Invited Reviews

The Assessment of Clinical Skills/Competence/Performance

GEORGE E. MILLER, M.D.

It was just 20 year ago, at the 8th annual RIME conference, that I last delivered an invited lecture, offering what was then labeled "A Perspective on Research in Medical Education". Now, after more than a decade of absence from the front lines of that craft, the invitation to make this presentation was a high compliment but one that generated no small measure of uneasiness, for there seem to be so many others better qualified through personal experience to offer the scholarly review that you have come to hear: David Swanson, or Geoff Norman, or Paula Stillman, or Howard Barrows, for example. However, the organizers have made their choice, for reasons that may be obscure but probably relate to the fact that one who has finally achieved the biblical span of years can once more offer a perspective. At least that is what I will attempt to do.

Although it was suggested that the presentation focus upon standardized patients, it seems important to start with the forthright acknowledgment that no single assessment method can provide all the data required for judgment of anything so complex as the delivery of professional services by a successful physician. And so let me begin by suggesting a framework within which that assessment might occur.

At the base of the pyramid I will use for illustrative purposes (Figure 1) is some assurance that a student, a resident, a physician *knows* what is required in order to carry out those professional functions effectively. There are many who appear to believe that this *knowledge* base is all that needs to be measured. And it is unquestionably measurement of knowledge, largely through objective test methods, that dominates current institutional and specialty Board examination systems. But as Alfred North Whitehead pointed out many years ago, there is nothing more useless than a merely well informed man. Tests of knowledge are surely important, but they are also incomplete tools in this appraisal if we really believe there is more to the practice of medicine than knowing.

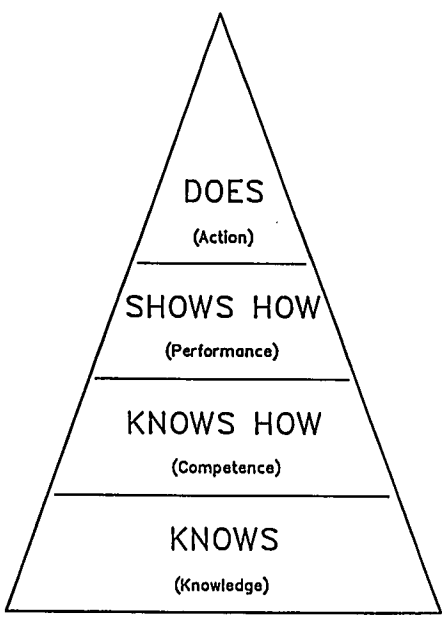

Figure 1. **Framework for clinical assessment.**

To fulfill that broader objective, graduates must also *know how* to use the knowledge they have accumulated, for otherwise they may be little more than "idiot savants." They must develop, among other things, the skill of acquiring information from a variety of human and laboratory sources, to analyze and interpret these data, and finally to translate such findings into a rational diagnostic or management plan. It is this quality of being functionally adequate, or of having sufficient knowledge, judgment, skill, or strength for a particular duty that Webster defines as *competence.*

Despite the significant advances in testing procedures that probe these qualities, skeptics continue to point out that such academic examinations fail to document what students will do when faced with a patient, i.e., to demonstrate not only that they *know* and *know how* but can also *show how* they do it. The evaluation of this *performance* objective represents a challenge now being addressed most aggressively, even though many clinical teachers still claim that they make just such judgments about student performance through encounters on the wards or in ambulatory settings. Such a claim regrettably ignores a growing body of evidence suggesting that these judgments are generally based upon limited direct observation and equally limited sampling of clinical problems (which means an inadequate database); they seem more often related to the product of student interaction with patients, that is, to the accuracy of diagnosis and the nature of management, than to the process through which these conclusions were reached.

Finally, however, the question remains whether what is done in the artificial examination setting ordinarily used to assess any of these elements can accurately predict what a graduate *does* when functioning independently in a clinical practice. This *action* component of professional behavior is clearly the most difficult to measure accurately and reliably. While the diligent efforts of recent years to perfect this final stage of the assessment system have produced mixed results, they must continue with unabated vigor.

