STANDARD-SETTING IN WRITTEN AND INTERACTIVE (ORAL) SPECIALTY CERTIFICATION EXAMINATIONS
Issues, Models, Methods, Challenges

JOHN A. MESKAUSKAS
JOHN J. NORCINI
American Board of Internal Medicine

This article addresses standard-setting for written and interactive (oral) examinations in the health professions. The currently used methods are explained and classified, and their strengths and weaknesses are discussed. We argue that standard-setting must be understood as an interactive system involving decision maker(s), subject area, and methods. Thus, standard-setting is a methods. Thus, standard-setting is a psychological/social psychological process as well as a psychometric one. It rests upon a foundation of judgment. For written examinations, normative and content-referenced (absolute) methods are discussed. In interactive examinations, judges' standards are inherently absolute; design considerations are presented to systematize the context for these judgments.

Standard-setting processes are so ingrained in the daily functioning of our society that they seem mostly to engender uncritical acceptance. This state of affairs is surely true of much of education. The present controversy over the back-to-basics movement is clearly an issue of educa-

EVALUATION & THE HEALTH PROFESSIONS, Vol. 3 No. 3, September 1980 321-360

tional goals, but it is equally an issue of standards, since without a standard, how does one know that the goal has been attained? Grading practices and standards are rarely covered in educators' curricula, a fact which explains but does not condone the vagaries of local-level evaluation processes to which the nation's children are subjected. When educational achievement is assessed by way of state- or national-level examinations, standard-setting becomes an increasingly controversial as well as a complex topic. A recent special issue of the *Journal of Educational Measurement*, devoted to standard-setting, served to highlight a number of basic disagreements among respected authorities in the field, as well as a sorry lack of development of scientific bases for standard-setting methods and processes. In the future, public concern over testing and standard-setting will undoubtedly increase; the health field is unlikely to escape its own share of attention concerning these matters.

In this article, we hope to provide a somewhat different framework for the consideration of the standard-setting problem. We will consider various types of standard-setting methods being used currently in the health professions, develop a classification system for them, review the strengths and weaknesses of various methods, and suggest some directions for future inquiry. We approach this task from psychological and social psychological perspectives as important partners to psychometric concerns. Our essential position is that standard-setting cannot be fully understood until the decision maker(s), the subject area, and standard-setting methods are treated as an interactive system. We will attempt to make a beginning at such an integration here. In so doing, we frequently will be stressing the role of judgment, for judgment is the foundation upon which the methodological superstructure is built. We cannot escape the necessity to make judgments in the standard-setting process, most certainly not by the use of methodology. Nor can we ignore the fact that judgment is a human process and hence includes some degree of variability.

What we can and should do is to strip away irrelevant factors so that judgmental aspects are as systematic (hence reproducible) as possible.

Since standard-setting methodology is closely linked to the type of examination format used, we discuss written examinations and "oral" examinations in separate parts. As the term *oral examination* is not general enough to include all the methods we will be discussing in the second part of this article, we will use *interactive examination* instead. It might also be noted that this article does not address standard-setting issues existing in a third type of examination methodology used in certification and licensure examinations: clinical simulations. These methods, of which patient management problems (PMPs) are the most widely used, are beyond the scope of the present discussion.

STANDARD-SETTING FOR WRITTEN EXAMINATIONS

Written examinations composed of multiple-choice questions are widely used in certification processes of health professionals. These examinations involve a collaboration between two disciplines—the particular health field about which the questions are written and the field of psychometrics, which supplies a theory of testing and scoring. Because psychometrics is heavily based in statistics, it is hardly surprising that Gaussian notions are often adopted for standard-setting purposes as well. These will be discussed below under the heading of normative standards. An alternative of more recent development is that of content-referenced standards, a part of the movement toward mastery learning (see, for example, Block, 1971; Bloom et al., 1971) that goes under a variety of names: criterion-referencing, domain-referencing, and competency-based measurement. The philosophical difference between normative and content-referenced standards is based on the performance of a person relative to peers (normative standards) or on performance relative to a standard developed by expert judgment prior to examination. Both approaches have their positive features, as we will see below, but for the long run the philosophical base of content-referenced standards is stronger (Cahn, 1974).

NORMATIVE STANDARDS

On the surface the application of normative standards is, very simply: Given a distribution of test scores, the decision makers determine that score which separates passing from failing performance, thereby determining the success of all examinees. This process includes several major factors: what aspects of the measurement situation are expected to be stable enough to peg standards to from one year to the next, and the method of cutting score determination.

Stability of Performance: Content or People

A certifying agency seeks as a major objective the reproducibility of its standards over multiple administrations of its certifying mechanism. It would be patently unfair to candidates and patients if individuals with the same level of ability had different likelihoods of success from one year to the next. There are basically two major sources of information upon which expectations of stable performance could be based. One may expect, based on the procedures that are followed to assure content validity of examinations, that content sampling in an examination is equivalent to that of another year. Therefore, it would follow that individuals achieving equal or higher scores to those who passed previously should be certified. But does equivalent content imply equivalent performance on examination? Not at all. The post-World War II era has brought an explosion of new knowledge in the health field. Some have claimed that the half-life of medical knowledge is five years. A rapidly changing knowledge base might mean that one year individuals who answered a question correctly were engaging in creative problem-solving based on knowledge of an emerging area, while at some later time they were simply regurgitating what had become rather common knowledge. The same numerical indices of the difficulty of an item therefore may have different meanings because of different contexts. While one suspects that the effects of rapidity of change are often overstated, nevertheless those health professions examining in areas of rapid knowledge change have to be wary of too great a reliance on equivalence of content as a mechanism for assuring equivalence of standards.

The other possibility is to use the performance of the examinees as the reference standard. Schumacher (see Hubbard, 1978: 68) notes that "relative standards achieve stability of failure rates from subject to subject and from year to year if these standards are based upon the performance of fairly large reference groups, the 'quality' of which is stable over time." A reference or norm group consists of a group of examinees (usually a subgroup of the total number) that is sufficiently well described for decision makers to be able to have some confidence that groups are reasonably equivalent from one year to the next. For physician certification examinations, these descriptors often include a mixture of personal background (e.g., graduate of a U.S. medical school) and training factors. Having defined these groups, the application of a consistent cutting score rule should allow similar standards to be maintained from year to year. Since the rule remains fixed, standards are putatively equivalent over the years. Whether they are so in fact depends on the constancy of the relationship of these quality indicators to the ability of the succeeding groups.

Cutting Score Determination

Within the framework of normative standard-setting, there are several approaches to the determination of the *cutting* *score*—the score separating passing from failing performance. The norming-group approach, to be discussed in detail below, is typically thought of in this context. However, it is also possible to have content-based normative standard-setting. This would occur in a situation in which a standard had been determined for some previous administration of a certifying examination, and the present task was to reproduce that standard with a somewhat different examination.

a. Content-based normative standards: If one expects stability to be based on content, the cutting score needs to be decided by some process of equating the current examination with previous ones. For the moment we will beg the question of how the standard was originally obtained. What concerns us now is the maintenance of that standard in the form of the current examination.

