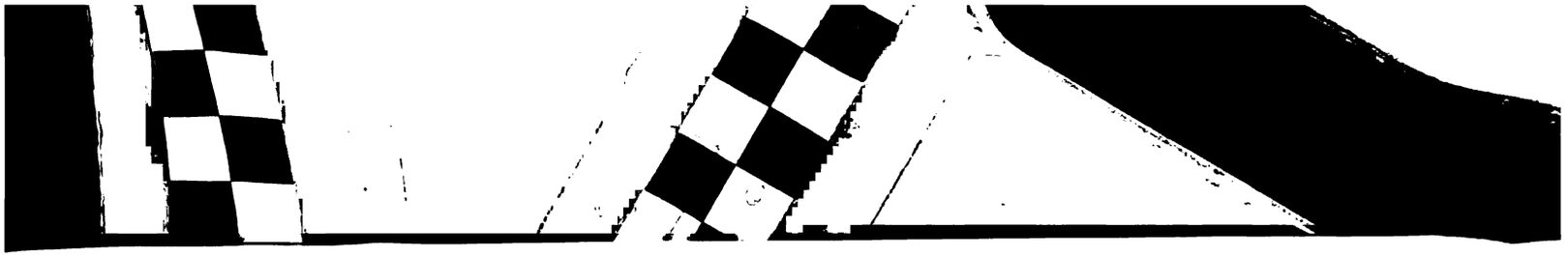
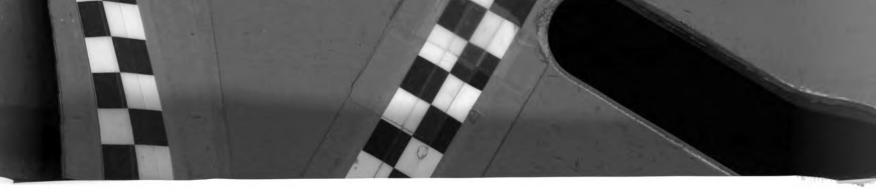


Table 5

Intercorrelation Matrix of Ratings from Independent and Group-Process Condition Reviewers

	Group-Process Condition Reviewers					Independent Condition Reviewers				
	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
R1	—	.339	.365	.289	.264	.425	.336	.451	.283	.328
R2		—	.243	.258	.253	.331	.290	.247	.172	.292
R3			—	.346	.225	.398	.347	.413	.317	.443
R4				—	.421	.301	.306	.351	.233	.351
R5					—	.299	.345	.175	.353	.414
R6						—	.371	.337	.296	.446
R7							—	.270	.308	.340
R8								—	.275	.388
R9									—	.320
R10										—

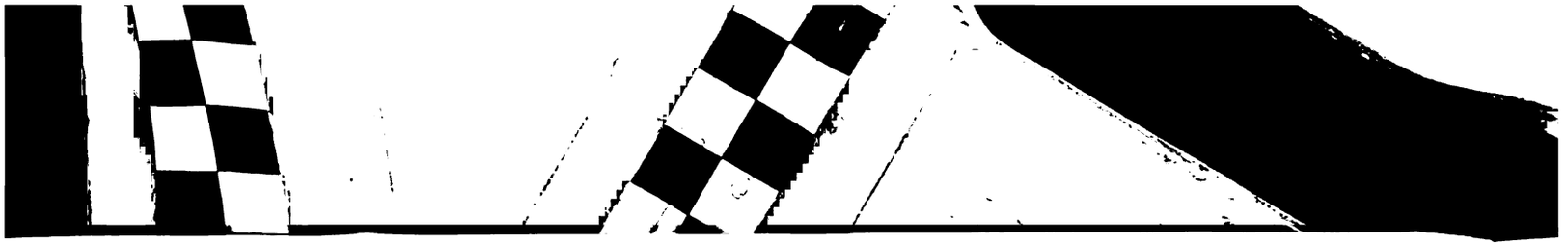


reviewer's ratings would generally correlate more highly with other within-condition reviewers' ratings than they would with ratings provided by reviewers assigned to the other condition. In fact, the correlations of highest and lowest magnitude were observed between conditions (highest: r 1,8 = .451; lowest: r 2,9 = .172). However, all correlations were significantly different from zero ($p < .01$).

To test for significant differences between the average within-condition correlations, the within-condition correlations (enclosed by dashes in Table 5) were first transformed using Fisher's r to Z transformation. After transformation, a mean correlation for each condition was calculated and a test for significant difference between the mean independent and group-process correlations was conducted. Although the mean correlations differed (mean group-process condition correlation = .348; mean independent condition correlation = .311) the difference was nonsignificant.

Decision Consistency

The extent to which exposure to the group-process condition and exposure to the independent condition results in differing levels of classification consistency was also examined. Indices of classification consistency, $\hat{p}_o$ and $\hat{k}$, were calculated and are presented in Table 6. As the table shows, the group process condition exhibited a slightly greater index of overall consistency than the independent condition ($\hat{p}_o$ = .958 and .930, respectively). However, it should be noted that the contribution of the examination itself to consistency of classification decisions was slightly reduced under the



group-process condition as compared to the independent condition ($\hat{k}$ = .647 and .681, respectively).

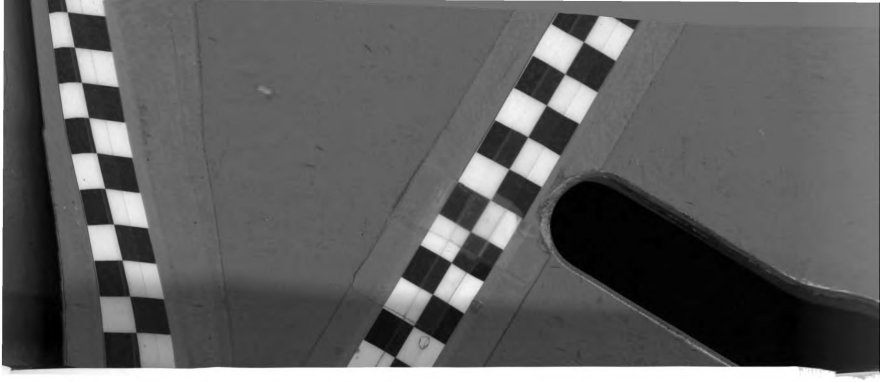
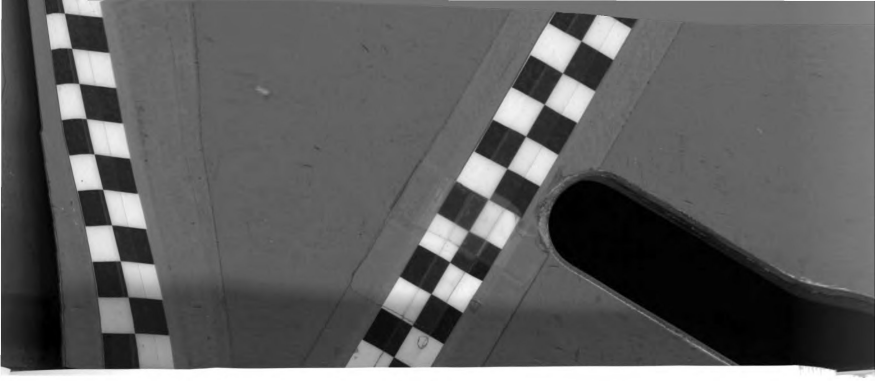


Table 6

Indices of Decision Consistency for Independent and Group-Process Conditions

Condition	Suggested Passing Score	$\hat{P}_o$	$\hat{\kappa}$
Independent	111	.930	.681
Group-process	102	.958	.647

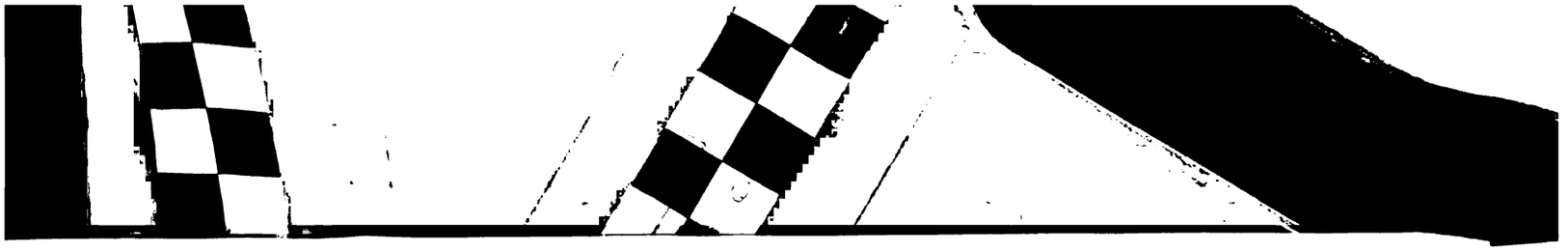


Relationship to Obtained Item Statistics

For reviewers in the group-process and independent rating conditions, overall item ratings for each of the 200 items were compared to the item difficulty values obtained from actual administration of the examination. For all analyses comparing the reviewers' ratings to obtained difficulty indices, however, modified p-values (MODP) were used. The modification consisted of calculating the p-values based only of the responses of examinees (n = 217) whose total score was within two standard errors of the operational passing score.

First, correlations were calculated between the overall independent and group-process condition ratings (INDEPRATE and GROUPRATE) and MODP. Correlations were also calculated for individual item reviewer's ratings and MODP. For both conditions, individual reviewer's ratings were found to be only weakly related to the modified p-values. All individuals' correlations with MODP were significantly different from zero at $p < .001$, but ranged only from a low of .306 (for independent condition Reviewer 2) to a high of .423 (for group-process condition Reviewer 5). Overall condition item rating correlations with MODP were only somewhat greater. The correlation for the overall group-process condition ratings and modified p-values was slightly but not significantly lower than the correlation between independent condition ratings and modified p-values ($r = .535$ and .544, respectively).

Two indices were also created to reflect the degree of agreement between reviewers' ratings and two important criteria. The first variable, E, was created to reflect the extent of agreement between a



reviewer's ratings and the modified p-values. The variable E can be conceptualized as an index of absolute error of specification. The second variable, E', reflects the degree of agreement between a reviewer's ratings and the mean ratings provided by reviewers within a particular condition. Computationally, E' for a reviewer j in condition c was calculated by averaging the ratings of reviewers except reviewer j within condition c. This variable, E', can be conceptualized as an index of relative error of specification.

Table 7 presents the obtained values of E and E' for the five reviewers in each condition. Examination of the table leads to some interesting observations. First, comparison of the values of E and E' across conditions indicates that, in general, item reviewers exhibited disconcertingly large errors of specification, although it is clear that they were much better at estimating how other reviewers in their condition would rate items than they were at predicting how the hypothetical minimally-competent group would perform. Second, when

Table 7

Absolute and Relative Errors of Specification for Item Reviewers in Independent and Group-Process Conditions

Independent Condition			Group-Process Condition		
Reviewer	E	E'	Reviewer	E	E'
1	23.46	14.31	6	28.10	15.99
2	26.42	15.58	7	23.49	15.74
3	23.72	13.51	8	24.56	17.75
4	24.53	13.26	9	25.33	13.47
5	27.90	19.59	10	32.37	19.86
Mean	25.21	15.25		26.77	16.56
Standard Deviation	1.90	2.59		3.57	2.39

evaluating the overall performance of the two conditions, it appears that the independent condition results in slightly improved accuracy of specification in both the absolute and relative sense.

Relationship between E and E' and Reviewer Characteristics

Responses provided by the 10 item reviewers on the Post-Meeting Questionnaire were used to help assess whether any background variable or characteristic of the item reviewers was associated with increased precision in estimation of item ratings. Correlations were calculated between background variables (obtained from the Questionnaire) and the two indices of error of specification (E and E').

For the total group of reviewers employed in the study, no background variable or reviewer characteristic appeared to be related to accuracy in specification. Correlations between all variables including years of service on examination committee, perceptions of helpfulness of the informational materials, perceptions of ease of implementing the passing score methodology, etc., and error of specification variables, E and E', were all small and not significantly different from zero, ranging from, -.401 to .026 with most of the correlations near zero.

This result is at once discouraging and unsuprising. On the one hand, the result indicates that the study did not isolate any relevant reviewer characteristics that might help predict which reviewers would produce the most accurate item ratings. On the other hand, this finding also provides some evidence that the homogeneous backgrounds and perceptions of the panel of reviewers utilized were not likely to have contributed to the variability in item ratings.

Generalizability Analyses

Two separate completely crossed (items x raters) generalizability analyses were performed, one each for the group-process and independent rating conditions. A summary of the G-study results is presented in Table 8. The table shows that the variance components for items and the variance components for the items by raters interaction are somewhat similar in magnitude across conditions and similar to each other, indicating that each of these two factors contributes roughly equally to the dependability of the measurement process (i.e., to the rating of items). In general, standard errors for the Items and Items x Raters variance components were relatively small, indicating that they were fairly well estimated and that subsequent d-study analyses are likely to be fairly accurate. Less confidence in the precise estimation can be assigned to the variance components associated with raters in the presence of the relatively large standard errors for these components.