In the meantime, while it may be reasonable to assume that either action or performance implies achievement of the more basic elements of the triangle, measurement of the infrastructure (i.e., knowledge and competence) cannot be assumed to predict fully and with confidence the achievement of the more complex goals. When this fact is coupled with the inescapable truth that examinations drive the educational system, because they convey in the most clear and realistic terms what students must learn or do in order to succeed, then it follows that faculties should seek both instructional methods and evaluation procedures that fall in the upper reaches of this triangle.

With this multidimensional complex in mind, let us turn to what we know about the individual assessment techniques em-

ployed in documenting professional behavior, whether it be of student, resident, or practitioner. First, the evaluation of knowledge, particularly by objective test methods, has been so thoroughly studied and the findings so widely disseminated that no more than a summary statement is required here. Suffice it to say that these procedures, skillfully employed, have such a high level of reliability and sampling validity that virtually universal adoption attests to their usefulness, limited in scope though they may be.

It is at the next level, that of assessing the intellectual skill with which knowledge is applied, or the technical skill with which diagnostic and therapeutic procedures are carried out, that some measure of uncertainty begins to intrude. Adoption of the Bloom Taxonomy of Educational Objectives as a guide to the preparation of multiple-choice test questions has surely facilitated the refinement of techniques to probe something more than the recall of informational fragments. However, there remains some disagreement about whether an item that purports to assess analysis, interpretation, or synthesis, for example, can be used confidently to document achievement of such objectives without some knowledge of whether an examinee has previously experienced a comparable challenge, for if such exposure has occurred then what might for the novice require some higher level process may demand no more than simple recall for one well informed or experienced.

To combat this kind of objection, the sequential format illustrated by modified essay questions (MEQ) or patient management problems (PMP) has often been employed. Each of these is introduced by a clinical vignette to set the stage for subsequent actions. In the former those actions are either affirmed, or a revised database provided, before the next step in solving the problem is taken. In the latter no such feedback is provided and subsequent steps depend upon the effect of initial interventions. Scoring of the relatively standardized MEQ has generally shown a reasonably high level of reliability, while that of the PMP has been fraught with problems. Among these are the difficulty of gaining consensus among independent judges on the positive or negative weights to be assigned each possible intervention and even to the optimal path that should be followed. When well prepared, with clear and unambiguous scoring keys and well trained scorers, comparable numbers of MEQs and PMPs in an examination should be about equally reliable. However, with the PMP there is the further confounding element of cueing that is virtually unavoidable in the printed form.

There have been notable efforts to resolve some of the logistical and psychometric problems of these techniques, and to extend their range of usefulness, through the application of more advanced technology. Most prominent among these developments has been the computer-based testing program of the National Board of Medical Examiners, which incorporates both clinical simulations and multiple-choice questions. One new dimension that the computerized simulations may offer is an opportunity to introduce the dynamic element of time in examinee analysis and management of clinical problems. The library of these test materials is steadily expanding, and the procedure is currently under critical scrutiny in more than 70 medical schools.

Despite lingering psychometric questions, to the extent that these procedures have a higher face validity, that is, more closely resemble real performance and action requirements than the simpler techniques, a limited sacrifice of reliability may in some instances be acceptable. What those instances are, however, will remain a matter of intense debate while efforts to achieve higher levels of reliability for these efficient, machine scorable test formats continue.

Less debatable may be the role that models can play in the appraisal of technical competence to carry out specific procedures. Although these devices have been used more often for instruction than for evaluation, as assessment tools they have the advantage of stability and consistency in the challenge with which students are faced. The common ones include such devices as Resusci-Anni, genital and rectal and breast models, others that allow examination of eye grounds or ear drums, simple heart sound simulators, or the more complex cardiovascular system simulator called Harvey. Whatever shortcomings these tools may have lie not so much in the accuracy of what they are designed to represent as in the reliability of the checklists and rating scales required for scoring and of the raters who use them. Such impediments can be significantly reduced, although not entirely overcome, by careful design of the scoring instruments and training those who will use them.