Techniques have been available for a long time (see, for example, Gulliksen, 1950) to equate one examination to another as long as some number of questions are shared by both examinations. For a number of technical reasons, these methods have not been entirely satisfactory. More recently, latent trait theory (Lord and Novick, 1968; Hambleton et al., 1978; Wright, 1979) has provided some promising, more powerful, methods for solving this problem. Latent trait theory, unlike classical test theory, develops an expectation with regard to how examinees will perform on each item by generating a mathematical function of the expected relationship between the ability of examinees and their likelihood of success on that item. Once determined, this information can be used to estimate the ability level of future groups from their performance on the calibrated items. This (item characteristic) curve, or ICC, is said to be the result of the functioning of an unobservable (hence the word *latent*) psychological trait. However, since the theory does not require the verification of the existence of that trait, but merely that the items behave as if one trait were responsible for the data, this writer prefers the term *item characteristic curve* (*ICC*) *theory*. One particular model, first postulated by Rasch (1966) and investigated extensively by Wright and co-workers (Wright, 1977, 1968; Wright and Douglas, 1977; Wright and Panachapakesan, 1969; Wright and Stone, 1979) is particularly simple and powerful. The only information formally needed by the model is person ability, β, and item difficulty, δ; both parameters are placed on the same log scale. Consider the plot of success likelihood by ability (or difficulty) in Figure 1. Item 1, easier than item 2, is such that examinees of ability + 1.0 ("logits") are expected to answer it with a likelihood of essentially 1, while those of ability −1 have a likelihood of essentially zero. A person who has a 50-50 chance of answering items like item 1 correctly is said to have an ability of a. Another person, who has 50-50 odds of answering questions like the more difficult item 2 correctly, has ability b. Since items and ability are on the same scale, we can also speak of item 1 as having difficulty a and item 2 as having difficulty b. If items 1 and 2 are given to another group of examinees, results may be different in that the second group may be more or less successful than the first group, on which we calibrated the questions and established the scale. Since the items were not changed, their inherent difficulty should be unchanged. Therefore any differences in the performances of the groups can be attributed to differences in the abilities of the groups. For example, if group 1 responds to a question as shown by the item characteristic curve for item 1, and group 2 responds to the *same* item with the ICC labeled item 2, then group 2's ability is of magnitude b − a greater than that of group 1. If we construct two tests that share a set of items in common, or give two different tests to a common group of examinees, we can use ICC theory to establish the equivalence between the two tests. This equivalence allows us to maintain standards once they are established, as long as significant shifts in the difficulty of the items or ability of examinees do not occur.

ICC theory in the form of the Rasch model has received practical application to standard-setting problems of several certifying agencies (Schumacher et al., 1979) and will undoubtedly be used by more in the future. The field is still in its adolescence; there are still conceptual and computational problems to be addressed. However, once these are solved it is to be expected that the use of these methods will be widespread.

b. Distribution-based normative standards. This type of standard-setting method is commonly used by medical speciality boards; it involves determing the pass-fail cutoff according to the distribution of scores. This is accomplished by either of two rules: the fixed-percentage rule or the fixed-formula rule. In the former, a fixed percentage of the reference distribution is passed and the rest must fail. Fixed-formula rules, typically stated in terms of the mean (or median) score and the standard deviation of some (reference) group are probably more commonly used. The statistician will recognize that if score distributions are Gaussian in form, the two rules are equivalent. On the other hand, if a distribution is markedly skewed, some anomalies occur with the fixed-formula rule—these largely because of the sensitivity of the standard deviation (and the mean) to extreme scores. Consider the formula for the standard deviation, $SD = [\Sigma (X - X)^2]/N$, where X = a score, $\overline{X}$ = the mean, and N = the number of examinees. Since differences are squared before they are added into the numerator, extremely large or extremely small scores differ greatly from the mean. Thus, if people of very high or very low ability are present in a test population in unusual numbers, the standard deviation is larger than expected. In the typical situation, the distribution is skewed to the left (low) side; this produces an unexpectedly large standard deviation and a lower mean. The effect of the overly large standard deviation and low mean is to lower the cutoff score. On the other hand, skewing means that there are more people below a particular point in the distribution than expected, so the net effect may well balance out, over repeated administrations, to something closely approximating Gaussian-predicted values.

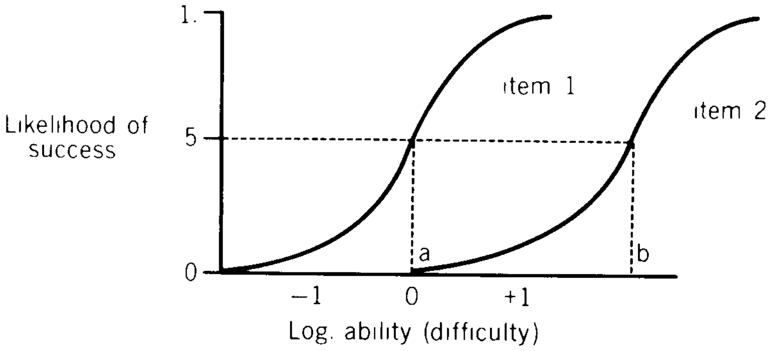

Figure 1: Item Characteristic Curves

But this is faint solace. It is hard to justify using Gaussian statistics on noticeably non-Gaussian distributions. When such occur, either the use of a fixed-percentage rule or normalization of the distribution to Gaussian shape prior to the application of a fixed-formula rule is indicated. Computer programs (SAS Institute, Inc., 1979) are widely available to accomplish this at low cost.

CONTENT-REFERENCED STANDARDS

Since normative standards are *relative*, dealing as they do with a person's rank relative to others, the alternative should, by logic, be *absolute* standards. However, the meaning of an absolute, according to Webster's Seventh New Collegiate dictionary, includes these definitions as: "free from imperfection," "independent of arbitrary standards of measurement," and "having no restriction, exception, or qualification." Since the standards set for any particular examination will be imperfect and arbitrary (in the sense of "depending on choice or discretion") this term applies more as an ideal than as a reality to currently available standard-setting methodology. We prefer the term *content-referenced standards*—standards that define passing or failing in terms of what constitutes minimally acceptable performance on the particular questions of a particular examination. Other terms dealing with the same general question are *criterion-referenced* and *domain-referenced* (Glass, 1978). Both of the latter terms require that domains of knowledge be very completely specified, that the questions sampling a domain be sampling it according to a prespecified plan, and that each domain be sampled. These assumptions are seldom met in medical certification examinations. Consequently, we shall use the term *content-referenced* to refer to standards that are based solely upon the content of the examination, are developed a priori, and it is accepted that the proportion of people passing the examination may well vary from one year to the next.

Let us turn, then, to methods of setting content-referenced standards. The methods below have been applied to medical examinations. For others which may be applicable under some circumstances, see Glass (1978), Meskauskas (1976), and Millman (1973).

Counting Backwards From 100%

This method (Glass, 1978) naturally comes to mind when one thinks about "absolute" standards. Drawing upon universal experience in elementary and high schools, it would seem very simple to set some reasonable lower bound (say 65% correct answers on the test) and be done with the standard-setting problem. Glass (1978: 244), directing his remarks to that experience, has this acerbic comment:

> Many criterion scores appear to have been established in a manner appropriately, though perhaps facetiously, referred to as "counting backwards from 100%." An objective is stated and a test item is written to correspond to it. Since the objective is felt to be important—or else it wouldn't have been stated—its author readily endorses the proposition that everyone should be able to answer the test question based on it; that is,

> the "desired performance level" is 100%. But reason and experience prevail and it is quickly recognized that perfection is impossible and concessions must be made for mental infirmity, clerical errors, misinformation, inattention, and the like. Just how great a concession is to be made becomes distressingly arbitrary, with some allowing a 5% shortfall and others allowing 20% or more.

Generating a standard by picking a number out of the air is, ipso facto, contrary to the goal of fairness in any standard-setting process. What is often not appreciated is the major difference small changes in the percentage-correct passing level may make in numbers passing or failing. Consider the situation at the medical certification examination level. Medical specialists are a highly selected group. The competitiveness of medical school admissions is legendary. Three or more years of post-M.D. training further narrow the field, by both winnowing and self-selection processes. Given all this, not only would it be surprising if there were much variance among individuals, but it would also be an indictment of the educational process. Consider the following real example. On one administration of a subspecialty examination the mean number of true-false questions answered correctly was 80%; the standard deviation of these scores was approximately 7%. For the administration, the committee changed its content emphasis slightly and the mean became 75%—a difference of only 5%. The standard deviation remained roughly 7%. The calculation in Table 1 shows the consequences in the two groups, had the particular board chosen to adopt a 70%-correct passing standard.