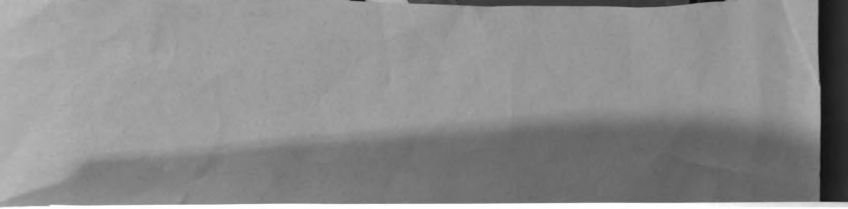


Table 8

Summary of Generalizability (G-study) Results for Independent and Group-Process Conditions

	Independent Condition			Group-Process Condition		
Effect	df	Variance Component	Standard Error	df	Variance Component	Standard Error
Items	199	109.25	16.59	199	98.18	13.99
Raters	4	22.14	13.57	4	144.07	83.76
Items x Raters	796	273.79	13.71	796	202.74	10.15

Of particular practical note, however, is the relatively large variance component (and associated large standard error) for raters in the independent and group-process conditions. The variance component for raters in the independent condition is larger than is normally desirable, relative to the variance component for items (about one-fifth its magnitude). The variance component for raters in the group-process condition is extremely large (144.07) relative to all of the other group-process variance components, exceeding even the magnitude of the variance component for items (98.18). Taken together, the magnitude of these variance components suggests that, were the independent or group-process item rating processes to be utilized in future studies, better rater training or an increase in the number of raters would be profitable in terms of improved dependability.

Table 9 presents a summary of d-study results. The entries in the table include estimated variance components for five to 20 raters for each of the two conditions. As one would expect, the variance components associated with raters and raters by items interaction decrease as the number of raters increases (with the number of items held constant). Absolute error variance [$\sigma^2(\Delta)$], relative error variance [$\sigma^2(\delta)$], index of dependability (Φ^2), and signal-to-noise ratio are also presented in the table. Absolute error variance represents the variance of the difference between the unit of measurement's observed and "true" scores or, in this case, the variance of the difference between the observed and "true" item ratings. Relative error variance is another type of error variance, whose magnitude depends of the differences between observed and true score variances relative to population means for observed and true

Table 9

Summary of Generalizability Analyses (d-Study) Results

Independent Condition Results

No. of Raters	V.C. Items	V.C. Raters	V.C. I x R	Absolute Error Variance	Relative Error Variance	S/N Ratio	Index of Depend.	Mean Error Variance	S.E. Mean
5	109.25	4.43	54.76	59.19	54.76	2.00	.649	5.25	2.29
6	109.25	3.69	45.63	49.32	45.63	2.39	.689	4.46	2.11
7	109.25	3.16	39.11	42.27	42.28	2.79	.721	3.90	1.98
8	109.25	2.77	34.22	36.99	34.22	3.19	.747	3.48	1.87
9	109.25	2.46	30.42	32.88	30.42	3.59	.769	3.16	1.78
10	109.25	2.21	27.38	29.59	27.38	3.99	.787	2.90	1.70
11	109.25	2.01	24.89	26.90	24.89	4.39	.802	2.68	1.64
12	109.25	1.84	22.82	24.66	22.82	4.79	.816	2.51	1.58
13	109.25	1.70	21.06	22.76	21.06	5.19	.828	2.35	1.53
14	109.25	1.58	19.56	21.14	19.56	5.59	.838	2.22	1.49
15	109.25	1.47	18.25	19.73	18.25	5.99	.847	2.11	1.45
16	109.25	1.38	17.11	18.50	17.11	6.38	.855	2.02	1.42
17	109.25	1.30	16.11	17.41	16.11	6.78	.863	1.93	1.39
18	109.25	1.23	15.21	16.44	15.21	7.18	.869	1.85	1.36
19	109.25	1.17	14.41	15.58	14.41	7.58	.875	1.78	1.34
20	109.25	1.11	13.69	14.79	13.03	7.98	.881	1.67	1.29

Table 9 (Cont'd.)

Summary of Generalizability Analyses (d-Study) Results

Group-Process Condition Results

No. of Raters	V.C. Items	V.C. Raters	V.C. I x R	Absolute Error Variance	Relative Error Variance	S/N Ratio	Index of Depend.	Mean Error Variance	S.E. Mean
5	98.18	28.81	40.55	69.36	40.55	2.42	.586	29.51	5.43
6	98.18	24.01	33.79	57.80	33.79	2.91	.629	24.67	4.98
7	98.18	20.58	28.96	49.54	28.96	3.39	.665	21.22	4.61
8	98.18	18.01	25.34	43.35	25.34	3.87	.694	18.63	4.32
9	98.18	16.01	22.53	38.53	22.53	4.36	.718	16.61	4.08
10	98.18	14.41	20.27	34.68	20.27	4.84	.739	15.00	3.87
11	98.18	13.09	18.43	31.53	18.43	5.33	.757	13.68	3.70
12	98.18	12.01	16.90	28.90	16.90	5.81	.773	12.58	3.55
13	98.18	11.08	15.60	26.68	15.60	6.29	.786	11.65	3.41
14	98.18	10.29	14.48	24.77	14.48	6.78	.799	10.85	3.29
15	98.18	9.60	13.52	23.12	13.52	7.26	.809	10.16	3.19
16	98.18	9.00	12.67	21.68	12.67	7.75	.819	9.56	3.09
17	98.18	8.47	11.93	20.40	11.93	8.23	.828	9.03	3.00
18	98.18	8.00	11.26	19.27	11.26	8.72	.836	8.55	2.92
19	98.18	7.58	10.67	18.25	10.67	9.20	.843	8.13	2.85
20	98.18	7.20	10.14	17.34	10.14	9.69	.850	7.74	2.78

scores. The index of dependability (sometimes also referred to as a generalizability coefficient) in this case takes into account the between-rater variance component, since systematic bias in reviewers' ratings would contribute error to the estimation of an item's "true" rating. The signal-to-noise ratio is the ratio of true score variance to error variance and provides another "reliability-like" index. For example, a signal-to-noise ratio of 2.00 would indicate that the estimated universe score variance is twice as large as the estimated error variance.

Both the group-process and independent conditions appear to produce moderate signal-to-noise ratios and indices of dependability with as few as five raters. Of significant interest, however, are the relatively large variance components for raters in the group-process condition and the correspondingly larger absolute error variance. Interestingly, due to the relatively smaller rater variance component associated with the independent condition, higher overall indices of dependability for the independent condition compared to the group-process condition were observed. A practical application of this relationship means, for example, that to achieve the same level of dependability found with only eleven raters in the independent condition (approximately .80) would require 14 raters using the group-process procedure.

Table 9 also includes the mean error variance and the standard error of the mean for five to 20 raters in each of the two conditions. These values provide information on how well the passing score is estimated across the samples of items and raters. The independent condition rating procedure appears to result in a more precise estimate

of the passing score than the group-process ratings; for a 200-item test, the standard error of the mean indicates that the confidence interval around the passing score would vary from about three to five items for the independent condition (.0129 x 200 = 2.58; .0229 x 200 = 4.58) but from six to 11 items for the group-process condition (.0278 x 200 = 5.56; .0543 x 200 = 10.86).

In evaluating these results, a standard error of the cutting score of less than two items was used as a minimally acceptable standard, following criteria suggested by Norcini, et al. (1987). Applying this criterion, only the ratings generated using the independent condition with at least seven raters would approach acceptability for standard setting procedures (see Table 9). However, for both conditions, signal-to-noise ratios and indices of dependability tended to be smaller than would be desirable for standard setting.

Cost Analysis

As Norcini, et al. (1987) and Lockwood, et al. (1986) have noted, factors other than psychometric concerns can influence decisions regarding the conduct of passing score studies. Specifically, the cost of empaneling item reviewers (SMEs) may be prohibitive in many cases. Because differing costs would likely be associated with implementation of the group-process or independent conditions, an examination of the relative costs for each condition was undertaken.

For the following cost analyses, several assumptions were made. First, it was assumed that a group of SMEs (n = 10) were to be empaneled to provide ratings for a 200-item examination. For analysis of the group-process condition, it was assumed that nine of the ten

reviewers would incur air travel, lodging, and meal expenses in order to travel to the passing score study site and participate in an item review procedure lasting two days.

Two variations of the independent rating condition were explored in the analysis, in addition to the group-process condition. For one variation (hereafter called the "without-meeting" condition), it was assumed that the panel of reviewers would be mailed informational materials explaining the passing score methodology, then the test items to be reviewed would be rated independently and returned by mail to a central site. The second variation of the independent condition (hereafter called the "with-meeting" condition), assumes that reviewers would travel to a single site for a one-half day meeting in order to become familiar with the passing score methodology. Reviewers in this condition would then receive a booklet of test items to be rated, would return to their cities of origin, and would return their ratings by mail.

Table 10 presents a summary of cost comparisons for the group-process condition and the two variations of the independent condition. Costs estimated in Table 10 are based upon figures published in the Corporate Travel Index for 1988, the most recent year for which complete information was available. (To adjust for inflation, figures listed in the Index were increased by a factor of 1.1236. The adjustment factor assumes a uniform 6% per year increase in costs due to inflation and was applied to all travel expense categories.)

Table 10

Comparison of Costs for Conducting a Passing Score Study under Group-Process and Independent Conditions

	Group-Process Condition	Independent Conditions	
		With Meeting	Without Meeting
Meeting Time	2 days/2 nights	1 day/0 nights	0 days/0 nights
Air Travel	$4000.00	$4000.00	n/a
Lodging	1309.99	n/a***	n/a
Meals *	1014.84	507.42	n/a
Transportation **	200.00	200.00	n/a
Informational Mailing +	12.50	12.50	12.50
Test Items Mailing ++	n/a	85.00	85.00
Test Items Return Mailing ++	n/a	85.00	85.00
TOTALS	$6537.33	$4889.92	$182.50

Notes:

* = includes tax and gratuity.

** = assumes $10.00 per person each way to and from meeting site.

*** = assumes travel to and from meeting site in one day.

\+ = first-class postage costs only.

++ = secure-method postage costs only.

The Index provides costs associated with travel expense categories ranked by major city. To determine costs for this study, the median city listed in the Index was used. For example, for meal costs, Harrisburg, Pennsylvania represented the median of 100 U.S. cities for meal costs. Thus, meal costs were estimated using the data supplied for Harrisburg; however, for other travel categories, the median major city within the category was used (i.e., not necessarily Harrisburg).

It should further be noted that expenses associated with travel and consultation by a psychometrician or testing organization representative have not been included in the following analyses. Because licensure and certification boards vary in the extent to which they utilize in-house psychometric services or contract with external consultants, it was decided to omit this variable cost from each condition presented. Also excluded because of wide variability are expenses for conference room rental and equipment rental for the group-process and with-meeting conditions. Like the area of psychometric services, organizations vary widely in the extent to which they utilize "home office" facilities or conduct meetings off site. It is recognized that the exclusion of these expenses probably results in a downward bias in the overall cost estimates for the group-process and with-meeting conditions.

Two additional assumptions should be noted. First, because air travel costs are extremely variable, depending on the city of origin, destination, class of service, and time of week, the costs for air travel were estimated to be $400.00 per person using figures obtained from a national travel service agency for round-trip weekend travel to and from "Anywhere, U.S.A." Also, the with-meeting variation assumes

that reviewers would not require overnight lodging in order to participate in the meeting. In situations where overnight lodging is required due to distance travelled, flight connections, etc., lodging costs would be incurred. Thus, the with-meeting cost listed in Table 10 represents a lower bound estimate for that variation.

Examination of Table 10 suggests, based upon total costs for each of the three conditions, that the "Without-meeting" condition is by far the least costly method of conducting a passing score study. Total estimated costs for the three conditions are: Group-process condition, $6537.33; With-meeting condition, $4889.92; and Without-meeting condition, $182.50.

While it is true that the without-meeting condition is the least costly way of conducting a standard setting procedure when only monetary expenditures are considered, there are certainly other factors that should be discussed. For example, earlier in this section it was observed that the group-process conditions and independent conditions resulted in statistically and practically meaningful differences in passing scores. This is certainly not a factor that should be ignored. Another factor beside monetary cost that should be considered is the cost in terms of time. For many professions, it is quite difficult to identify SMEs who would be willing to forego two days of personal time or time away from professional activities in order to participate in a passing score study. For this reason, the "Without-meeting" condition, which would not require set-aside meeting time, could be viewed as the most economical. However, it should be noted that no data were collected as a part of Experiment 1 to address the psychometric properties of

the without-meeting condition. For that reason, this alternative can only be evaluated in terms of its economic feasibility and no conclusions regarding the accuracy or variability of without-meeting procedure results can be offered.

In summary, it was observed that expected savings in terms of time and financial resources were observed for the "With-meeting" and "Without-meeting" variations of the independent condition when compared to the group-process condition. However, it should be further noted that any savings incurred under any method would result in trade-offs that should be considered when those responsible for standard setting actually select a procedure. Also, some investigation of actual results from a without-meeting standard setting study seems warranted before any statements regarding its propriety should be made.

Experiment 2

Between-condition Mean Differences

Of primary interest in Experiment 2 was whether exposure to additional information (i.e., the distribution of the reviewers' own initial item ratings) would result in differing overall passing standards. Table 11 provides descriptive statistics comparing the ratings produced under the two conditions: "no-information" and "with-information." The no-information condition is defined as the independent provision by reviewers of Angoff ratings for the 100 items. These ratings were collected as part of Experiment 1. The with-information condition is defined as the independent provision of

a second set ratings for the same 100 items by the same reviewers, who were subsequently provided with the distribution of ratings generated under the no-information condition. Figure 2 shows a plot of the reviewers' means across 100 items, observed under each of these two conditions.