Yet each of these methods is at least one step removed from an encounter with a human subject. It is for this reason above all others that many faculty members cling to the evaluation they feel comfortable making in the course of working with students or residents in the wards, clinics, and private pavilions where so much clinical teaching occurs. And there is undeniable appeal to the argument that this is closer to the reality of independent practice than any of the devices that probe components of that performance in artificial and isolated settings. What is not so generally acknowledged by proponents of this evaluation procedure is the lack of standardization, the limitation of sampling, and the infrequency with which observation of performance itself (rather than discussion of outcome) provides the basic data upon which judgments are made. It is essentially a method that depends on clinical impressions rather than systematic accumulation of reliable information. Direct observation of a candidate performing a history and physical examination, by a trained rater, using a standardized checklist or rating scale, does address the reliability issue but it does not deal with the sampling question, which is critical if generalized conclusions about performance are to be reached. Applied occasionally it may have great usefulness in formative evaluation, but it has distinct limitations for summative assessment.

This brings us to the next step up the pyramidal structure, the use of patient substitutes that allow some of the perplexing psychometric questions associated with real clinical encounters to be answered. Among the first efforts to move in this direction was the introduction of role playing by the American Board of Orthopedic Surgery, later by the Royal Canadian College of General Practitioners, and more recently by the American Board of Emergency Medicine. Here physician-examiners are programmed to portray the historical features of specific patient problems and to convey, upon request, precise information about physical and laboratory findings. Coming out of role, they may then conduct further oral examination of candidates and subsequently make judgments about overall performance using predetermined and standardized criteria. While the evidence is persuasive that these techniques provide insights that cannot be obtained through more conventional methods, it is also clear that large-scale examinations of this kind are costly both in money and manpower.

For specific technical procedures, an alternative approach has been the employment of non-physician gynecologic and urologic teaching associates upon whom genital and rectal examinations may be performed and who can offer immediate feedback on the accuracy of those manipulative techniques as well as an examinee's sensitivity to patient comfort and understanding. While employed most frequently for instructional purposes, these individuals have also been successfully trained to use checklists or

rating scales in judging and recording the quality of candidate performance.

But the most effective substitute for reality is probably the simulated clinical encounter using standardized patients (SP). When Howard Barrows introduced such normal, trained simulators more than two decades ago, there was widespread skepticism about their ability to portray abnormal clinical states accurately and convincingly. I was among the skeptics, but it took no more than a few minutes in my first such encounter to erase any doubts about the reality of the portrayal. By now most of you have probably had a similar experience, and with similar reactions. It has certainly been affirmed by large numbers of students, residents, and practitioners who, in retrospect, have usually been unable to distinguish the real from the simulated patients they met during a series of encounters in an examination setting, a clinic, or a private office.

It is now clear that there are few limits to who can be trained as patient simulators, at least for the portion of a simulation that deals with communication of medical-history facts, emotional states, ethnic and cultural differences, or patient types. The simulation can occur in direct confrontation, in exchanges by telephone, or through third persons who might be required when dealing with infants and children, unrepsonsive patients, or families.

Even an astonishing array of physical abnormalities can be successfully simulated by the most gifted standardized patients: altered reflexes, tics, abnormal gaits, hot and painful joints, and limited thoracic expansion, for example. But for those things that cannot be simulated, many investigators have employed real patients with stable physical abnormalities, trained to deliver a standardized history consistent with those findings.

But just as the encounter with a single patient cannot be used to draw generalized conclusions about overall clinical performance, neither can a single encounter with a standardized patient serve this purpose. The issue of appropriate sampling must still be dealt with. Some ten years ago, Ronald Harden at the University of Dundee, Scotland, introduced the Objective Structured Clinical Examination (OSCE) as a means of increasing the sample of clinical behaviors that might be evaluated in a reasonable period of time, using facilities and resources generally available in most medical schools.

Harden used as a model the familiar multi-station laboratory examinations so long employed by anatomists and pathologists. In this clinical version the stations might, for example, include patients on whom a focused history or physical examination would be performed (with judgments made by one or more observers); x-rays or microscopic slides or electrocardiograms to be interpreted (and reported in some written document); clinical data analyzed, and diagnostic or management conclusions drawn (and evaluated through responses to written questions). As that multi-station format has been further exploited by many other groups, real patients have often been replaced by standardized patients to assure consistency of challenge to examinees. All of which means that the OSCE is not an examination technique per se but represents a format within which a variety of techniques (from multiple-choice questions to simulations) can be employed.