The Year 1 cutting score would have been 1.43 standard deviations below the mean, the Year 2 score .71 standard deviations. This would result in a 92% pass rate one year, 76% the next. Is this fair? Well, perhaps. Cahn (1974) examined standard-setting from a philosophical perspective. We have taken the liberty of quoting a version (Cahn, 1973: 14-15) which

TABLE 1
The Effect of Distributional Characteristics on Passing Rates Set by a "Counting Backwards" Strategy—An Example

	Year 1	Year 2
Mean	80%	75%
"passing" score	70%	70%
difference	10%	10%
÷ 7%, the S.D.	1.43	.71
hypothetical % passing in a normal distribution	92%	76%

contained a useful analogy dropped from the later-published version:

> But why is it difficult to explain this degree of variability? It is just what might have been expected, for why should we assume that students always perform in approximately the same way? Consider the fact that each year the NFL drafts hundreds of players from the thousands of football players graduating from college. It is a well-known fact that some years many fine quarterbacks are available, but there are few top-notch defensive linemen. Other years outstanding defensive linemen are a dime a dozen but excellent quarterbacks are rare.

The analogy is very compelling but, upon close scrutiny, not very appropriate. The random variation which resulted in a "large" crop of (five? ten?) good quarterbacks has no relation to the kind of process that produced many hundreds of subspecialists, all theoretically capable of competent performance. It is also well known that it takes three to five years of professional experience to develop a good NFL quarterback; the subspecialists have already had equivalent training experiences. Nevertheless, the quotation is very important because it points out the key role of *expectations* in judging the reasonableness of the outcome. The 16% difference between the hypothetical results is excessive if we expect that the training processes preparing the two cohorts were dealing with students of equal timber and prepared them equally well. If these assumptions are false, then the figure may be just what one might expect. Here again, we come to a realization of the importance of judgment.

The above numerical example, based as it is on actual data, points out that, in some situations, pass/fail rates can be extraordinarily sensitive to apparently minor shifts in the data. If the variance of scores is small, the burden of precision is great not only for the "counting backwards from 100%" strategy, but for all other content-referenced cutting-score methodologies as well.

Ebel's Passing Score Estimation Method

A method for deriving a passing score by considering the characteristics of items along relevance and perceived difficulty dimensions was proposed by Ebel (1972). His example uses four relevance categories: essential, important, acceptable, and questionable. Three difficulty levels (easy, medium, and hard) are specified. Since the two dimensions are operationally independent, this forms a 4 × 3 matrix. Each question is classified by expert judges into that cell which represents the combination of these two parameters felt to be characteristic of the question. Once all questions are classified, the next task is to review the questions placed into each cell and to decide what percentage of these questions candidates should answer correctly. As Ebel gives no particular new method for accomplishing these decisions, we expect that judges will adopt what amounts to a "counting backwards from 100%" strategy. Finally, the number of questions in each cell is multiplied by the appropriate percentage, and the sum across all twelve cells is divided by the total number of questions to derive the lowest passing score. The test itself is scored in the usual way: 0 credit for wrong answers, 1 point for correct answers.

On reflection, it seems quite clear that this method will produce judgments which are more thoughtful than those of the counting backwards strategy applied to the exam as a whole. This is an expected consequence of a careful scrutiny of the specific content the examinees will see. On the whole, however, the questions raised several years ago (Meskauskas, 1976: 138) still hold.

> In Ebel's method, the judge must simulate the decision process of the examinee to obtain an accurate judgment and thus set an appropriate standard. Since the judge is more knowledgeable than the minimally-qualified individual, and since he is not forced to make a decision about each of the alternatives, it seems likely that the judge would tend to systematically over-simplify the examinee's task. Whereas the examinee has to choose among a number of alternatives, the judge's tendency is to consider only the correct answer in relation to the stem. Thus, the judge's rating process is transformed from a consideration of how difficult a question is when considered in relation to its distractors to merely the difficulty of the correct answer. Even if this occurs only occasionally, it appears likely that, in contrast to the Nedelsky method, the Ebel method would allow the rater to ignore some of the fine discriminations that an examinee needs to make and would result in a standard that is more difficult to reach. However, perhaps the most troublesome feature of Ebel's method is the requirement that a separate judgment be made about the percentage of items in each cell along the relevance/difficulty continuum that the minimally-qualified examinee should be required to answer. Unless there are external criteria upon which to base this judgment, it seems entirely arbitrary.

The Nedelsky Minimum-Pass-Level (MPL) Method

The most widely used method of setting content-referenced standards was developed by Nedelsky (1954). This technique, which involves judgments about the plausibility of each of the alternatives supplied with a multiple-choice question, was developed for use with a university physics course. Since a number of different instructors taught the same subject matter, there was need for a standard-setting process that represented consensus. There are a number of references to the use of this of this method in the health professions: Andrew and Hecht, 1976; Taylor et al., 1971; Levine and Forman, 1973; Meskauskas and Webster, 1975. The method involves judging each incorrect answer to a multiple-choice question with regard to whether the examinee who has just enough knowledge to pass should be able to recognize that alternative as incorrect. The incorrect alternatives which should be recognized as such are indicated; the reciprocal of the number of remaining alternatives determines the minimum passing level (MPL) for that item. Thus, if, in a five-choice question, two alternatives are marked as ones that should be recognized by the minimally competent candidate as incorrect, three remain. The MPL for the question is the reciprocal of 3—1/3, or 33. When all items have been judged in this way, the sum of MPLs across all the items in the test determines the lowest passing score.[1] The candidate gets one point credit for each question answered correctly; no penalty is assessed for wrong answers. Thus, implicitly, a person who passes is expected to correctly answer one-third of all questions with MPSs of 1/3, one-half of those with MPLs of 1/2, and so forth.

It might be noted that while this method can be used with true-false questions, in such a case it basically becomes a question of deciding whether the incorrect response is acceptable. Thus, if the Nedelsky method is used with this type of question, the MPLs can only take on values of 1/2 and 1. The passing score is therefore determined by the relative proportions of these two values.

Discussion: Normative and Content-Reference Written Exams

As practiced, a normative standard-setting process has a major advantage over content-referenced standards: simplicity of generation. However, we wonder whether this approach, if based upon the performance of a group that is only loosely described, really assures equivalence over administrations of an examination. Normative standard-setting assumes that, from one year to the next, it is possible to select a subgroup from the total candidate pool that is of equivalent talent and has been exposed to the same quality of educational process. We are unaware of any published work that investigates whether the typical kinds of readily available descriptors do, in fact, result in the identification of equally capable reference groups from one year to the next. In the absence of supportive data, we believe that one can have no assurance that normative standard-setting methods result in truly equivalent standards from one year to the next. Research is sorely needed here.

Content-referenced procedures base standards upon consideration of the particular material that candidates will be asked to deal with. This is laudatory, since a person's success is determined entirely by his or her proficiency with that material rather than on standing relative to other examinees. However, that does not, ipso facto, mean that content-referenced standards are more fair. Fairness depends more on the details of the method and procedures involved in the standard-generation process. To date, the outcomes of work with the Ebel and Nedelsky methods have yielded mixed success. Some applications have been reasonable and useful, others have yielded unrealistically high fail rates. A high degree of consensus has not been found, typically, among the standards set by different judges. Thus, while it cannot be said that methods exist which are guaranteed to work, these results are useful directions for future development.

One of the major implications of work to date is that, lacking careful design of the judge's task, efforts to treat judges as "black boxes" whose standards would be extracted by Ebel or Nedelsky procedures are likely to be failures. Glass (1978) points this out rather strikingly. In retrospect, this is no surprise. A standard-setting method is incomplete without consideration of the psychological context of the judging process as well as certain psychometric issues. It is hoped that a comprehensive approach to the judging process will contribute a workable standard-setting methodology. Some of the aspects to be considered are discussed below.

Task Congruence

The importance of context on behavior is a widely accepted principle. Further, Miller (1956) and others studying human information processing (see Klatzky, 1975) have shown that while humans can deal with extremely complex phenomena, they can deal with only a few aspects of it at any time. Therefore can we expect that a judge is engaging in mental tasks that are equivalent when attempting to set a standard on an entire examination ("counting backwards" method) or a question/right answer combination (Ebel) or a question with each of its associated alternatives (Nedelsky)? We think not. Andrew and Hecht (1976) compared equivalance of outcomes of the Ebel and Nedelsky methods, using comparable questions and groups of judges in a counterbalanced design. Within each method, differences between the standards set by the two groups of judges were small. Differences between the results of the two methods were large: The Ebel method resulted in a standard of 68% correct answers required to pass, the Nedelsky in 49%. The investigators interpreted this as arising from differences in the method, but another interpretation is to view this as arising from differences in the psychological tasks of the judges associated with each method.