Table 11

Descriptive Statistics for No-Information and With-Information Reviewers Across 100 Items

	First Rating (No Information)			Second Rating (With Information)		
Reviewer	Mean	Standard Deviation	Skew	Mean	Standard Deviation	Skew
1	57.35	17.56	-.141	62.50	18.46	-.382
2	52.95	19.93	.015	62.64	18.45	-.784
3	55.75	17.02	-.053	61.99	21.77	-.288
4	51.80	18.11	-.092	59.75	14.64	-.450
5	56.64	25.70	-.076	53.50	22.59	-.024

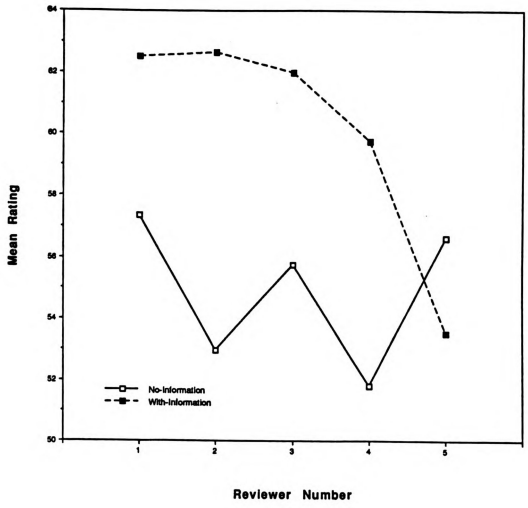


Figure 2

Plot of No-Information and With-Information Reviewers' Means

As shown in both Table 11 and Figure 2, overall ratings produced with knowledge of the other reviewers' ratings are somewhat greater and less variable. Again, as in Experiment 1, one reviewer's pattern of ratings diverged from the trend suggested by the rest of the group. Examination of Figure 2 reveals that Reviewer 5's overall ratings decreased under the with-information condition compared to the other reviewers whose overall ratings increased. Reviewer 5's overall with-information rating was also quite different from the fairly uniform overall with-information ratings provided by the other reviewers.

Two variables were created (NOINFO and WITHINFO) to reflect each item's overall rating under the two conditions. NOINFO represents the initial overall rating provided by the reviewers for each of the 100 items. WITHINFO represents the second (with information) rating provided by the reviewers for the same items. In each case, NOINFO and WITHINFO were obtained by calculating the mean rating for each item across reviewers within the no-information and with-information conditions. This procedure resulted in 100 pairs of ratings (one for each item).

The overall means for each condition and other descriptive statistics are presented in Table 12. The correlation between the overall ratings provided under each condition is also reported in Table 12. The magnitude of this correlation indicates fairly strong intra-reviewer agreement between initial and subsequent item ratings.

Table 12

Descriptive Statistics for No-Information and With-Information Condition Passing Scores

	NOINFO	WITHINFO
Mean	54.90	60.08
Standard Deviation	13.38	14.31
r NOINFO, WITHINFO =	.890	$p < .001$
Mean Difference =	5.178	

It should again be noted that the overall condition means represent the proposed passing score, expressed as a percentage, that would result from each condition. For example, the no-information condition would result in a passing percentage of approximately 54.9% compared to the 60.1% that would result if the passing standard were established using the with-information ratings. This means that the no-information standard of 54.9% would require examinees to respond correctly to approximately 110 items on a full test of 200 items in order to pass the examination. The passing standard suggested by reviewers in the with-information condition (60.1%) would translate into a passing score of approximately 120 items correct on a 200-item test. Regardless of the statistical significance of the difference between the two condition means, the 10 raw score unit difference in passing scores over a 200-item examination is clearly of practical significance.

To test whether the difference in overall condition means was statistically significant, a repeated measures analysis of variance (ANOVA) was conducted. The full model specified three factors: Items ($n = 100$); Raters ($n = 5$); and Conditions (replication) ($n = 2$). Results of the repeated measures ANOVA are presented in Table 13. Despite the substantial practical significance noted earlier, results of the repeated measures ANOVA failed to reveal a statistically significant difference between the two conditions. However, inspection of Table 13 shows an expected significant effect for items and a significant effect for raters. Clearly, the results indicate that both items and raters affect overall passing scores.

Table 13

Repeated Measures ANOVA Results for No-Information and With-Information Conditions

Source	Sum of Squares	Mean Square	df	F
Between Subjects				
Raters	9526.25	2381.56	4	11.71*
Within Subjects				
Conditions	881.75	881.75	1	0.77 ns
Raters x Conditions	4557.75	1139.44	4	
Items	179325.73	1811.37	99	8.91*
Items x Raters	80505.22	203.30	396	.72
Items x Conditions	10940.29	110.51	99	.39 ns
I x R x C, e	111660.91	281.97	396	
Total	397398.00	397.80	999	

* = $p < .001$

A test for homogeneity of variances with paired (dependent) observations was also performed. The calculated statistic ($t = 1.46$) did not exceed the critical value of 1.98 at alpha = .05 with 98 degrees of freedom. This result further suggests that overall item ratings were not more or less variable under either condition.

Relationship between With-Information and No-Information Ratings

As shown in Table 12, a significant Pearson product-moment correlation between no-information ratings and with-information ratings was observed (r NOINFO, WITHINFO =.890, $p < .001$). Calculation of the rank order correlation coefficient yielded similar results (r NOINFOrank, WITHINFOrank = .871. $p < .001$). These results indicate that the no-information and with-information conditions provided item ratings that were highly similar.

An intercorrelation matrix of item reviewers' first and second ratings was also produced and is presented in Table 14. Visual inspection of Table 14 reveals that reviewers' first and second ratings (i.e., under no-information and with-information conditions) are generally moderately correlated, ranging from a high of .759 (for Reviewer 5) to a low of .485 (for Reviewer 4) with a mean of .673.

Table 14

Intercorrelation Matrix of Ratings from No-Information and With-Information Condition Reviewers

	No-Information Reviewers (Initial Rating)					With-Information Reviewers (Second Rating)				
	R11	R21	R31	R41	R51	R12	R22	R32	R42	R52
R11	—	.389	.367	.249	.231	.650	.484	.306	.255	.273
R21		—	.305	.259	.342	.522	.749	.372	.474	.399
R31			—	.338	.294	.455	.449	.722	.251	.452
R41				—	.457	.276	.473	.296	.485	.559
R51					—	.323	.449	.436	.244	.759
R12						—	.554	.433	.321	.401
R22							—	.483	.516	.517
R32								—	.294	.519
R42									—	.320
R52										—

Two groups of correlations are enclosed by dashed lines in Table 14. The encircled values correspond to the correlations based only on no-information ratings (upper left) and those based only on with-information ratings (lower right). After transforming these correlations using Fisher's r to Z transformation, a mean correlation for each condition was computed and the two overall means were compared. As hypothesized, the average with-information correlation exceeded the average no-information correlation (.473 > .337); however, the difference between the two mean correlations was not statistically significant.

Decision Consistency

The extent to which exposure to the no-information and with-information conditions resulted in differing levels of classification consistency was also examined. Indices of classification consistency p and k were calculated for each condition using the passing scores suggested by each. The results are shown in Table 15. As Table 15 shows, application of the no-information passing score would result in a higher overall index of classification consistency ($\hat{p}_o$ = .934), compared to the with-information condition index ($\hat{p}_o$ = .898). Accordingly, the contribution to classification of the examination itself to consistency of pass/fail classifications was greater under the with-information condition ($\hat{k}$ = .706) compared to the no-information condition ($\hat{k}$ = .678).

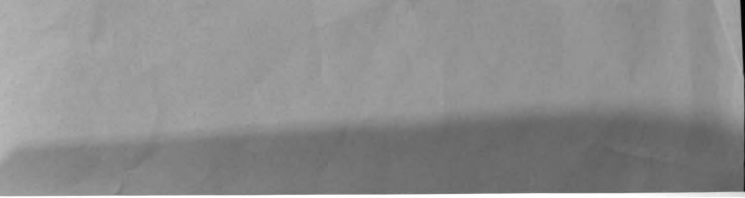


Table 15

Indices of Decision Consistency for No-Information and With-Information Conditions

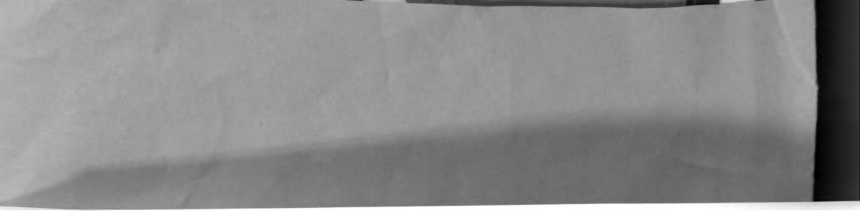
Condition	Suggested Passing Score	$\hat{P}o$	$\hat{k}$
No Information	54.9% (110)	.934	.678
With Information	60.1% (120)	.898	.706

Relationship of Ratings to Obtained Item Statistics

For item reviewers in the no-information (NOINFO) and with-information (WITHINFO) conditions, overall item ratings for the 100 items were compared to item difficulty indices resulting from the actual administration of the examination. As in Experiment 1, modified p-values (MODP) were used, obtained by calculating each item's difficulty based only on the responses of examinees whose total score was within two standard errors of the passing score.

Correlations were calculated between the overall NOINFO and WITHINFO ratings and MODP. Correlations were also calculated between individual item reviewers' ratings and MODP. For both conditions, individual reviewers' ratings were found to be moderately related to MODP. Interestingly, the lowest correlation with MODP ($r = .197$) was observed for a reviewer in the with-information condition, while the highest correlation with MODP ($r = .505$) was observed for a reviewer in the no-information condition. Also, surprisingly, the no-information condition produced a higher (though non-significantly) overall correlation with modified p-values ($r = .590$) than the with-information condition ($r = .573$).

The two indices created to reflect the degree of agreement between reviewers' ratings and certain criteria (E and E') were also calculated for each reviewer. Table 16 presents the obtained values of absolute error of specification (E) and relative error of specification (E') for the five reviewers under no-information and with-information conditions. Comparison of the values displayed in Table 16 indicates that, generally, absolute errors of specification are only slightly reduced through the provision of additional



information. The mean absolute error of specification for the with-information condition (24.12) was quite close to the mean for the no-information condition (24.93). However, relative errors of specification were also sightly reduced under the with-information condition (mean = 13.43) compared to the no-information condition (mean = 14.81).

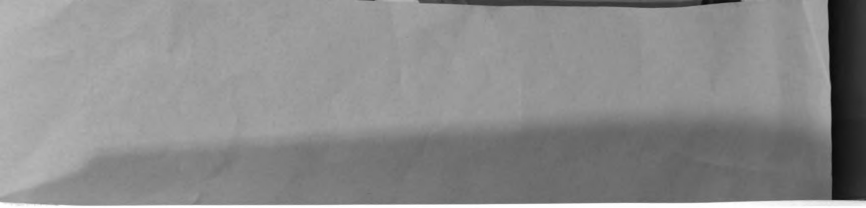


Table 16

Absolute and Relative Errors of Specification for Item Reviewers in No-Information and With-Information Conditions

	No-Information Condition		With-Information Condition	
Reviewer	E	E'	E	E'
1	23.52	14.09	22.48	12.98
2	26.27	14.76	22.89	10.99
3	23.95	13.24	24.95	14.36
4	25.94	13.77	25.84	12.78
5	24.99	18.17	24.46	16.04
Mean	24.93	14.81	24.12	13.43
Standard Deviation	1.20	1.96	1.41	1.89

In evaluating the effect of the provision of additional information, it is again observed that individual item reviewers were more proficient at estimating the overall group rating for the items than they were at predicting how the hypothetical minimally-competent examinee group would perform.

Regression Analyses

In order to further evaluate the effect of providing additional information to item reviewers, five regression analyses were performed. A regression model was developed which reflects the hypothesis that an individual reviewer's second (i.e., with-information) rating can be predicted by knowledge of his original (without-information) rating and with knowledge of the group's original mean rating (with the group mean calculated excluding the individual reviewer). These two ratings were used as the independent variables in the regression equations with the reviewer's revised (with-information) rating used as the dependent variable. Theoretically, the model assumes that reviewers' make their judgments about item ratings based upon their own procedure-related knowledge; that is, knowledge regarding the hypothetical minimally-competent examinee group and the difficulty of the items being rated. And, reviewers take into account information gleaned from other reviewers; in this case, from the distribution of reviewers' initial ratings that was provided for their use in the second round of ratings.

To assess the likelihood of such an effect, five regression analyses were conducted, one for each reviewer according to the procedure described above. Results of the analyses are presented in

Table 17. Raw (non-standardized) multiple regression equations are presented in the table, along with the correlations between the two independent variables, the Multiple R, and R squared. In each case, the correlations between the independent variables are low to moderate, suggesting that the choice of independent variables does not pose a threat of multicollinearity. For each regression performed, analyses of plots of predicted values against residuals revealed no disconcerting patterns; plots were broadly scattered and all residuals had means at or near zero.

Table 17

Regression Analyses for Individual Reviewers in Experiment 2

Reviewer	Regression Equation	r x1,x2	Mult. R	R^2
1	y = 8.805 + .535(x1) + .424(x2) + e	.425	.715	.511
2	y = 5.745 + .526(x1) + .524(x2) + e	.461	.827	.683
3	y =-1.073 + .789(x1) + .349(x2) + e	.456	.750	.563
4	y =29.528 + .290(x1) + .273(x2) + e	.480	.537	.288
5	y =-7.161 + .537(x1) + .555(x2) + e	.476	.807	.652

Notes: x1 = original rating for item i by reviewer j, and

x2 = group's original mean rating for item i computed with reviewer j excluded.