The growing pressure for medical educators to be as concerned about the documentation of clinical performance as traditionally they have been about the acquisition of knowledge has led an ever-increasing number of medical schools to adopt standardized patients or patient substitute methods in their instructional programs. The 1988 LCME questionnaire revealed that 97 U.S. schools now use gynecologic or urologic teaching associates, and 61 use standardized patients for other clinical skill instruction. Although not documented in that survey, it seems reasonable to infer from other sources that a majority of such use is in the Introduction to Clinical Medicine course. But 41 schools also employ such methods for the evaluation of clinical skills, and more than half of that group use them in making decisions about promotion or graduation. In all categories increasing use is projected for the coming year.

While psychometric issues may be a minor concern when these procedures are used for instruction, and create only limited uneasiness when they are employed in formative assessment, they are of major importance when standardized patients are introduced as summative assessment tools. Such questions will further intensify as these simulations are employed in high-stake examinations where certification or licensure are at risk. What, then, can be said about these issues at this relatively early stage of development? Here I will depend largely upon the superb critical review, now in press, by Karl van der Vleuten and David Swanson.

When any evaluation technique is introduced, one of the first questions asked is about the reliability of measurement. In this multi-station format it is apparent that the reproducibility of scores derived from standardized patients may be affected by lack of inter-rater agreement, inconsistency of standardized patient performance, or variation of examinee performance across stations. Each of these variables has to greater or lesser degree been investigated but the conclusions that have been reached must still be regarded as tentative pending further confirmation.

Initially there was a general feeling that the observers who would make judgments about the quality of examinee performance must be physicians, and in order to assure fairness as well as consistency two observers were commonly employed. This manpower-intensive approach raised serious questions about feasibility if the method were to be widely used. It now seems clear that interrater agreement, when raters have been trained in the use of standardized checklists or rating scales, is in the 0.5 to 0.9 range, generally falling between 0.75 and 0.85. Under these circumstances one rater is as good as two, given the usual length of such an examination. Any second rater is probably more wisely employed to increase the number of encounters.

Further it has been found that standardized patients themselves, or other non-physician personnel, when properly trained in the use of well designed checklists or rating scales, can describe examinee performance as accurately as physicians do. Whether medical faculty members at large will accept and act upon this finding remains to be seen.

There is now growing evidence that reproducible performance of the same role can be achieved by several standardized patients trained at a single site. Initial evidence suggests that such consistency can also be accomplished when training occurs at several sites or by different trainers. While this may be of little concern for individual institutions, it assumes great significance when cooperative, multi-institutional testing is contemplated, a development that will be of critical importance if economies of scale are to be realized.

On one matter there need be no further debate: examinee performance on a single case is a poor predictor of performance on others. The issue of content specificity looms as large here as it does in other examination methods. It now appears that to obtain acceptably reproducible scores a minimum of three to four hours of testing time will be required. Where SP-based stations are either associated with or followed by questions involving data interpretation, differential diagnosis, or lab skills, for example, an even longer total test will be needed. This suggests that SP testing might best be used for the documentation of direct patient interaction behavior, while other aspects of

clinical performance are assessed with more economical testing methods. Whether this compartmentalization of performance components distorts the overall assessment of professional behavior will require further investigation.

There has been considerable discussion about optimal station format and length. It seems reasonable to conclude from the evidence now available that these matters should be determined by what is to be measured rather than by any arbitrary decision in advance. The longer station may give more information, but shorter stations will provide wider sampling of patient problems in the same time period.

Finally, it should be noted that most reliability studies have focused upon the reproducibility of scores rather than of decisions. There has not yet been any significant amount of work on setting absolute standards for SP-based tests, yet a strong argument could be mounted that ranking examinees is not the goal of clinical performance assessment. The real objective is to determine whether a defined level of mastery has been achieved. Were such a pass-fail point to be the focus of reliability studies, one might predict that less testing time would be required to reach supportable generalized conclusions. Such a shift in focus might also offer the opportunity to explore the usefulness of sequential testing, for when most examinees perform well (as they could be expected to do in this situation) then short screening tests might reliably certify the majority and detailed attention could then be reserved for those whose performance is of questionable quality.