Level of Detail

How might we design the judges' task? The first step would be to see that the judge has to deal with the questions at the same level of detail that the candidate does. The candidate must choose between the alternatives provided for each question, taking into account not only what he may know about the subject, but any cues that may be available. Such cues may be written into the question or they may present themselves because one or more options the examinee might consider are not available. Judges have to deal with the same stimuli to be able to evaluate the candidate's performance.

What Should Be Judged?

There needs to be a conceptual match between the task of the examinee and that of the standard setter. Since examinees are asked to determine the correctness of each alternative, judges must do so as well. Judges should not be asked to determine the relevance and difficulty of questions and to base standards solely on them. Relevance and difficulty are really weighting issues; they allow adjustment of standards for vagaries of the content of the examination. That in itself is good, but it should not be thought of as a complete approach. The standard itself must be defined also.

Minimal Competence

As originally defined, judges using the Nedelsky approach had to develop a notion of a borderline student so as to be able to judge how that person might perform on a question. In health professions, this becomes the notion of a "minimally competent" professional. This is a very difficult concept. Whom can judges use for a model of the minimally competent? Even if a good model were available, such a person might do well in some areas and poorly in others, making it hard to utilize him or her in judging any specific question. This is an important but solvable problem; it suggests that a good deal of attention will need to be focused on the mental set given the judges.

Interjudge Variability

Very little has been written about how to deal with examiner variation in the standard-setting process. What little there is suggests that this is an important factor to take into account. Some judges require a very high level of performance, others a lower one. This should dismay no one. Since in any human activity there will be some error, so will there be in standard-setting. Accepting this, our goal becomes to reduce random error while retaining real differences of opinion. Three techniques suggest themselves. The first is to match carefully judges' experiences to the behavior being judged. For example, a tertiary-care subspecialist may not be the best person to serve as a judge on a primary-care exam. The second is to make sure that the number of judges used is sufficient to assure stability. Based on personal experience, we suggest a minimum of ten. Third, it may be useful to "handicap" judges. Recognizing that the overall standards are determined by, and are the responsibility of, the certifying body as a whole, it may be useful to establish the typical difference in standards between one judge and others. Future standard-setting work by that judge might be "corrected" to account for this difference. Stanley (1961) had some useful contributions to technique in this area.

Psychosocial Aspects

The standards set by any judge or group of judges are not only a product of their expertise but also a product of the social psychological environment in which the decisions are made. Two important factors which we expect to affect the decision-making process are the environmental features associated with the standard-setting process and the influences of the group on the standard setter.

In a landmark series of studies, Lorge (1936) presented subjects with a list of quotations followed by the names of two authors. Subjects were expected to rate each quotation on a five-point agree-disagree scale and to choose the true author. Two weeks later the same quotations were presented; however, they were attributed to only one of the authors. When the author was the same as the one chosen previously by the subject, the ratings of the quotation remained the same; when the author was different than the one the subjects chose previously, then there were significant shifts in the rating of the quotation. Lorge explained this phenomenon in terms of the attachment of a positive or negative feature (in this case prestige) to the object being rated. Asch (1948), on the other hand, argued that the attachment of such a feature to the object being rated, changed the object so that subjects were in fact responding to a different task. In either case the implications for standard-setting are clear. To the extent possible, those unwanted features of the entire social psychological environment within which the judgment was made, should be identified; "blind" judgment with respect to these factors should eliminate their effects. Other features which may be important aspects of the standard-setting process should be available in the same form to all the judges.

The nature of the group also influences the standard-setting process. A prevading aspect of all human behavior is the drive to evaluate one's own opinions and abilities by comparing them with those of others in the absence of objective, nonsocial means of valuing (Festinger, 1954). This drive expresses itself to varying degrees, contingent on the characteristics of the group and the group members. Group characteristics that influence individuals include factors such as closeness, cohesiveness, and group attractiveness. On the other hand, particular members influence the overall group process in relation to the relevance of their expertise, leadership ability, power outside of the group, and similarity to other group members. These findings strongly suggest that careful attention be given to psychosocial factors so as to assure the reproducibility and high quality of standards.

Facilitation of Judging

There is no armamentarium of tried-and-true methods for facilitating standard-setting judgments, although we suspect that social psychology has developed a few that might be applicable. Until better methods come along, it may be useful to make decisions by means of series of questions posing paired contrasts. Is equivalence of standards best maintained by equilibrating content or by drawing "equivalent" groups of examinees? Should we have normative or content-referenced standards? Fixed-percentage or fixed-formula cutting score? And so forth.

STANDARD-SETTING FOR INTERACTIVE (ORAL) EXAMINATIONS

In the history of evaluation, written examinations are relative newcomers. Preceding their development, the "oral" examination was the accepted method of evaluation of an individual. Indeed, examinations involving face-to-face interaction are still widely used for graduate-level evaluation, such as examinations of advanced-degree candidates and for a number of medical specialty board examinations. Here, the term *interactive examination* will be used for examinations which involve direct contact between a candidate and an examiner. The term subsumes the traditional medical oral examinations, but is intended to be broad enough to include other interactive examination designs as well.

Widespread use of interactive examinations continues in the face of criticism regarding their measurement properties (Abrahamson, 1975; Marshall and Ludbrook, 1972; Evans et al., 1966; Foster, 1969). While many of these authors and others (American Board of Medical Specialties, 1975; Van Wart, 1974) present suggestions for improvement on their experiences with implementing well-conceived designs, it seems that implementation of some of these suggestions is lagging. Thus, despite the fact that a number of boards have redesigned interactive examinations, a number of boards have abandoned them. Abandonment has often been based on the logistics of administering ever-increasing numbers of examinations, and, indeed, this is a serious concern. From the perspective of this discussion, however, redesign is an attractive alternative because for all their faults, interactive examinations include a key feature: they incorporate "absolute" standards. The examiner grades the examinee against his or her own internal standards, within the confines of the examination methods and policies. Thus, in written examinations, the standard exists as a reality external to judges and the examination, while in interactive examinations the standard is internal to the judge. Thus, the judge, along with the various factors that impinge on the judge, becomes the major focus. In the remainder of this section we will develop a perspective on the design of interactive examinations which will, we hope, further understanding in order to help maintain the strengths of the interactive approach while dealing in a positive way with the criticisms.

This approach to the analysis of interactive examinations will be based upon consideration of the various sources of differences in performance among individuals or, more technically, a variance components view. The reasons for doing so are twofold. First, the adoption of a variance components perspective drawn from experimental research, emphasizes the existence of empirical, quantifiable questions, the answers to which are helpful in the design of interactive examinations. Second, evaluations of the measurement properties of these examinations use the "sources of difference" notion in the determination of validity and reliability. Since the measurement properties, especially reliability, are such a key factor in the criticisms leveled at interactive examinations, let us begin with them.

The concept of reliability is eminently reasonable and useful; however, it should not be used as the sole criterion against which to judge the fitness of an examination. Many reliability formulas exist for application in various situations. When the assumptions made in the derivation of these formulas are met, they produce reasonably equivalent results (Raju, 1977). For interactive examinations, the applicable reliability statistic (Winer, 1971) seeks to assess the reproducibility of examination scores by taking into account the degree of relationship among the components of that total examination. The model that underlies this statistic is unidimensional; that is, it divides the total variability among the components into one part which is due to positive association between the components, and another part which is considered error or noise. Thus, it defines all variance other than that due to this single common factor as error. The reliability coefficient is the ratio of common variance to total variance. (Notice that it is impossible to assess reliability if only one measurement is made, since there nothing to relate that measurement to.) If the results are caused by two or more well-measured but unrelated factors, the unidimensionality assumption built into all reliability assessment will consider all variance beyond that assessed by the first factor to be unstable error variance. Thus high reliability implies high reproducibility, but low reliability may mean nothing more than an inappropriate selection of analytic model. It is very possible, therefore, that the low reliability often reported in the literature for interactive examinations may be due to such an event.