The hypothesized influence of additional information appeared to be evident in each of the regression analyses. For every reviewer, values of b_1 and b_2 were tested for significant difference from zero; in all cases, the t statistic were significant at $p < .01$. Further, the moderately high values of Multiple R and (with the exception of Reviewer 4) the moderate values of R squared suggest that the regression model has accounted for at least half of the variation in reviewers' ratings.

Combined Results

Selected results from Experiment 1 and Experiment 2 were combined to achieve an overall assessment of the effect of the various standard setting procedures. First, the group-process ratings from Experiment 1 were reanalyzed to obtain the passing standard that would result using rating for the first 100 items only. This was done so that direct comparisons could be made between the passing standards suggested by the group-process condition, the independent/no-information condition, and the independent/with-information condition, and the standards to be compared would be based upon ratings of the same 100 items.

Table 18 presents the results of the combined analysis. Several striking differences between the three procedures are apparent. First, the mean item ratings for the three procedures differ considerably, from a low of 48.88% (for the group-process condition) to a high of 60.08% (for the independent/with-information condition). The dramatic impact that differences of this magnitude would have on

actual classification decisions is also shown in Table 18. For example, the lowest passing rate (77.4%) was observed for the independent/with-information condition, while the highest passing rate (95.0%) was observed for the group-process condition. Accordingly, failure rates also varied dramatically, from 5.0% for the group-process condition to nearly 4 1/2 times as great for the independent/with-information condition (22.6%).

Table 18

Comparison of Experiment 1 and Experiment 2

Suggested Passing Standards

	Conditions		
	Group-Process	Independent No-Information	Independent With-Information
Mean Item Rating (across reviewers)	48.88	54.90	60.08
Standard Deviation of Reviewers' Overall Ratings	11.60	2.41	3.86
Standard Error	5.19	1.08	1.73
Passing Score (rounded)*	97.76(98)	109.80(110)	120.16(120)
95% Confidence Interval for Passing Score	88, 108	108,112	117, 124
Percent Passing (Failing)**	95.0(5.0)	86.8(13.2)	77.4(22.6)

* = adjusted to reflect passing standard for a 200-item test.

** = based on passing score obtained using Beuk (1984) method.

The issue of variability among individual reviewers' overall ratings (i.e, suggested passing standards) is also highlighted by the results displayed in Table 18. The wide variability across reviewers in the group-process condition is express statistically in the large standard deviation of group-process reviewers' ratings (11.60). This fairly large value for the standard deviation of group-process reviewers' ratings is also reflected in a correspondingly large standard error (5.19) and a very wide confidence interval (88, 108).

On the other hand, both of the independent conditions (i.e., the no-information and with-information conditions) displayed comparatively smaller standard deviations for reviewers' overall ratings and correspondingly smaller standard errors and confidence intervals. Surprisingly, the smallest standard error (1.08) and narrowest confidence interval (plus or minus 2 raw score units) was observed for the independent/no-information condition.

In summary, it should be emphasized that these fairly large differences may—or may not—be attributable to exposure to the experimental conditions. Because of the small panels of item reviewers utilized, it is possible that the results could be explained by random error. Although the social interaction hypothesis would predict the observed results, the failure to achieve statistical significance for group mean differences does not rule out the observation of these results due to chance.

V. DISCUSSION

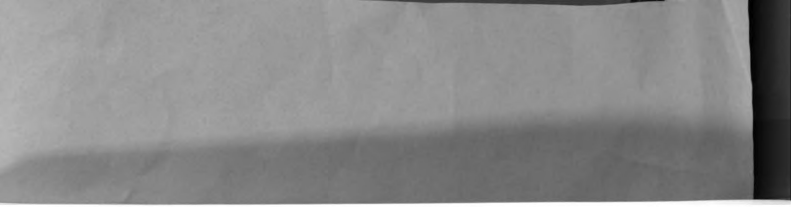
This study consisted of two experiments. The purpose of the first experiment was to examine one variation of the traditional group-process procedure of establishing passing standards using the Angoff standard setting methodology. The variation studied consisted of having item reviewers generate their Angoff ratings under an "independent" condition in which the usual effects of the group-process procedure (e.g., social comparison, sharing of information, etc.) could be controlled.

The purpose of the second experiment was to isolate the effect of one source of information that item reviewers use in generating their ratings--knowledge of the ratings provided by other (peer) reviewers. The results of each experiment are summarized below and a list of major findings and implications of these results is presented.

Experiment 1 Summary

Mean Ratings and Variability

Ten item reviewers in Experiment 1 provided Angoff ratings for 200 items on a medical specialty certification examination. Before providing their ratings, reviewers were given informational materials and participated in a training session to ensure their familiarity with the methodology. After this, reviewers were randomly assigned to



one of two conditions: an independent condition in which reviewers had no inter-reviewer interactions concerning their ratings, and a group-process condition in which reviewers freely discussed their ratings for items, item difficulty and relevance, and their conceptions of the hypothetical minimally-competent candidate group.

Exposure to the two conditions produced varied results. The primary question of interest was whether exposure to the conditions would yield differing passing standards. In Experiment 1, the passing standards obtained showed that the independent condition resulted in a standard that was approximately nine raw score points higher than the group-process condition. However, that difference was not statistically significant. Although the independent condition standard was higher, overall group item ratings provided by reviewers in each condition were nearly equally variable and fairly highly correlated.

A second variability issue addressed in Experiment 1 was whether the two conditions resulted in differential ratings for individual items. As hypothesized, independent reviewers exhibited, on average, a slightly wider spread of ratings for individual items than did reviewers in the group-process condition. This result complements the earlier observation of the higher standard suggested by the independent group in that the absence of reviewer interaction in the independent group may have contributed to this result. Conversely, the variability of the group-process condition ratings for individual items may have been reduced due to the effect of group interaction.

It is critical at this point, however, to highlight the failure to achieve statistical significance for observed differences between

mean ratings for the two conditions in Experiment 1. Although the results would surely result in large practical consequences for the examinee population, the profession, and the certifying board, confident statements regarding the reproducability of the result cannot be made. Specifically, the failure to reject the null hypothesis for group mean differences means that the results could be explained simply with reference to random error: Different groups of item reviewers could be empanelled and arrive at identical passing scores or even at different passing scores in the opposite direction as those observed in this study.

Decision Consistency

Both the independent and group-process conditions yielded high indices of decision consistency, as evidenced by the coefficients $\hat{p}_o$ and $\hat{k}$. However, neither the fact that both indices were high or the fact that the group-process condition yielded slightly higher coefficients is particularly noteworthy: These findings can be explained by simply noting that the examination itself was highly reliable and that both the independent and group-process passing scores were not very close to the overall mean score on the examination (with the group-process condition passing standard located slightly further from the overall mean score than the standard suggested by the independent group).

Relationship of Ratings to Obtained Item Statistics and Reviewer Characteristics

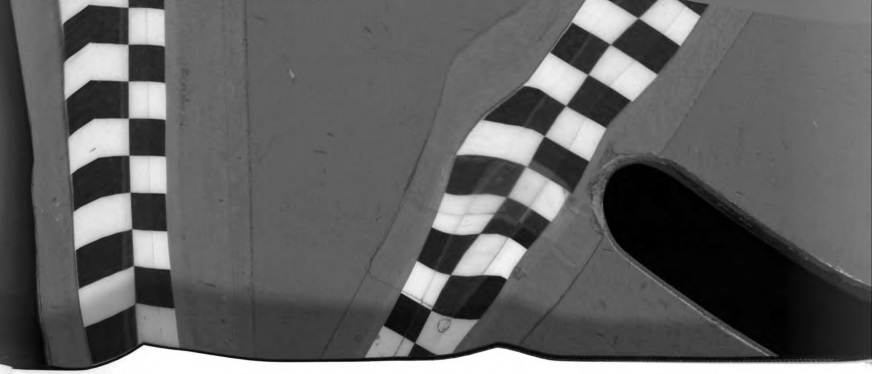
Ratings from item reviewers in the independent and group-process conditions were compared to p-values which were calculated using only

the responses of the hypothetical minimally-competent candidate group. Although, for all individual reviewers, correlations between item ratings and modified p-values were significantly different from zero, all of the correlations were uniformly low. Even when combined to form group average item ratings, correlations with modified p-values were moderate at best.

Similarly, the magnitude of the variables E and E' (conceptualized as average errors of specification for item ratings) indicated that individual item reviewers, in general, exhibited a fairly large degree of error when attempting to estimate the performance of the minimally-competent group, as evidenced by the large values of E. It is of small consolation that reviewers could more accurately provide estimates of their overall group item ratings, as evidenced by the relatively smaller values of E'.

These findings, taken together, all confirm one common criticism of the Angoff standard setting methodology--that item reviewers often experience some difficulty in accurately conceptualizing the minimally-competent examinee group.

Further, precision in estimation of item ratings does not appear to be dependent upon any of the reviewer characteristics measured in this study. For example, one might suspect that the more experience a reviewer had with producing and reviewing test items would lead to more accurate specification in item ratings. This result was not observed. Likewise, neither was a significant relationship observed between the extent to which reviewers reported to understand the Angoff methodology or their confidence in its results and the precision of their ratings. These results do not rule out the



possibility that other reviewer characteristics do contribute substantially to accuracy in item ratings; perhaps other significant background variables exist that were not measured in this study. On the other hand, it is also somewhat encouraging that the measured variables do not appear to influence reviewer accuracy. If standard setting bodies can be less concerned about these variables when empaneling reviewers, the pool of potential reviewers might be larger, possibly widening to include participation by able reviewers who may have otherwise been excluded.

Generalizability Analyses

Generalizability analyses were conducted to investigate differing sources of variation in item ratings so that potential future applications of either the independent or group-process procedures could be developed to yield increased dependability of measurement (i.e., dependability of item ratings). G-study results indicated that variance components were fairly well estimated (except for the group-process condition raters component) and would be useful for subsequent d-study analyses. D-study results for the independent and group-process conditions were obtained, varying the number of reviewers while holding the number of items constant. The results showed that slightly increased measurement dependability was achieved using under the independent condition as compared to the group-process condition, with acceptable results for operational purposes achieved with approximately 11 to 15 reviewers. This finding is contrasted with the suggestions of some that at least six to seven reviewers be empaneled for passing score decisions, although others (cf., Cross, et al.,

1984, p. 116) have also suggested that empaneling 15 or more reviewers is desirable.

D-study results also suggested that adding more item reviewers (or, possibly, more extensive reviewer training) would likely result in increased measurement dependability under either the independent or group-process conditions, though more so in the group-process condition. In practice, of course, increasing the number of test items would also, generally, improve overall dependability. However, with test length for the test under study already fairly long (n = 200 items), more and better-trained reviewers would likely be a more practical, less costly, and more efficacious method of addressing the issue of increasing the accuracy of item ratings.

Cost Analysis

Because the independent item review procedure was proposed as an efficient alternative to the group-process procedure, a cost analysis was also conducted. As expected, the financial costs associated with implementation of an independent/with-meeting rating procedure were lower than the costs associated with conducting the traditional group-process procedure for a 200-item examination. Substantially lower costs yet were estimated for an independent/without-meeting procedure. However, it is noted that some control over the standard setting process is surely lost when either independent condition is utilized. One potentially important element that is excluded from the independent/without-meeting condition is the ability of reviewers, as a group, to arrive at some consensus regarding their conception of the minimally-competent examinee group—an important aspect of the Angoff

methodology. And, it is unknown how standards established using an independent/without-meeting procedure would compare to the independent/with-meeting or group-process procedures examined in this research.

Promising results were observed for the one variation of the independent procedure in which item reviewers assemble only long enough to receive group training, become familiar with the methodology, and develop common referents regarding the minimally-competent group. This variation was also less expensive that the traditional group-process method, but would require a greater time commitment on the part of potential item reviewers. This option, however, should probably be considered by groups contemplating the need for a standard setting study in light of earlier findings regarding the importance of reviewer training.

Experiment 2 Summary

Mean Ratings and Variability

Five item reviewers--the same reviewers who participated as independent item reviewers in Experiment 1--were each provided with the five ratings generated for each of the first 100 items form the 200-item examination used in Experiment 1. The reviewers were asked to reread the 100 items, to review the distribution of initial ratings for each item, and to independently provide a second rating for each item. This procedure created two conditions: a "No-Information" condition represented by the initial ratings generated independently before any normative information was provided, and a "With-Information" condition represented by the subsequent ratings generated

with knowledge of the distributions of initial ratings for each item.

Fairly consistently, ratings generated under the with-information condition were higher than ratings generated by the same reviewers under the no-information condition. Differences between the condition means were of statistical and practical significance. However, overall mean item ratings across reviewers were roughly equally variable for the no-information and with-information conditions, although at the individual item level, a slight reduction in variability for the with-information ratings was observed.

These findings generally complement the findings presented for Experiment 1. For example, the provision of additional information--in the form of the distributions of item ratings--may have had the effect of communicating to reviewers a group "expectation" or conceptualization regarding minimal competence levels which they used in generating their second set of ratings. Accordingly, reviewers whose ratings may have been extreme initially were subtly induced to converge on the standard implied by the distributions of item ratings, making their subsequent ratings for individual items somewhat less variable. This effect is similar to what some have termed the "reality check" aspect of the modified Angoff method in which item reviewers, after providing an initial set of ratings, are given empirical item difficulty levels and asked to generate a second (revised) set of ratings.