Let us now turn to the question of design, keeping the above in mind. In the discussion to follow, the term *factor* will be used to identify broad classifications of sources of differences on the outcome of interactive examinations. Each factor may contain one or more variables or dimensions. Four factors can be identified: examination environment / design, examiner, examinee, and clinical material. The strategy will be to consider the implications of each factor for the design of interactive examinations. The goal will be to attempt to identify those aspects which affect the examiner-examinee encounter, and to suggest ways of dealing with them when appropriate. Following this discussion, various models of interactive examinations will be discussed.

Factor 1a: Examination Environment

Evidence relating to the importance of environment on medical interactive examinations is either nonexistent or, at best, fragmentary. Yet, environmental variables may well be more important than is generally appreciated. For example, Evans et al. (1966) found a correlation of .67 between the percentage of words spoken by the examinee during the interaction, and the grade received. While one would surmise that those who spoke more also had a greater amount of substantive information to impart to the examiner, it seems unlikely that this could account for that strong relationship. Possibly, the "seductive speaker" phenomenon (Ware and Williams, 1975) is at work. Since we probably should not allow this type of personality variable to affect measurements of clinical competence, we need to design the interaction environment in such a way that, to the extent possible, this type of effect is controlled.

Another factor to consider is whether candidates should be examined under conditions that are likely to evoke optimal behavior, typical behavior, or pressured behavior. In the psychological literature, the effects of stress on performance have been well documented and they would be expected to apply to interactive examinations. Increasing stress facilitates performance to a point; additional stress has the effect of decreasing performance. While specialists who typically practice under a great deal of pressure may or may not wish to design a pressured interaction, certainly there is little justification for so doing among those specialties where stress is not a practice factor. Whatever decision is made in this regard, the type of environment should be as equal as possible for all examinees.

Factor 1b: Design Considerations

The objective of interaction design is to develop an evaluation procedure that measures the desired candidate attributes (and only those attributes) and is practical. Execution is much more difficult. The so-called "reliability" problem keeps cropping up. However, let us not be misled. The evidence is strong that, within a particular interaction consisting of examiners observing or interacting with an examinee on the same clinical problem, reliability is high. Evans et al. reported reliability coefficients ranging between. 77 and .85 for inter-observer reliability. McGuire (1975) reported coefficients of .7 or .8 in another study. Van Wart (1974), using six reliability indexes, concluded that three examiner teams were "extremely reliable" in terms of the grades they awarded the same examinees, five were "reliable," and four "unreliable." Thus, there seems to be little gain in having examiners work in pairs, since agreement will be good. On the other hand, relationships across interactions are not likely to be very high. When scores across pairs of different cases are considered, reliability coefficients on the order of .4 are found, and correlations tend also to be of the same or lower magnitude (Marshall and Ludbrook, 1972; Meskauskas, 1975). If we are open to the possibility of behavior being influenced by more than one attribute operating simultaneously, then the above findings are explainable by performance which is consistent but situation-specific. The implication for design is that more interactions are needed to assure a certifying body that the total picture obtained about a person is accurate.

In situations where expansion of the number of interactions to which a person is exposed is limited by examiner time, two solutions are available: shorter interactions and variable-interaction examinations. The shortening of interactions suggests itself if examiners commonly make up their minds on the grades they will give well before a session is completed. Variable-number examinations are made possible by the ready availability of sophisticated handheld calculators. After an examinee undergoes two interactions, the likelihood of a change of status (pass to fail and vice versa) could be computed and, if the results are sufficiently clear-cut, the candidate could be excused from further interactions. This approach is particularly useful for those who score consistently well. However, if such a strategy is followed, care must be taken to make sure that examiners are unaware of the status of their examinees, lest it influence their judgment.

Some useful suggestions regarding design and utilization of interactive examinations have been made by Ebel (1972: 266-267):

> Avoid using oral examinations if a written examination can be devised to do the job with reasonable effectiveness.
>
> Define clearly the purpose of the examination and the basis on which the examinee performances are to be judged.
>
> Prepare the examinees to give an accurate account of themselves on the examination.

In summary, even though our knowledge of the effects of the examination environment and other design issues is far from complete, the present evidence suggests that they are likely to be very important. Three suggestions have been offered. The first is to decide what environment will be conducive to the observation of the desired behaviors. The second is to control the environment to assure consistency of interaction. The third is to use enough interactions to assure a stable decision.

Factor 2: Examiners

In interactive examinations, the examiner is the major examination delivery factor. Given this important role of the examiner, the paucity of research on the impact of examiner characteristics on the outcome of the interaction is disturbing. Our comments will relate to four issues: who should be an examiner; how to deal with differences in what examiners know; how to deal with differences in examiners' standards; and psychological factors impacting the interaction.

The first of these, who should be an examiner, is an important question. Our perspective cannot include a full appreciation of the medical aspects of the choice. We can offer the suggestion that the individuals chosen should be mature, broad, experienced professionals with a type and level of expertise that closely matches the area being examined. It must also be recognized that the certifying board will be making an investment in the examiner, so the relationship should be long enough that the investment can be recovered.

No matter how excellent the cadre of examiners, there will be differences in their expertise. This must be acknowledged beforehand, and provision made for the examiners to familiarize themselves fully with the medical content of the examination. Otherwise, examiners may be synthesizing the material while the interaction is in progress, and they may miss some of the more subtle aspects of the candidate's performance. Also, examiners need to be educated about the essentials of the evaluation process and the board's procedures in that regard. This is a time-consuming, perhaps unrewarding, but nevertheless necessary process. The College of Family Physicians of Canada (Van Wart, 1974) has developed a particularly good approach to this problem. The inclusion of several practice interactions, followed by a critique, is a highly recommended means of consolidating all of the information a new examiner faces.

Differences in standards among examiners is commonplace in the folklore of interactive examinations. Some of these differences are undoubtedly caused by factors that can be controlled by careful design of the interaction. But ultimately the fact must be faced that irreducible differences of opinion are inherent in the judgment of most situations. This is perfectly acceptable in some matters, but fairness is questioned when certification decisions are affected. The key to the resolution of this problem lies in a consideration of the locus of responsibility for certification. Since that responsibility is the board's and not the examiner's, it makes sense for the board to decide that the standards set by some are too stringent and by others, too lenient. Thus we can contemplate a data collection system that "handicaps" examiners. These handicaps could then be automatically applied to the scores of examinees or, alternately, as a correction for those whose performance falls near the pass/fail dividing line. Handicapping is likely to be unpopular with examiners. It must be stressed, however, that the final responsibility for certification rests with a board rather an individual examiner, and therefore that board must take whatever steps it sees fit to assure all candidates an equal opportunity for success.

While there is not a great deal of information concerning the psychological effects of the examiner on the interaction, a considerable amount of work has been done on an analogous situation: the effects of an experimenter on the behavior of subjects. Since several parallels can be drawn between the interactive testing situation and the experimental situation, this work appears to be particularly relevant to the issues at hand.

Several studies, especially those conducted by Rosenthal (1964, 1966) and his colleagues, have indicated that factors such as the warmth, expectations, sex, race, and status of the experimenter affect the behavior of subjects. For instance, they found that a warmer experimenter tends to elicit responses that are more adequate and agreeable than a hostile experimenter. In addition, a high-status experimenter tends to elicit responses that are less pleasant and more conforming than a low-status experimenter (Wrightsman, 1972).

The major thrust of Rosenthal's investigation has been an examination of the impact of experimenter or teacher expectations on behavior. In the classic work, *Pygmalion in the Classroom* (1968), Rosenthal and Jacobson randomly selected a group of students from an elementary school and told their teachers that these pupils were about to "bloom." By the end of the year the selected students had gained significantly more in achievement and I.Q. than had the other pupils. Other studies have confirmed the importance of experimenter expectations.

While the effect of these psychological factors cannot be completely eliminated from human interactions, their impact can certainly be minimized in the testing situation. Rigorous training and selection procedures can go far toward this end. In addition, the examiner should not be informed of the educational and personal characteristics of the examinees.

From the perspective of the particular board, it is important to overcome these obstacles, because doing so increases the fairness of examinations. Moreover, we must recognize the certification helps to perpetuate the specialty through maintenance of its values and standards. Abrahamson (1975: 27) has reminded us of the importance of interactive examinations from this perspective:

> I would see one other phase in the case for the oral examination, and I really think it is important, although I present it somewhat apologetically. This is concerned with the rites of passage, the ritual associated with achieving another level of competence or recognition of that level of competence. There may be some who say, "My God, that's all it is, a ritual," to which I respond that as a ritual it may be serving a very important purpose. It may encourage a quality of practice just through the pride of belonging that is generated by the passage through a ritual. Lasting professional relationships and professional identification may be established during the course of an oral examination. The candidate recognizes that he is admitted to a group, that the group holds certain values, and therefore he adopts those values.

Factor 3: Examinees

From the perspective of the examinee, the best success strategy in approaching an interaction is to present himself, his knowledge and skills, in the best possible light. There can be very little question but that some strategies adopted by the candidate are likely to be more successful than others. One hopes that the degree of impact such interpersonal strategies have on the final result are minimal, but under some circumstances it could be a problem. Examiners need to be clearly aware of the possible effects of these aspects of behavior on their opinions of a candidate. In particular, as it becomes difficult to make a distinction between two grades, examiners may resort to paired-comparisons considerations. Is person A better than person B on this? What was given as a grade for B when that person was examined? Is B worse than A, and if so by how much? This type of decision process is inescapable in some circumstances, and may be quite useful if carried out carefully.

Since all examinations are stressful, and interactive examinations are particularly so, examinees may adopt hostility, excessive assuredness, or other mechanisms as an unwitting way of handling the situation. Unless heavy stress is characteristic of the specialty, such reactions are undesirable because they lead to an uncharacteristic impression of the examinee's typical functioning. Examiners will wish to handle these mechanisms early in an interaction lest an inaccurate perception result.

Factor 4: Clinical Material

Recent studies with clinical simulations (Elstein et al, 1978; Meskauskas and Norcini, 1979; LaDuca, 1979a, 1979b) have shown that performance varies very markedly from one case to another. While it has been common to assign this to variation among candidates' knowledge of the material, it may well be that different cases draw on different mixtures of knowledge, problem-solving abilities, skill, and so on. In any case, one would strongly suspect that these findings of situation-specific behavior apply to interactive examinations as well. Increasing the number of cases to which a candidate is exposed is one way of handling this. However, this approach should be supplemented by attention to the clinical material itself, along the lines of assuring that the material is likely to evoke the desired behaviors for observation, and that all examinees have similar material to deal with. This is easier said than done, but the attempt to meet these requirements is an important aspect of the assurance of a fair and equitable interactive examination.

MODELS OF INTERACTIVE EXAMINATIONS

The purpose of this section is to present a classification system of the types of interactive examinations reported in the literature. Within a particular examination environment, the players are the candidate, examiner, and the patient/clinical material. The interactive examination contains many of the essential features of the medical encounter (American Board of International Medicine, 1979). A patient with one or more medical illnesses is presented to the candidate, either personally or as a case data base. The candidate applies his or her problem-solving abilities, intellectual tools, skills, and attitudes, to the tasks of data-gathering, problem definition, and therapy. The examiner observes this process and reaches conclusions about how well the candidate does. In addition, the examiner may seek to assess other aspects (Abrahamson, 1975) such as the person's acceptability to the specialty, ability to assume and perform the appropriate role, ability to respond to a change in situation, and ability to react quickly. We will first present the models themselves and then discuss their relative utility for each of these aspects.

The traditional medical oral examination is shown in Model 1 (see Figure 2). The essence of this model is to allow the examiner to observe the candidate-patient interaction and to have access to the patient for independent verification of findings. The interaction between examiner, patient, and candidate occurs within the same time period.

The examiner can observe the process of interaction between candidate and patient, so the candidate's approach can

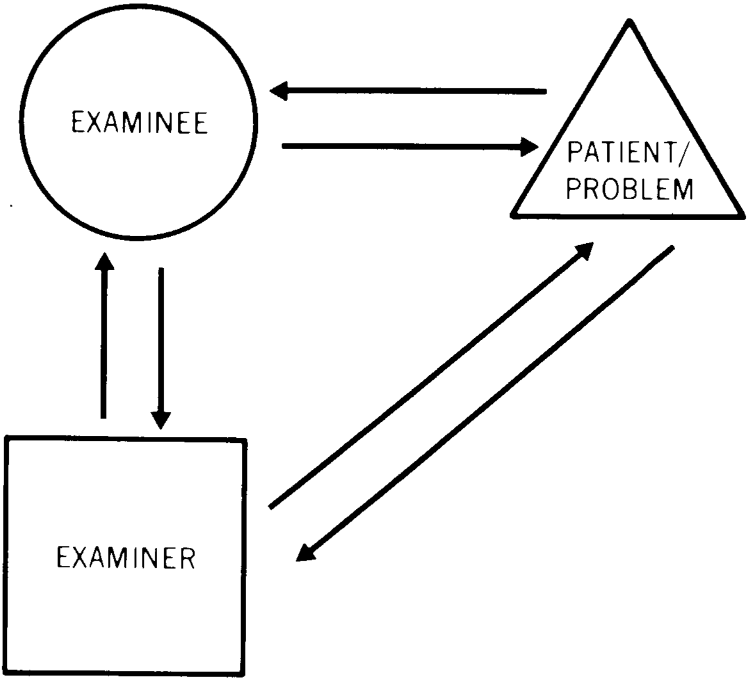

Figure 2: Model 1

be assessed instantaneously. Access to the patient allows the examiner the opportunity to validate the findings which the candidate reports. On the other hand, the presence of the patient may well have some influence on the candidate-examiner interaction. Thus, after candidate and examiner have satisfied themselves that they fully understand the patient's problems, it would be best to adjourn the candidate-examiner interaction to another room. The presence of the examiner undoubtedly has some effect on the candidate's handling of the medical problem. To the extent that the candidate behaves in an atypical fashion in response to the presence of the examiner,

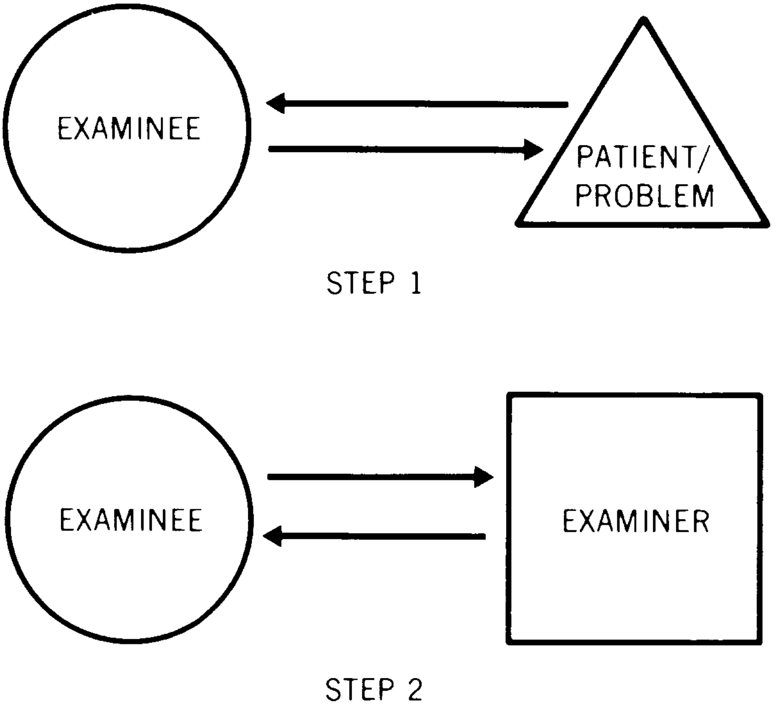

Figure 3: Model 2

the ability to generalize from this one encounter to other medical encounters will be limited.