Relationship of Ratings to Obtained Item Statistics

Ratings from item reviewers in the no-information and with-information conditions were compared to p-values which were calculated

again using only the responses of the hypothetical minimally-competent candidate group. Although, for all individual reviewers, correlations between item ratings and modified p-values were significantly different from zero, all of the correlations were of low to moderate magnitude. Also, average correlations between ratings provided under the two conditions and the modified p-values did not differ significantly. These results likely mean that the provision of additional information did not influence the reviewers to converge on the standard that would be suggested by the actual performance of the minimally-competent group (as operationalized in this study). Rather, reviewers converged on their own—somewhat inaccurate—conception of the level at which an appropriate minimum standard should be set. In fact, reviewers in both the no-information and with-information conditions had similar and fairly large absolute errors of specification. Mean relative errors of specification for the two conditions were also quite close.

Because the same reviewers who provided ratings for Experiment 1 also provided ratings for the second experiment, the results are somewhat dependent; the low degree of accuracy found in Experiment 1 is, to some extent, carried over to Experiment 2. The results presented here, however, strongly suggest that providing item reviewers with additional information in the form of distributions of initial ratings—although this has been promoted by other researchers in the area of standard setting as a means of decreasing the variability of ratings—does not contribute substantially, if at all, to the accuracy of those ratings.

Regression Analyses

Several regression analyses were performed to ascertain what impact certain factors had on individual reviewer's with-information ratings. A model was proposed that suggested individual reviewer's with-information ratings could be predicted with knowledge of the reviewer's initial ideas about an item (i.e., the reviewer's initial, no-information rating) and knowledge of the initial group opinion about the item (i.e., the initial group item rating gleaned form the distribution of initial ratings).

In all five cases, reviewers' subsequent ratings were substantially and significantly affected by their initial ratings and by the group opinion. In each case, the multiple regression equation for an individual reviewer explained approximately 50% of the variation in reviewers' subsequent ratings. Although this result is partially encouraging, it does leave considerable room for improvement in predicting power.

One possible factor that may have moderated this result is the time period that passed between initial (i.e., no-information) and subsequent (with-information) ratings. In this study, approximately four weeks passed between the time initial ratings were gathered and the time distributions of initial ratings were mailed to reviewers. It is possible that during the interim time period, reviewers lost some of their familiarity with concepts central to the standard setting methodology that this contributed one source of error to the second sets of ratings. It would be of interest to learn if varying the time period between no-information and with-information ratings is related to the degree to which the second set of ratings can be

accurately predicted.

Discussion of Combined Analysis

Experiment 1 compared Angoff item ratings generated by two groups of content expert item reviewers: Group-process reviewers who shared opinions about individual items and information concerning the construct of "minimal competence," and independent reviewers who provided their item ratings without such interaction (sharing of information). A check on the reasonableness of Experiment 1 results was undertaken in Experiment 2, in which reviewers were prohibited from personal interaction, but were provided with information concerning others' ratings of items. Such a check was indicated primarily because sample sizes used in Experiment 1 were small and some verification that the effect of providing information could exert a predictable effect on item ratings was desired. Also, because the group-process condition was susceptible to possible over-influence by one or more reviewers with strongly-held opinions, an attempt was made—through Experiment 2—to examine how suggested passing standards might change when the social influence of individual reviewers was controlled.

To accomplish this, the same independent reviewers from Experiment 1 rerated a subset of the same items they rated as part of Experiment 1. However, for their second sets of ratings, reviewers in Experiment 2 were provided with relevant information in the form of distributions of their original ratings generated during Experiment 1.

A combined analysis of the two experiments across the three

differing conditions (i.e., group-process, independent/no-information, and independent/with-information) presented some unexpected results. First, each of the conditions resulted in different overall passing standards. However, and possibly due to the relatively small sample sizes, differences between the suggested passing standards were not statistically significant. This observation can, in practice, result in fairly great practical consequences, though: Application of the differing passing standards would yield substantially differing passing and failing rates and, consequently, would result in different certification decisions for fairly large proportions of examinees.

The three conditions also displayed differences with respect to the variability of individual reviewers' passing standards. And, the observed differences in variability were not always in the hypothesized direction. Specifically, although the presence of information (whether in the form of group-process interaction, or through the provision of initial item ratings only) tended to result in less variability across reviewers for individual item ratings, it did not have a predictable effect on overall passing standards. For example, reviewers overall ratings (i.e., passing standards) were least variable under the independent/no-information condition and most variable in the group-process format--a finding that would not be expected if the influence of information gained in the group-process setting exerted its influence as would be expected. Also, surprisingly, in comparing the independent/with-information and independent/no-information conditions, variability in overall ratings was reduced--though slightly--in the no-information condition.

Finally, it was expected that the independent/with-information

condition would result in a suggested passing standard that would fall somewhere between the standards suggested by the group-process procedure and the independent/no-information procedure. Or, conceptually, it was expected that the effect of providing the distributions of initial item ratings would be to moderate the social effects of the group-process condition, while also tempering the wide variability anticipated for reviewers who rate items independently. These expected results were not observed. Instead, reviewers who rated items in the independent/with-information condition provided the highest overall standard of the three groups--a standard higher that either the group-process standard and higher than their own initial (no-information) standard.

In evaluating these results it is observed again that caution is indicated because of the size of the sample employed. For example, interpretation of some of these findings can, to some extent, be explained with reference to the extreme ratings provided by a set of one or more reviewers in each of the three rating situations. However, extreme or aberrant ratings are a characteristic of most standard setting applications. In practice, as many others have noted in psychometric analyses, variation in individual and overall ratings will certainly be observed and will contribute to the dependability of the standard setting process.

From another perspective, however, the results presented above help to make explicit some of the often implicit policy considerations in standard setting. Certainly choice of standard setting methodology is a policy decision, given what is already known about the likely effects of methodology on the magnitude of resulting passing

standards. Also, in this study, within the Angoff methodology, three variations for implementing that methodology were examined. A policy decision to utilize one of these or other Angoff variations will also surely have an impact of the resulting suggested passing standard. Finally, policy decisions might arise if consideration is given to the extent of interreviewer variation that is acceptable, either in terms of individual item ratings or for overall passing standards across reviewers. Results presented in this study have made this policy consideration especially salient, with the effects of individual reviewers highlighting the need for further attention to variance-reducing measures.

Summary of Findings and Implications

This section presents a distillation of the twelve key findings of the study, with implications for future research and standard setting practice.

1. Finding: The group-process procedure for establishing a passing standard did not result in a passing score that was significantly different from the passing score generated by reviewers in the independent condition.

Implication: Because of the failure to achieve statistical significance, observed differences between the independent and group-process procedures

may be attributable to chance. However, in the context of standard setting, even chance differences can present practical concerns. For example, it is often a practical concern for standard setting bodies that the standard setting methodology utilized might yield a standard that results in an unacceptably high failure rate. Use of the independent procedure studied here might heighten such concerns, although because of the failure to observe statistically significant differences, the same result may not be observed in other applications of the procedure. It is clearly desirable for further research to be done comparing the independent and group-process procedures so that assurance regarding any true differences between the procedures can be attained.

Additionally, the present research addressed only group-process and independent variations of the Angoff methodology; however, of the absolute standard setting methodologies in prevalent usage (Angoff, Ebel, Nedelsky), it is often reported that the Angoff method yields higher passing scores than the others. It is possible, therefore, that in some instances the independent variation of the Angoff method described in this study would yield passing standards that are not politically or practically feasible. Replication of this research with other, larger samples, and replication using other methodologies seems

warranted.

2. **Finding: Large practical consequences were observed for applications of the passing standards suggested by the group-process and independent procedures.**

Implications: Despite the fact that mean differences for the independent and group-process conditions were not statistically significant, the observed differences in passing standards for the two groups would result in substantial practical consequences. For example, it was noted that the percentage of examinees who would pass under the two standards varied from 86.8% for the independent condition to 95.0% for the group-process condition. Accordingly, the failure rates more than doubled for the independent condition (13.2%) compared to the group-process condition (5.0%). Standard errors also varied substantially for the two conditions (group-process = 5.19; independent = 1.08).

It is noteworthy that such large differences can occur in the absence of statistical significance. Obviously, this result is related to the small sample of reviewers employed for the rating of items. However, it is also worth mentioning that fairly small samples are often utilized for standard setting

purposes. Accordingly, it is suggested that the uncertainty regarding "true" classifications associated with passing standards established with judgmental methodologies should become more prominent in standard setting discussions. Specifically, the confidence that can be placed in the accuracy of the passing standard (i.e., the size of the standard errors) directly translates into confidence about decisions made for individual examinees (e.g., certify/do not certify) and into sometimes strong inferences about those examinees (e.g., safe to practice/unsafe practitioner).

It is recommended that the uncertainty associated with passing score estimation assume a more prominent place not only in deliberations by standard setting entities, but also in reporting on procedure results to those responsible for the standards, and possibly in reporting to those affected by the standards (e.g., professional groups, individual examinees, the public). It is interesting to note that "information on how to interpret the reported score, and any cut score used for classification" is listed as a primary standard in the Standards for Educational and Psychological Measurement (AERA/APA/NCME, 1985, p. 53). However, the notion that "where cut scores are specified for selection or classification, the standard error of

measurement should be reported for score levels at or near the cut score" is listed as a secondary standard by the Standards (p. 22). Secondary standards are described as those that are "likely beyond reasonalbe expectation in many situations" (p.33). No mention is made of reporting the accuracy with which the passing score is estimated in the section of the Standards dealing with professional and occupational licensure and certification testing (Section 11). It might be appropriate for that section of the Standards to be revised to include reporting, to those affected by the pass/fail decisions, of the confidence that can be placed in the accuracy of passing score estimation.

3. Finding: Interrater agreement (within items) was somewhat higher for the group-process procedure than for the independent procedure.

Implication: Because the group-process procedure tended to produce ratings for individual items that were somewhat less variable across reviewers than did the independent procedure, the group-process procedure may be preferrable to the independent procedure for its variance reducing effect. This suggestion would be true if interrater agreement on ratings for individual items remains a goal of

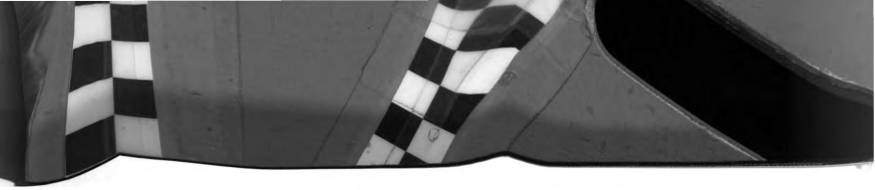
standard setting methodologies. However, for standard setting applications in which there is concern that the group-process procedure might be unduly biased by one or more dominant raters, the independent procedure may be preferable.

4. Finding: Variation between reviewers' means across all items was roughly equal for the independent and group-process procedures.

Implication: The use of either the independent or group-process procedure does not seem to constrict the spread of item ratings when viewed across items. Reviewers using either procedure appear to be able to make fairly consistent discriminations between relatively easier and more difficult items.

5. Finding: Neither item reviewers in the independent rating condition nor reviewers in the group-process condition exhibited desirable levels of accuracy in estimating the performance of the minimally competent examinee group.

Implication: Reviewers in both the independent and group-process conditions were fairly poor at predicting the p-values that were actually obtained from administration of the examination. Thus, it



appears that the effect of group interaction, while increasing precision somewhat, has no positive effect on reduction of rating bias. It is possible that efforts to reduce variation between reviewers may not be as profitable as efforts to help reviewers internalize an accurate conception of minimally competent performance. The provision of item p-values during the rating process has already been suggested by others as a means of increasing accuracy; however, the results of this study suggest that modified p-values, rather than conventional p-values, should be used when it is decided to provide normative information.

6. Finding: Relatively large variance components for raters were observed for both the independent and group-process conditions.

Implication: Large variance components associated with item reviewers suggests that large differences in observed passing scores could be noted if different groups of item reviewers are empaneled. As mentioned earlier (see Finding 2), the attendant uncertainty with which the passing score is estimated should probably become a matter of wider acknowledgement, discussion, and reporting.

At least two methods for reducing the magnitude

of variance associated with reiviewers' item ratings are known: improving reviewer training and increasing the number of reviewers. The importance of appropriate reviewer training has been stressed by several researchers in the area of standard setting. This study acknowledges the criticality of appropriate training in the passing score methodology to be used, prior to data collection. Certainly, it was also observed that increasing the number of item reviewers leads to increased measurement dependability. However, it is noted that, even when the number of item reviewers is increased--the solution suggested most frequently to reduce this source of variation--only modest gains in overall measurement dependability were observed. An obvious implication of this finding is that standard setting bodies would do well to increase the number of reviewers as well as expend the time necessary to ensure that all item reviewers clearly grasp the mechanics of the methodology employed and possess clear conceptions of elements central to the methodology (such as that of the minimally competent candidate, acceptable performance, etc.).

7. Finding: For both the independent and group-process conditions, acceptable dependability of measurement was obtained with 11 to 15 item

reviewers.

Implication: This finding is contrasted with the earlier work of Smith, Smith, et al., (1988) that suggests six to seven raters as an acceptable number of reviewers for conducting the passing score methodology, but agrees with the recommendations of others (cf. Cross, et al., 1984).

Certification and licensure entities considering implementing a passing score methodology should expect results to become very unstable as fewer item reviewers are utilized. Greater numbers of reviewers should, of course, be utilized whenever feasible (as suggested by Cross, et al, 1984) to improve the dependability of the standard setting procedure. For example, this research found that, using the independent or group-process procedures, a dependability index of approximately .70 could be attained with six to eight item reviewers, respectively. A dependability index of approximately .80 was obtained with 11 to 14 reviewers. Although the numbers of reviewers in these cases are similar to the numbers of reviewers commonly employed in standard setting procedures, the corresponding indices of dependability seem somewhat low for decisionmaking purposes. On the other hand, it was observed that a dependability index of approximately

.90 could be attained using at least 20 or more reviewers for each of the procedures. It is, however, uncommon for such a large group to be empaneled for standard setting procedures, especially in the areas of health and business professions credentialling. This fact points to the unavoidable need for standard setting entities to carefully weigh the importance of precise passing score estimation with their own practical, financial, and logistical considerations. In most cases, an increase in confidence about the passing score can be "bought" with better item reviewer training, an increased number of reviewers, or both of these. In any case, it is again noted that, regardless of the configuration of the procedure in terms of training, number of reviewers, etc., those responsible for standard setting would do well to recognize, evaluate, and to report the trade-offs that played a part in the determination of how the standard setting process was configured (e.g., which methodology was used, what training was implemented, how many reviewers were empaneled, etc.).

8. Finding: The independent condition was far less costly to implement than the group-process condition.

Implication: For standard setting bodies that

have severely constrained financial resources, the independent procedure presents an economically efficient alternative to a multiple-day meeting for purposes of establishing a passing standard. Additionally, if it is desired to increase the number of item reviewers, the independent procedure also presents an economical way to widen participation. However, the necessity of sound reviewer training is not obviated and should again be emphasized. The variation of the independent condition in which reviewers are convened only long enough to receive methodological training and to arrive at consensus on key conceptual issues before providing their ratings in isolation would seem to be a good alternative if the full (group-process) procedure is not possible.

9. Finding: Providing independent condition reviewers with information about the distribution of initial item ratings resulted in subsequent ratings that were generally higher and less variable.

Implication: The provision of additional information to item reviewers, in the form of distributions of their original item ratings, tends to cause reviewers to converge on an implicit standard of performance. It is not known, however, what degree of confidence can be expressed that the

provision of this kind of additional information will always result in subsequent ratings that are higher than the original ratings. Confirmatin of the directionality of the effect could be addressed by further research. It is likely though, that this result and the reduction in intra-item variability of ratings are fairly dependable outcomes of providing the additional information.

10. Finding: The no-information and with-information conditions produced overall item ratings that were of roughly equal variability.

Implication: The provision of additional information did not have the effect of reducing the spread of item ratings when viewed across items. Thus, it appears that reviewers do not reduce their notions of a rating "floor" or "ceiling" due to the provision of additional information and can still make fairly consistent discriminations between items they perceive as easier or more difficult for the minimally competent examinee group.

11. Finding: The provision of additional information to item reviewers in the form of distributions of original ratings does not appear to result in more precise estimates of the performance of the minimally competent examinee group.

Implication: Although the provision of additional information had the effect of enabling reviewers to better approximate the eventual group average rating for an item, it did not appreciably increase the reviewers' accuracy in estimating the performance of the minimally competent examinee group. As with the group-process procedure, additional information of the kind provided in this study does not appear to have the desired effect of helping reviewers to better estimate the "true" standard. Perhaps, in addition to the provision of modified p-values during the rating process, procedures for identifying item reviewers who are more familiar with the knowledge, skills, and abilities of the minimally competent examinee group should be investigated and utilized in future standard setting studies.

12. Finding: Knowledge of other item reviewers' ratings is a significant source of information that is used by individual reviewers to arrive at their own item ratings.

Implication: A reviewer's own initial opinion about the difficulty of an item for the minimally competent examinee group is often revised after exposure to information provided regarding peer

reviewers' opinions about the item. Although a reviewer's final rating can be fairly well predicted from knowledge of the reviewer's initial rating and knowledge of the initial group opinion, other factors contributing to the final rating certainly exist; further research is needed to identify which other key variables incline reviewers to alter their initial ratings when an iterative process is utilized. Because of the relatively small interrater variation in overall passing standards observed in this study using the independent procedure, an independent, iterative procedure appears to offer some promise. Future applications of independent methodologies using Delphi techniques might serve to take advantage of the positive characteristics of independent rating generation noted in this study: provision of relevant normative information and restriction of unwanted social comparison and other sources of irrelevant influences on item ratings.

Limitations and Suggestions for Future Research

The primary limitation of this study was the relatively small sample sizes employed and the consequent risk of a failure to be able to identify true differences between groups when such differences existed. Specifically, this means that some of the seemingly large

group differences in unexpected directions that were observed in this study could be attributed to chance.

However, despite this limitation in terms of statistical power, it was observed that chance differences would have substantial practical consequences for those affected by the passing score estimation procedure (e.g., examinees, the profession, the public, etc.). This situation has been cause for reexamination of the distinction between statistical and practical significance in the context of standard setting. In this study, even statistically nonsignificant findings resulted in often strikingly disparate practical consequences, such as differences in pass/fail rates and individual classification decisions. This situation has also been cause for reexamination of the suggested standards for documenting and reporting the attendant uncertainty in the results of judgmental passing score methodologies.

Another limitation of the study was that item reviewers were all male. It is not known if the same results would be obtained for female reviewers. Also, it would be advisable to investigate the applicability of the results described in this study with other, different medical specialty certification groups, as well as with other areas altogether (e.g., teacher licensure examinations, business credentialling programs, industrial selection applications, etc.).

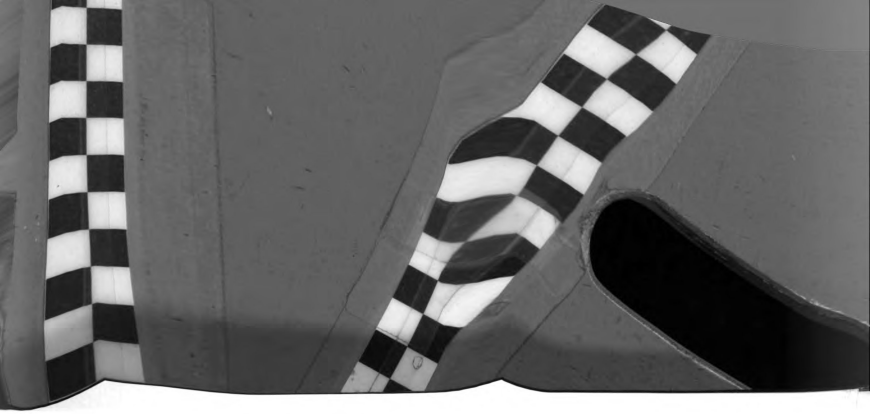
Similarly, the results described in this study were obtained using the Angoff standard setting methodology. One might wonder if the same results would have been observed if another absolute methodology (e.g., Ebel or Nedelsky) or if a common variation of the Angoff methodology (e.g., one of the "modified Angoff" approaches) had

been utilized.

As mentioned earlier, some investigation of the effect of the amount of time that passes between initial item ratings, the provision of additional information, and the generation of subsequent ratings seems warranted. It would be of interest to learn if the effect of providing the additional information was stable over time or if there were an optimal time period for providing the data to item reviewers.

Also, results of this study have served to make salient the often implicit policy considerations inherent in standard setting applications. Among these considerations are the choice of standard setting methodology, the manner in which the methodology is actually implemented, the kind of training provided, the number of item reviewers utilized, and the degree of variability that will be judged acceptable. This study has highlighted these concerns, and it is suggested that entities responsible for standard setting begin their investigation and planning for standard setting procedures sufficiently in advance of the time that results are needed so that policy considerations such as those listed above can be debated, made more explicit, and considered when the operational passing standard is set.

Finally, a recurring recommendation in this study has been that item reviewers require better training in the standard setting methodology in order to accurately predict performance of the minimally competent examinee group. If the ability for item reviewers to "zero in" on a standard delimiting some "true" line between acceptable and unacceptable performance is truly desired--and not just the ability of reviewers to provide similar ratings--then serious effort should be



expended toward identifying methods of assisting reviewers to succeed at that task.

It was suggested earlier that providing reviewers with modified p-values during the rating process and purposefully selecting reviewers who already possess a keen conception of the knowledge, skills, experiences, and abilities of the minimally competent examinee group would be beneficial. Possibly, the integration of already accepted and well-researched practice guidelines from other areas of education would also help to reduce error in item reviewers' estimates; it is apparent that simply increasing the number of reviewers is not enough. For example, applying principles of instructional design and enlisting the assistance of experts in the training field for designing and/or conducting the initial methodology orientation sessions might help address the issue of rating accuracy.

Undoubtedly, the necessity for setting fair, defensible, and accurate standards will remain. As long as needs for professional recognition, certification of competence, public protection, and personnel selection exist, criteria will need to be established that delineate acceptable from unacceptable performance. As long as strategies for setting standards that hinge on subjective conceptualizations of a hypothetical group are employed, variability of human judgments will coexist. This study reaffirms the notion that reduction in the variability of those judgments should be a goal of standard setting applications: Surely, a standard could not be strongly argued to be a valid standard if the measurements contributing to it (i.e., the reviewers' judgments) were not reliable.

However, this study has also demonstrated that individual and

collective reviewer judgments about the "truth" regarding that line delimiting minimal competence can themselves be fairly inaccurate. Perhaps, a redirection of effort, away from (or in addition to) attempts to make more consistent subjective judgments, and toward attempts to make such judgments less subjective will prove to be a rewarding research agenda.

APPENDIX A

Inter-Methodological Comparison Studies of Standard-Setting Procedures Involving One or More Absolute Standard-Setting Methodologies

Methodologies Compared	Study
Ebel v. Nedelsky	Andrew & Hecht, 1976 Schoon, Rosen & Jones, 1979 Schoon, Gullion, & Ferrara, 1979
Angoff v. Nedelsky	Behuniak, Archambault, & Gable, 1982 Brennan & Lockwood, 1980 Cross, Impara, Frary, & Jaeger, 1984 Harasym, 1981 Kleinke, 1980 Mills & Melican, 1987 Rock, Davis, & Werts, 1980 Smith & Smith, 1988 Subkoviak & Huff, 1986 van der Linden, 1982
Angoff, Ebel and Nedelsky	Colton & Hecht, 1981 Halpin, Sigmon, & Halpin, 1983 Poggio, Glasnapp, & Eros, 1981
Angoff, Modified Ebel and Normative	Skakun & Kling, 1980
Angoff v. Modified Angoff	Garrido & Payne, 1987
Angoff, Ebel and Normative	Greenberg & Smtih, 1988
Angoff v. Direct Standard	Jones, Rosen, & Schoon, 1988
Angoff, Beuk, and Hofstee	Bowers & Shindoll, 1989
Angoff, Ebel, Test Specifications, and Contrasting Groups	Mills & Barr, 1983
Nedelsky v. Contrasting Groups	Koffler, 1980

APPENDIX B

Establishing a Standard of Performance

for the

[ ]

Informational Materials

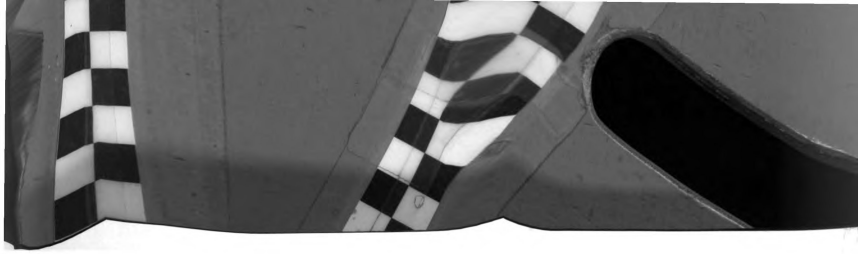
ESTABLISHING A SUGGESTED STANDARD OF PERFORMANCE FOR THE [ ] EXAMINATION

Overview

The [ ] Examination [ ] has been designed to assess the [ ] knowledge and skills of Board candidates. The Board has determined that a criterion reference standard of performance should be established on the [ ] so that candidate scores can be compared to a level of mastery that a group of experts in [ ] has judged to be sufficient for a given level of specialty practice. Setting a sufficient level of performance, or a standard of mastery, on the [ ] requires experts in [ ] to determine what constitutes a mastery level of performance for effective practice at the specified level.

There are a number of procedures available for establishing standards of performance, each based upon subjective judgments of a group of content experts selected to be representative of important perspectives in the profession. One of the most popular is the Angoff method. In this method, each expert examines each question in the test and estimates how many examinees whose level of knowledge is sufficient and acceptable for entry into practice will respond correctly. When the estimates for all items are summed and averaged across all experts, the result is the suggested standard of mastery for the test.

Before specific instructions for conducting this method are presented, a brief explanation of the notion of a "sufficient and acceptable level of knowledge" might be helpful.



Notion of an "Acceptable Level of Knowledge"

The purpose of most examinations is to measure an examinee's level of knowledge in the content area covered by the examination. The purpose of a suggested standard of performance for the [ ] would be to differentiate between those physicians who have a sufficient and acceptable level of knowledge for the safe practice of [ ], and those who do not. In setting the standard of performance, it is essential to keep in mind the concept of the examinee whose knowledge is right at a "sufficient and acceptable level" for safe practice.