One response to the potential for examiner influence inherent in the previous model is shown in Model 2 (see Figure 3). The candidate interacts with the patient, then subsequently interacts with the examiner. The examiner has not seen the patient but of course is knowledgeable in the general area of the patient's problem. Evans et al. (1966) utilized this model for the evaluation of third-year medical students; at least one medical specialty board has utilized it as well. The influence of the time delay between steps 1 and 2 has not been studied. It allows time for integration or synthesis to take place if that did not occur at the bedside. This is particularly useful for the assessment of students or when the patient has a rare condition, because it allows for the measurement of optimal performance.

The fact that the medical encounter is not observed removes any stress that might be caused by the examiner, allowing the encounter to proceed in a more natural way. In the event that the candidate responds to observation with anxiety-related atypical behavior, this is useful.

The inability to observe has several consequences for the examiner. Cues from subtle, inexplicable, or patient-specific behavior which may be useful for orienting the examiner to key aspects of the encounter are unavailable. The examiner can only cast a wide net at the beginning of the examination, and refine his or her approach as time goes on. Further, note that this model relies completely on the candidate's expressiveness as input to the examiner. Habitual behaviors may well not be reported to the examiner. If the board wishes to assess interpersonal relationships, the examiner will need to form an impression about physician-patient relationships from physician-examiner relationships—a dangerous practice at best.

Model 3 (see Figure 4) will be recognized as a refinement of the previous one. Each of the diadic interactions occurs in a separate time period. It gives the examiner an opportunity to interact with the patient, to assess the medical problems he or she may have. Input from colleagues can also be received, assuring that the examiner has considered all aspects of the problem. The only negative feature is that considerably more examiner time is required. If the patient can be seen by more than one candidate, the investment of examiner time can be made more efficiently, however.

Throughout the discussions of the above models, it has been understood that the triangle could stand for either a patient or a clinical case. In Model 4 (see Figure 5), only the second of these is possible, as the examiner presents the case material.

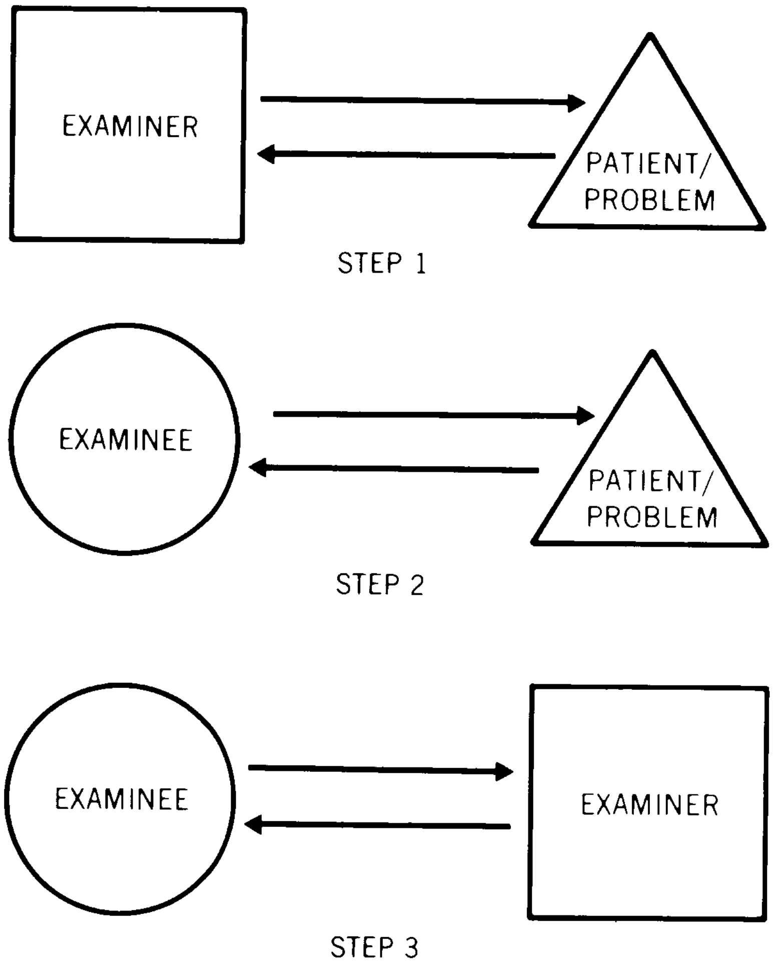

Figure 4: Model 3

Van Wart (1974) described such a design used by the College of Family Physicians of Canada. The clinical content and examination protocols are determined in advance. The ex-

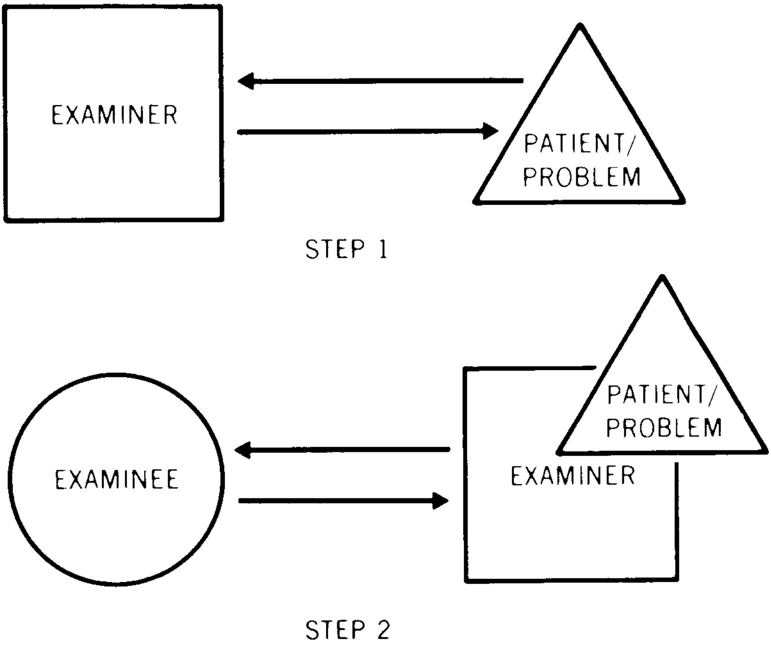

Figure 5: Model 4

aminer becomes familiar with the material to be presented, and plays the role of the patient to the extent of providing the candidate with whatever information may be sought about the patient. The examiner does not question the candidate, and can only depart from his or her role if the candidate is unable to proceed further. In that event, the examiner provides the needed help and the roles are resumed. Such a model elicits the logic of problem-solving, but of course cannot assess interpersonal aspects of the clinical encounter.

DISCUSSION

The preceding has emphasized the multidisciplinary aspects of interactive examinations—as psychological and psychometric processes occurring within a medical framework. Even though they are incompletely understood at present, interactive examinations hold very bright promise, for they provide a direct expression of the standards of the profession when designed and executed well. While some may be discouraged by the complexity of the picture painted here, it seems to us that if subtle and complex behavior is to be assessed, then for the present an evaluation mechanism which employs human judges is inescapable.

This treatment has emphasized the many intricacies which may effect the evaluation process. If all were operating at once, stability of measurement would be impossible. That they do not, in the usual case, is attested to by the fact that reliability coefficients (for two-interaction examinations) reported in various studies tend to be in vicinity of .4. Thus, there is agreement between the results of one interaction and another. We would hope that a consideration of some of the potential sources of variability, and the various tradeoffs inherent in the four models shown, would be useful in increasing the level of agreement in future examinations.