It is important to have a clear conceptualization of this hypothetical examinee. Suppose a group of physicians who seek to practice [ ] is assembled. These physicians are lined up, from the most knowledgeable physician to the least knowledgeable physician. The challenge is to start at the end of the line with the physician who has the greatest knowledge and walk toward the other end of the line to the physician who has the least knowledge. At some point it is possible to stop and say, "All the physicians whom I have walked past have a sufficient and acceptable level of knowledge. All of the physicians whom I have not walked past do not have a sufficient level of knowledge."

Now consider two physicians in the line: the last one you walked past (Physician A) and the next one whom you did not walk past (Physician B). Physician A will be considered prepared to practice as a [ ]. This is the physician who has the least knowledge of all those considered to have a sufficient and acceptable level of knowledge. Physician A knows just enough to practice safely. Physician B has the greatest knowledge of all

those who will be judged NOT to have a sufficient and acceptable level of knowledge. This physician does not know quite enough to practice safely as an [ ].

Now think again about Physician A, who knows just enough to begin practicing safely as a [ ]--that is, he has a sufficient and acceptable level of knowledge for entry into the profession. How would this physician be described? How much does this physician know? What kinds of problems should be entrusted to this physician? What are this physician's skills? Thinking about the knowledge and skills a physician must have to perform effectively is important as you participate in the Angoff standard-setting method. The following descriptors and questions may help to further conceptualize the borderline examinee:

An [ ] who has a sufficient and acceptable level of knowledge for entry into practice will:

1) demonstrate a knowledge base sufficient to diagnose and manage disease.

2) know the boundaries of the specialty and the profession. What types of problems should such a physician refer to other professionals?

3) make some errors. What types of errors cannot be tolerated?

4) be aware of the standards, laws, and ethical issues related to specialty practice. What do these include?

Once the notion of a "sufficient and acceptable level of knowledge" becomes clearer, the Angoff method can be conducted.

Instructions for Conducting the Angoff Method

Suppose that a hypothetical group of 100 physicians who have a sufficient and acceptable level of knowledge--physicians just like Physician A--are gathered in a room. These 100 physicians have been asked to respond to each question in the [ ].

To conduct the Angoff method, you will need to complete two basic steps for each item in the WQE:

1) Read the item thoroughly. Think about how frequently the knowledge or skill tested in the item is used in the practice of [ ]. Also think about how critical that knowledge or skill is to the practice of [ ]. For example, if a piece of knowledge or a skill is always critical, there will always be a serious adverse effect (for example, [ ]) if it is not known or is used incompetently. You might expect that a large percentage of examinees would respond correctly to critical-knowledge test items.

2) Next, estimate the percentage of sufficiently prepared examinees--examinees like Physician A--who will answer the question correctly. Write this percentage in the blank labeled Item 1 on your rating sheet. Remember that some of these examinees will answer correctly by guessing.

Please provide your estimates in multiples of 5. If you are not familiar with the content of a particular item and feel uncomfortable about rating it, you may leave the item blank. Please try, however, to rate as many items as you can.

Example:

1. Melanin is synthesized from which of the following amino acids?

   A. Lysine
   B. Leucine
   *C. Tyrosine
   D. Histidine
   E. Phenylalanine

2. A patient who has insulin-dependent diabetes experiences early-morning hyperglycemia that is not preceded by hypoglycemia. The insulin dosage need not be changed, because the hyperglycemia is due to:

   A. insulin resistance.
   B. waning of the insulin's effect.
   C. excessive levels of glycosylated hemoglobin.
   D. adrenocorticoid fluctuation.
   *E. a surge of growth hormone.

Suppose a rater reads question 1 and determines that it is testing relevant knowledge that is critical in certain situations, but only occasionally needed in the practice of [ ]. The rater estimates

that 40 out of 100 examinees who are prepared for practice will answer this question correctly. The rater then reads question 2 and determines that the information it is testing is fundamental and nearly all adequately prepared examinees will answer it correctly. The rater's rating sheet would be completed as follows:

Item No.	Estimated Percentage of Examinees Who Will Answer the Item Correctly
1.	40
2.	90

Once each rater has completed all the estimates for each item, the estimates will be averaged across raters and then across items. The result will be the suggested standard for this particular form of the [ ]. To illustrate, the following hypothetical example involves five content experts rating a 10-item test. Each rater's estimates are provided below, by item number.

Item No.	Rater 1	2	3	4	5	Item Average
1	70	80	65	70	50	67
2	70	90	60	60	75	71
3	85	90	70	70	60	75
4	80	70	45	70	40	61
5	75	80	50	50	70	65
6	75	30	40	40	50	47
7	70	90	45	60	65	66
8	65	80	60	50	75	66
9	70	80	60	60	50	64
10	75	40	80	60	50	61
Rater Average	73.5	73.0	57.5	59.0	58.5	643/10 = 64.3% or 6.43 items

The standard of mastery for this test would be set at 6 items out of 10. Examinees answering 6 or more items correctly would meet the performance standard for the test.

APPENDIX C

Sample Item Rating Collection Form

Directions:
Consider a hypothetical group of 100 physicians who have a sufficient and acceptable level of knowledge for safe practice of [ ]. What percentage of these physicians will answer each question correctly? Please enter your estimates clearly beside each item number. Please keep your estimates in multiples of 5 (e.g., 45, 60, 65,...).

Item No.	Estimated Percentage of Examinees Who Will Answer Item Correctly
1	___
2	___
3	___
4	___
5	___
6	___
7	___
8	___
9	___
10	___

APPENDIX D

Sample Post-Meeting Passing Score Study Questionnaire

Directions: For questions 1-3, please write your response in the underlined space provided.

1) Please indicate the number of years you have served as a [ ] Board Director. _____ years

2) Please indicate the number of years you have served on the [ ] Written Examination Committee. _____ years

3) Please indicate your primary practice setting. (e.g. private practice, teaching hospital, etc.) _____

For questions 4-7, please circle the number of the response that best characterizes your level of agreement with the statements below.

4) The informational materials were easy to understand and helpful.

1	2	3	4	5
Strongly Disagree	Disagree	Unsure	Agree	Strongly Agree

5) The item rating practice session was clear and helpful.

1	2	3	4	5
Strongly Disagree	Disagree	Unsure	Agree	Strongly Agree

6) The standard setting method used is easy to implement.

1	2	3	4	5
Strongly Disagree	Disagree	Unsure	Agree	Strongly Agree

7) The standard setting method used will result in a standard that adequately distinquishes between acceptably and unacceptably prepared examinees.

1	2	3	4	5
Strongly Disagree	Disagree	Unsure	Agree	Strongly Agree

Please write any additional comments or suggestions on the lines provided.

_____

_____

_____

_____

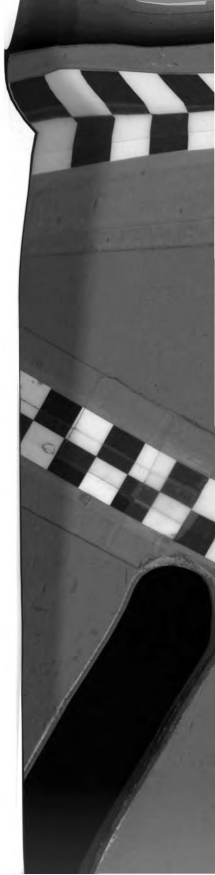
_____

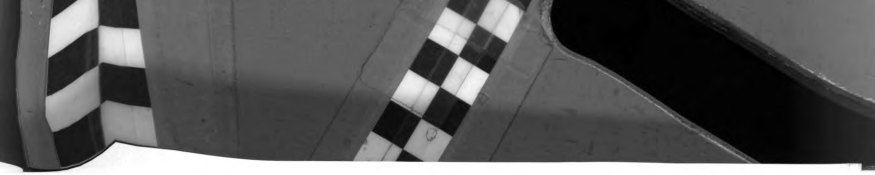
Thank you. Please place this questionnaire in the enclosed envelope and return it to [ ].

APPENDIX E

ITEMS	CONDITION 1 RATERS							CONDITION 2 RATERS							
	R1	R2	R3	R4	R5	Item Means	Item Variances	R6	R7	R8	R9	R10	Item Means	Item Variances	Item Difficulty
1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	$\bar{x}_{1,1}$	$s^2_{1,1}$	x_{16}	x_{17}	x_{18}	x_{19}	x_{110}	$\bar{x}_{1,2}$	$s^2_{1,2}$	p_1
2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	$\bar{x}_{2,1}$	$s^2_{2,1}$	x_{26}	x_{27}	x_{28}	x_{29}	x_{210}	$\bar{x}_{2,2}$	$s^2_{2,2}$	p_2
3	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	$\bar{x}_{3,1}$	$s^2_{3,1}$	x_{36}	x_{37}	x_{38}	x_{39}	x_{310}	$\bar{x}_{3,2}$	$s^2_{3,2}$	p_3
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
200	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
$\bar{x}_j$ =	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$	$\bar{x}_5$	$\bar{x}_{\cdot 1}$	$S^2_{\cdot 1}$	$\bar{x}_6$	$\bar{x}_7$	$\bar{x}_8$	$\bar{x}_9$	$\bar{x}_{10}$	$\bar{x}_{\cdot 2}$	$S^2_{\cdot 2}$	$\bar{p}$

Data Layout for Experiment 1

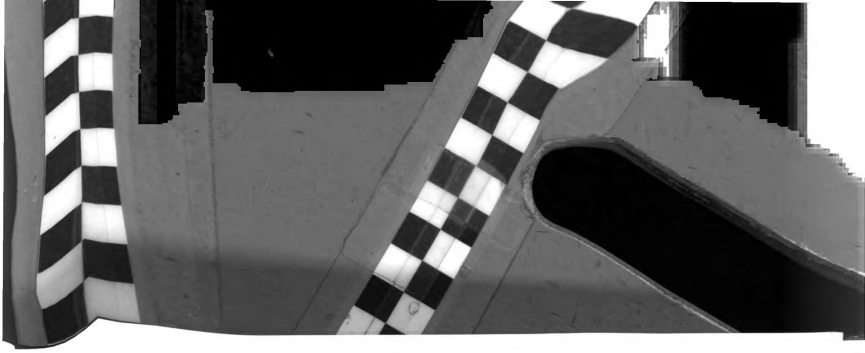




APPENDIX F

Sample Rating Form for Experiment 2

Item No.	Original Ratings by Independent Group					Second Rating
001	30	60	60	70	90	____
002	60	30	60	40	90	____
003	75	40	25	80	20	____
004	40	65	55	85	75	____
005	65	65	65	60	80	____
006	40	20	45	30	20	____
007	50	50	45	70	50	____
008	70	40	70	30	25	____
009	30	25	25	30	20	____
010	30	20	60	50	15	____
011	75	50	65	30	40	____
012	20	30	35	60	30	____
013	30	20	40	30	20	____
014	50	60	40	70	50	____
015	65	50	30	60	80	____
016	40	70	45	70	70	____
017	80	80	90	70	100	____
018	75	40	65	50	20	____
019	20	80	55	60	75	____
020	55	65	60	50	40	____
021	85	80	70	70	75	____
022	60	80	30	40	15	____
023	35	40	45	30	60	____
024	65	60	25	70	80	____



REFERENCES

American Board of Medical Specialties (1987). Recertification for medical specialties. Evanston, IL: Author.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 45-50.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (pp. 508-600). Washington, DC: American Council on Education.

Angoff, W. H. (1988). Proposals for theoretical and applied development in measurement. Applied Measurement in Education, 1(3), 215-222.

Babbie, E. R. (1973). Survey research methods. Belmont, CA: Wadsworth.

Behuniak, P., Jr., Archambault, F. X., & Gable, R. K. (1982). Angoff and Nedelsky standard setting procedures: Implications for the validity of proficiency test score interpretation. Educational and Psychological Measurement, 42, 247-255.

Berk, R. A. (1980). A framework for methodological advances in criterion-referenced testing. Applied Psychological Measurement, 4, 563-573.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.

Beuk, C. H. (1984). A method for reaching a compromise between absolute andrelative standards in examinations. Journal of Educational Measurement, 21, 147-152.

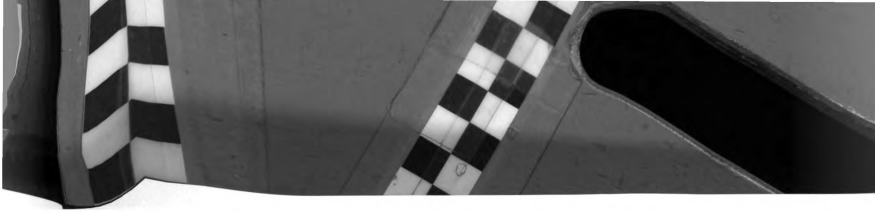
Bowers, J. J. & Shindoll, R. R. (1989, March). Angoff, Beuk, and Hofstee: A comparison of multiple methods for setting a passing score. Paper presented at the annual meeting of the National Council for Measurement in Education, San Francisco, CA.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City, IA: American College Testing Program.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4, 219-240.

Campbell, D. T. & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Boston, MA: Houghton Mifflin.

Cangelosi, S. S. (1984). Another answer to the cut-off score question. Educational Measurement: Issues and Practice, 3(4), 23-25.



Cizek, G. J. (1989). A report on the April 20, 1989 administration of the [...] examination. Iowa City, IA: American College Testing Program.

Collins, R. (1979). The credential society. New York: Academic Press.

Colton, D. A. & Hecht, J. T. (1981, April). A preliminary report on a study of three techniques for setting minimum passing scores. Symposium presentation at the annual meeting of the National Council on Measurement in Education, Los Angeles, CA.