We also wish to make clear that the status of research in this area is appalling. The psychometric problems are clear; however, psychometric theory does not hold the key to solutions. Rather, these must come from a melding of several perspectives. For these reasons, we have stressed a classification system which can incorporate multiple perspectives and is drawn from experimental research. We urgently need to address the question of the relative impact of the four factors (environment/design, examiner, examinee, and patient) on the outcome of the interactive examination. In an area like this, practice will almost always precede theory; those conducting the practice, however, have an obligation to further the understanding of that practice.

NOTE

1. The method described by Nedelsky involves a further consideration which is omitted here, since none of the applications in the literature used it. See Nedelsky (1954), Glass (1978), or Meskauskas (1976) for a more complete discussion.

REFERENCES

ABRAHAMSON, S. (1975) "The oral examination: the case for the case against." Presented at American Board of Specialties Conference on the Oral Examination, Chicago, March.

American Board of Internal Medicine (1979) *Clinical Competence in Internal Medicine.* Philadelphia: American Board of Internal Medicine, 1979.

American Board of Medical Specialties (1975) "Conference on the oral examination." Chicago, March.

ANDREW, B. J. AND J. T. HECHT (1976) "A preliminary investigation of two procedures for setting examination standards." *Educ. and Psych. Measurement* 36: 45-50.

ASCH, S. E. (1948) "The doctrine of suggestion prestige, and imitation in social psychology." *Psych. Rev.* 55: 250-276.

BLOCK, J. H. (Ed). *Mastery Learning: Theory and Practice.* New York: Holt, Rinehart, and Winston, 1971.

Bloom, B. J., J. T. HASTINGS, and G. F. MADAUS (1971) *Handbook on Formative and Summartive Evaluation of Student Learning.* New York: McGraw-Hill.

CAHN, S. M. (1974) "Philosophical reflections on evaluation." *American J. of Medicine* 57: 152-156.

——— (1973) "Philosophical reflections on evaluation." Presented at the annual conference of the American Board of Internal Medicine, August.

EBEL, R. L. (1972) *Essentials of Educational Measurement.* Englewood Cliffs, NJ: Prentice-Hall.

ELSTEIN, A. S., L. S. SHULMAN, and S. A. SPRAFKA (1978) *Medical Problem Solving: An Analysis of Clinical Reasoning.* Cambridge, MA: Harvard Univ. Press.

EVANS, L. R., R. W. INGERSOLL, and E. J. SMITH (1966) "The reliability, validity, and taxonomic structure of the oral examination. *J. of Medical Education* 41: 651-657.

FESTINGER, L. (1954) "A theory of social comparison processes." *Human Relations* 7: 117-140.

FOSTER, J. T., S. ABRAHAMSON, S. LASS, R. GIRARD, and R. GARRIS (1969) "Analysis of an oral examination used in specialty board certification." *J. of Medical Education* 44: 951-954.

GLASS, G. V. (1978) "Standards and criteria." *J. of Educ. Measurement* 15: 237-261.

GULLIKSEN, H. (1950) *Theory of Mental Tests.* New York: John Wiley.

HAMBLETON, R. K., H. SWAMINATHAN, L. L. COOK, D. R. EIGNOR, and J. A. GIFFORD (1978) "Developments in latent trait theory: models, technical issues, and applications." *Rev. of Educ. Research* 48: 467-510.

HUBBARD, J. P. (1978) Measuring Medical Education: *The Tests and the Experience of the National Board of Medical Examiners.* Philadelphia: Lea & Febiger.

KLATZKY, R. L. (1975) *Human Memory: Structures and Processes.* San Francisco: Freeman.

LaDUCA, A. (1979a) "The structure of competence in health professions." Presented at annual meeting of American Educational Research Association, San Francisco, April.

——— (1979b) "Solving the medical problem solving problem." Presented at the Association of American Medical Colleges Conference on Research in Medical Education, November.

LEVINE, H. G., and P. J. FORMAN (1973) "A study of retention of knowledge of neurosciences information." *J. Medical Education* 48: 867-869.

LORD, F. M. and M. R. NOVICK (1968) *Statistical theories of mental test scores."* Reading MA: Addison-Wesley.

LORGE, I. (1936) "Prestige, suggestion and attitudes." *J. of Social Psychology* 7: 386-402.

MARSHALL, V. R. and J. LUDBROOK (1972) "The relative importance of patient and examiner variability in a test of clinical skills." *British J. of Medical Education* 6: 212-217.

McGUIRE, C. (1975) "Experiences with orthopedic surgery." Presented at the American Board of Medical Specialties Conference on the Oral Examination, Chicago, March.

MESKAUSKAS, J. A. (1976) "Evaluation models for criterion-referenced testing: views regarding mastery and standard-setting." *Rev. of Educ. Research* 46: 133-158.

——— (1975) Cardiology and the American Board of Internal Medicine. Presented at the American Board of Medical Specialties Conference on the Oral Examination, Chicago, March.

——— and J. J. NORCINI (1979) "Validity PMP/CASE redundancy, and some comments about the structure of relationships among measures. Presented at the Association of American Medical Colleges Conference on Research in Medical Education, Washington, D.C., November.

MESKAUSKAS, J. A., and G. W. WEBSTER (1975) "The American Board of Internal Medicine recertification examination process and results." Annals of Internal Medicine, 82: 577-581.

MILLER, G. A. (1956) "The magical number seven, plus or minus two: some limits on our capacity for processing information." *Psych. Rev.* 63: 81-97.

MILLMAN, J. (1973) "Passing scores and test lengths for domain-referenced measures." *Rev. of Educ. Research,* 43: 205-216.

NEDELSKY, L. (1954) "Absolute grading standards for objective tests." *Educ. and Psych. Measurement* 14: 3-19.

RAJU, N. S. (1977) "A generalization of coefficient alpha." *Psychometrika* 42: 549-565.

RASCH, G. (1966) "An item analysis which takes individual differences into account." *British J. of Mathematical and Statistical Psychology* 19: 49-57.

ROSENTHAL, R. (1966) *Experimenter Effects in Behavioral Research.* Englewood Cliffs, NJ: Prentice-Hall.

———(1964) "The effect of the experimenter on the results of psychological research," pp. 79-114 in B. A. Maher (ed.) *Progress in experimental personality research,* Volume I. New York: Academic.

——— and L. JACOBSON (1968) *Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development.* New York: Holt, Rinehart, and Winston.

SAS Institute, Inc. (1979) *SAS User's Guide.* Raleigh, NC: Author.

SCHUMACHER, C. F. et al. (1979) "Applying the Rasch model to equate examinations in the field of medicine." Symposium presented at the annual meeting of the American Educational Research Association, San Francisco, April.

STANLEY, J. C. (1961) "Analysis of unreplicated 3-way classifications with applications to rater bias and trait independence." *Psychometrika* 26: 205-219.

TAYLOR, D. D., J. C. REID, D. A. SENHAUSER, and J. A. SHIVELY (1971) "Use of minimum pass levels on pathology examinations." *J. of Medical Education* 46: 876-881.

VAN WART, A. D. (1974) "A problem-solving oral examination for family medicine." J. Medical Education 49: 673-680.

WARE, J. E. and R. G. WILLIAMS (1975) "The doctor fox effect: a study of lecturer effectiveness and ratings of instruction. *J. of Medical Education* 50: 149-156.

WINER, B. J. (1971) *Statistical Principles in Experimental Design.* New York: McGraw-Hill.

WRIGHT, B. D. (1977) "Solving measurement problems with the Rasch model." *J. of Educ. Measurement,* 14: 97-116.

——— (1968) "Sample-free test calibration and person measurement," in Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.

——— and G. A. DOUGLAS (1977) "Best procedures for sample-free item analysis." *Applied Psychological Measurement* 1: 281-295.

WRIGHT, B. D. and N. PANACHAPAKESAN (1969) "A procedure for sample-free item analysis." *Educ. and Psych. Measurement* 29: 23-48.

WRIGHT, B. D. and M. H. STONE (1979) Best Test Design: Rasch Measurement. Chicago: MESA Press.

WRIGHTSMAN, L. S. (1972) *Social Psychology in the Seventies.* Monterey, CA: Brooks/Cole.