Conaway, L. E. (1979). Setting standards in competency-based education: Some current practices and concerns. In M. A. Bunda & J. R. Sanders (Eds.), Practices and problems in competency-based education (pp. 72-88). Washington, DC: National Council for Measurement in Education.

Cramer, S. E. (1990, April). Some practical solutions to standard-setting problems: The Georgia teacher competency testing experience. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Cronbach, L. J. (1989). Educational measurement, third edition [Review of Educational measurement, 3rd Ed.]. Educational Measurement: Issues and Practice, 8(4), 22-25.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurement: Theory of generalizability for scores and profiles. New York: Wiley.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Journal of Educational Measurement, 21, 113-129.

Curry, L. (1987, April). Group decision process in setting cut off scores. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

deGruijter, D. N. M. (1980, June). Accounting for uncertainty in performance standards. Paper presented at the International Symposium on Educational Testing, Antwerp (ERIC Document No. ED 199 280).

deGruijter, D. N. M. & Hambleton, R. K. (1984). On problems encountered using decision theory to set cutoff scores. Applied Psychological Measurement, 8, 1-8.

DeMauro, G. E. & Powers, D. E. (1990, April). Internal consistency of the Angoff method of standard setting. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Dillon, G. F. (1990, April). The relationship of item position and Angoff-based standard setting judgments. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Fabrey, L. J. (1988, April). Adjustment of Angoff passing points. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Fabrey, L. J., & Raymond, M. R. (1987, April). Congruence of standard setting methods for a nursing certification examination. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), Educational measurement (pp. 105-146). New York: Macmillan.

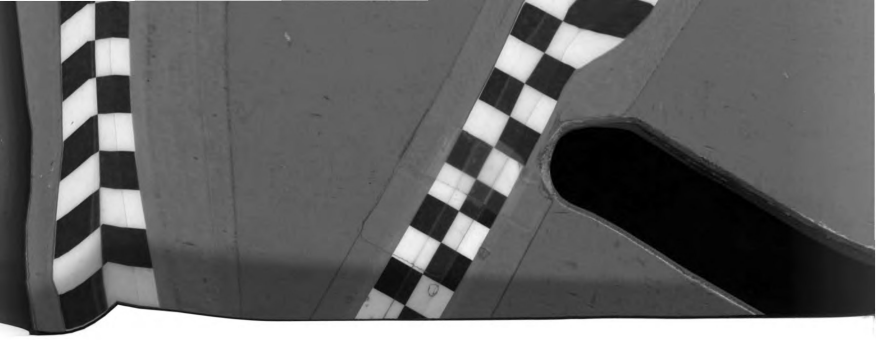
Fitzpatrick, A. R. (1984, April). Social influences in standard setting: Theeffect of group interaction on individual's judgments. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, IA.

Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. Review of Educational Research, 59, 315-328.

Francis, A. S. & Holmes, S. E. (1980, August). Criterion-referenced standard setting in certification and licensure: Defining the minimally competent candidate. Paper presented at the annual meeting of the American Psychological Association, Anaheim, CA.

Friedman, C. B., & Ho, K. T. (1990, April). Interjudge consensus and intrajudge consistency: Is it possible to have both in standard setting? Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Garrido, M. & Payne, D. A. (1987, April). An experimental study of the effect of judges' knowledge of item data on two forms of the Angoff standard setting method. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.



Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. American Psychologist, 18, 519-521.

Glass, G. V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-261.

Glass, G. V. & Hopkins, K. D. (1984). Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall.

Greenberg, S., & Smith, I. L. (1988, April). Strategies for reviewing existing standards in a high stakes profession. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Gross, L. J. (1985). Setting cutoff scores on credentialing examinations: A refinement in the Nedelsky procedure. Evaluation and the Health Professions, 8, 469-493.

Grosse, M. E., & Wright, B. D. (1986). Setting, evaluating, and maintaining certification standards with the Rasch model. Evaluation and the Health Professions, 9(3), 267-285.

Halpin, G., Sigmon, G., & Halpin, G. (1983). Minimum competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity. Educational and Psychological Measurement, 43, 185-196.

Hambleton, R. K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 15, 277-290.

Hambleton, R. K. & Eignor, D. R. (1980). Competency test development, validation and standard setting. In R. M. Jaeger & C. K. Tittle (Eds.), Minimum competency testing: Motives, models, measures, and consequences (pp. 367-396). Berkeley, CA: McCutchan.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 48, 1-47.

Harasym, P. H. (1980). A comparison of the Nedelsky and modified Angoff standard-setting procedure on evaluation outcome. Educational and Psychological Measurement, 41, 725-734.

Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson & J. S. Helmick (Eds.), On educational testing (pp. 109-127). San Francisco, CA: Jossey-Bass.

Hubbard, J. P. (1978). Measuring medical education: The test and the experience of the National Board of Medical Examiners. Philadelphia, PA: Lea and Febiger.

Huynh, H, (1976). On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 13, 253-264.

Jaeger, R. M. (1979). Measurement consequences of selected standard-setting models. In M. A. Bunda & J. R. Sanders (Eds.), Practices and problems in competency-based education (pp. 48-58). Washington, DC: National Council for Measurement in Education.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4, 461-475.

Jaeger, R. M. (1988). Use and effect of caution indices in detecting aberrant patterns of standard-setting judgments. Applied Measurement in Education, 1, 17-31.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), Educational measurement (pp. 485-514). New York: Macmillan.

Jones, J. P., Rosen, G. A., & Schoon, C. G. (1988, April). A preliminary investigation of the direct standard setting method. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kane, M. T. (1984, April). Strategies in validating licensure examinations. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kane, M. T. (1985). Strategies for validating licensure examinations. In J. C. Fortune (Ed.), Understanding testing in occupational licensure (pp. 45-63). San Francisco: Jossey-Bass.

Kane, M. T. (1986). The interpretability of passing scores (ACT Technical Bulletin No. 52). Iowa City, IA: The American College Testing Program.

Klein, L. W. (1984, April). Practical considerations in the design of standard setting studies in health occupations. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kleinke, D. J. (1980, April). Applying the Angoff and Nedelsky techniques to the National Licensing Examinations in landscape architecture. Paper presented at the annual meeting of the National Council for Measurement in Education, Boston, MA.

Koffler, S. L. (1980). A comparison of three approaches for setting proficiency standards. Journal of Educational Measurement, 17, 167-178.

Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2, 151-160.

Lerner, B. (1979). Tests and standards today: Attacks, counterattacks, and responses. In R. T. Lennon (Ed.), New directions for testing and measurement: Impactive changes on measurement (pp. 15-31). San Francisco, CA: Jossey-Bass.

Levin, H. M. (1978). Educational performance standards: Image or substance? Journal of Educational Measurement, 15, 309-319.

Likert, R. A. (1932). A technique for the measurement of attitudes. Archives of Psychology, 22(140).

Linn, R. L. (1978). Demands, cautions, and suggestions for setting standards. Journal of Educational Measurement, 15, 301-308.

Livingston, S. A., & Zieky, M. J. (1982). Passing scores. Princeton, NJ: Educational Testing Service.

Lockwood, R. E., Halpin, G., & McLean, J. E. (1986, April). Theoretical assumptions and situational constraints in the standard-setting process. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Maslow, A. P. (1983). Standards in occupational settings. In S. B. Anderson & J. S. Helmick (Eds.), On educational testing (pp. 91-108). San Francisco, CA: Jossey-Bass.

Mehrens, W. A. (1981, February). Setting standards for minimum competency tests. Presentation to the Michigan School Testing Conference, Ann Arbor, MI.

Melican, G. J. & Mills, C. N. (1987, April). The effect of knowledge of other judges ratings of item difficulty in an iterative process using the Nedelsky and Angoff methods. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 46, 133-158.

Meskauskas, J. A. (1986). Setting standards for credentialing examinations: An update. Evaluation and the Health Professions, 9, 187-203.

Meskauskas, J. A. & Norcini, J. J. (1980). Standard-setting in written and interactive (oral) specialty certification examinations: Issues, models, methods, challenges. Evaluation and the Health Professions, 3, 321-360.

Millman, J. (1973). Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 43, 205-216.

Millman, J. (1976). Reliability and validity of criterion-referenced test scores. In R. Traub (Ed), New directions for testing and measurement: Methodological developments (pp. 75-92), San Francisco, CA: Jossey-Bass.

Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. Educational Researcher, 18, 5-9.

Mills, C. N. (1983). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. Journal of Educational Measurement, 20, 283-292.

Mills, C. N. & Barr, J. E. (1983, April). A comparison of standard setting methods: Do the same judges establish the same standards with different methods? Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Mills, C. N. & Melican, G. J. (1986, March). A preliminary investigation of three compromise methods for establishing cut-off scores. Paper presented at the annual meeting of the National Council for Measurement in Education, San Francisco, CA.

Mills, C. N. & Melican, G. J. (1987, April). The effect of knowledge of other judges' ratings of item difficulty in an iterative process using the Angoff and Nedelsky procedures. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Mills, C. N. & Melican, G. J. (1988). Estimating and adjusting cutoff scores: Features of selected methods. Applied Measurement in Education, 1, 261-275.

Mills, C. N., & Melican, G. J. (1990, April). Equivalence of cut-off scores derived from randomly equivalent panels. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Nafziger, D. G. & Hiscox, N. D. (1976, April). A survey of occupational licensure and certification procedures. Paper presented at the annual meeting of the National Council for Measurement in Education, San Francisco, CA.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Norcini, J. J., Lipner, R. S., Langdon, L. O., & Strecker, C. A. (1987). A comparison of three variations on a standard-setting method. Journal of Educational Measurement, 24, 56-64.

Norcini, J. J., Shea, J. A., & Kanya, D. T. (1988). The effect of various factors on standard setting. Journal of Educational Measurement, 25, 57-65.

Plake, B. S., & Melican, G. J. (1986, April). Effects of item context on intrajudge consistency of expert judgments via the Nedelsky standard setting method. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Poggio, J. P., Glasnapp, D. R., & Eros, D. S. (1981, April). An empirical investigation of the Angoff, Ebel, and Nedelsky standard setting methods. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Popham, W. J. & Husek, T. R. (1969). Implications of criterion-referenced measurement. Journal of Educational Measurement, 6, 1-9.

Rock, D. A., Davis, E. L., & Werts, C. (1980, June). An empirical comparison of judgmental approaches to standard-setting procedures (ETS Research Report). Princeton, NJ: Educational Testing Service.

Rothman, R. (1989, November 8). States turn to student performance as new measure of school quality. Education Week, pp. 1, 12-13.

Saunders, J. C., Ryan, J. P., & Huynh, H. (1981). A comparison of two ways of setting passing scores based on the Nedelsky procedure. Applied Psychological Measurement, 5, 209-217.

Scheaffer, R. L., Mendenhall, W., and Ott, L. (1979). Elementary survey sampling. Boston, MA: Duxbury.

Schoon C. G., Guillion, C. M., & Ferrara, P. (1979). Bayesian statistics, credentialing examinations, and the determination of passing points. Evaluation in the Health Professions, 2, 181-201.

Schoon, C. G., Rosen, G., & Jones, J. P. (1988, April). A critique of difficulty estimation methodologies in the setting of cut points and a discussion of an alternative methodology: The direct standard setting method. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Scriven, M. (1978). How to anchor standards. Journal of Educational Measurement, 15, 273-275.

Shepard, L. (1979). Setting standards. In M. A. Bunda & J. R. Sanders (Eds.), Practices and problems in competency-based education (pp. 59-71). Washington, DC: National Council for Measurement in Education.

Shepard, L. (1980a). Standard setting issues and methods. Applied Psychological Measurement, 4, 447-467.

Shepard, L. (1980b). Technical issues in minimum competency testing. In D. C. Berliner (Ed.), Review of research in education, (pp. 30-84). Washington, DC: American Educational Research Association.

Shepard, L. (1983). Standards for placement and certification. In S. B. Anderson and J. S. Helmick (Eds.), On educational testing (pp.61-90). San Francisco, CA: Jossey-Bass.

Shepard, L. (1984). Setting performance standards. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 169-198). Baltimore, MD: Johns Hopkins University Press.

Shimberg, B. (1981). Testing for licensure and certification. American Psychologist, 36(10), 1138-1146.

Skakun, E. N. (1990, April). The effect of misinformation on judges' decisions in setting standards. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Skakun, E. N. & Kling, S. (1980). Comparability of methods for setting standards. Journal of Educational Measurement, 17, 229-235.

Smith, R. L. & Smith, J. K. (1988). Differential use of item information by judges using the Angoff and Nedelsky procedures. Journal of Educational Measurement, 25, 259-274.

Smith, J. K., Smith, R. L., Richards, C., & Barnhardt, S. (1988, March). The optimal number of judges to use in setting passing scores. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a mastery test. Journal of Educational Measurement, 13, 265-276.

Subkoviak, M. J. (1984). Estimating the reliability of mastery-nonmastery classification. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 267-291). Baltimore, MD: Johns Hopkins University Press.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. Journal of Educational Measurement, 25, 47-55.

Subkoviak, M. J. & Huff, K. J. (1986, April). Intrajudge inconsistency in the Angoff and Nedelsky methods of standard setting. Paper presented at the annual meeting of the National Council for Measurement in Education, San Francisco, CA.

Van der Linden, W. J. (1982). A latent trait method for determining the intrajudge inconsistency in the Angoff and Nedelsky techniques of setting standards. Journal of Educational Measurement, 19, 295-308.