Equally important are questions of validity. Here it may be possible to speak with confidence on the subjective assessment of this quality, but with less confidence on its empirical determination. Certainly standardized patients must have a high level of face validity (which Geoff Norman refers to as "faith validity") when residents and practitioners who meet them in the course of a series of clinical encounters are unable to detect which subjects are real and which represent simulations. And they also appear to have content validity, since the examinee performance being probed is that required in the practice of medicine. Whether the sampling of those behaviors is sufficiently large or diverse depends upon the care with which a blueprint has been devised and the extent to which the test matches that blueprint. But that is true of any test.

Empirical validation studies have thus far been relatively rare. Those which have been carried out appear to confirm that individuals with more advanced training perform better than beginners, and one might conclude that such findings confirm construct validity. Similarly with efforts to document concurrent validity: low correlations with more conventional tests are often offered as evidence that different qualities are being measured and higher correlations with faculty ratings of clinical performance as evidence that both are measuring the same critical quality. But in each instance two other issues intrude. First is the now generally accepted fact that performance is embedded in knowledge that can be expected to increase and thus influence performance as the stage of education advances. Second, correlation studies are usually derived from the scores or rankings of norm-referenced tests rather than the specific behavioral achievements of mastery-referenced appraisals.

When some of those mastery elements are specifically addressed, then the special contribution of standardized patients to the testing armamentarium becomes more apparent. For example, false positive findings on physical examination (such as heart murmur, papilledema, or joint effusion), or reporting findings when the appropriate examination has not been performed, may be infrequent numerically but represent significant deviations from acceptable standards, deviations that would otherwise go undetected. It is just such deficiencies that have too often been uncovered by these techniques, in students already judged qualified by faculty tutors at the end of clinical rotations.

A persistent question about SP-based tests is one of feasibility. It is an issue that cannot be evaded, but one for which only preliminary conclusions can be drawn since no common method for documenting costs has yet been agreed upon. The variables include training and utilization costs for whatever number of SPs are required to provide the necessary sampling of performance, the time and dollar cost of developing cases and scripts and checklists and rating scales, the cost of materials and supplies needed for the test, the cost of consolidating scores and reporting the results, and whether physicians or non-physicians (i.e., standardized patients themselves or others) are used in judging performance. Omitting developmental costs, current estimates for implementing a full-scale certifying examination range between $100 and $200 per student.

Such estimates, however, do not include the potential economies of scale that might be realized through cooperative test development by several schools or testing organizations. Efforts of this kind have been initiated at both Southern Illinois University and the University of Massachusetts and will be further examined in collaborative studies being encouraged and supported by the National Board of Medical Examiners. These undertakings are probably justified economically only when the objective is to create a summative examination of clinical performance, although creation of a pool of qualified SPs with accompanying scripts and checklists or rating scales might ultimately prove to be a welcome resource for instruction and formative assessment as well.

As promising and appealing as the SP examination method may be, any confident universal application of the technique to high-risk promotion and certifying procedures must probably await further research on some of the key questions that remain to be answered. Since that is the kind of work that so many in this audience might undertake, let me list some of the investigations that seem especially needed.

Among the most difficult problems is that of reaching agreement on what components of professional behavior should be addressed by an SP-based examination. From the variety of test formats now in use, it seems clear that different groups have different things in mind, and those differences may have significant influence on the time required to gather sufficient data for generalized conclusions to be drawn. In the light of cost-benefit concerns, should the SP component of a qualifying clinical examination be limited to assessing information gathering and communication skills (as several prominent investigators have suggested), or is some significant element lost by assigning the documentation of other aspects of professional behavior to more traditional testing methods?

Equally perplexing is the question of optimal methods for scoring an encounter with standardized patients. It is not simply a matter of checklists or rating scales, scoring by physicians or trained non-physicians, but rather of reaching agreement on what aspects of the encounter to observe and how to combine and weight these observations to yield scores that reflect, in a meaningful manner, the adequacy of observed performance. After reviewing many of the scoring forms currently in use, van der Vleuten and Swanson were moved to comment that "the potential for omitting important items and including unimportant ones is great. The former penalizes examinees who take indicated actions that are not listed; the latter rewards examinees who are unjustifiably thorough."

Whatever the behavioral dimensions of the test, or of the

scoring procedure employed, a still unresolved question concerns the most effective methods for developing performance standards. This issue has been successfully bypassed in the past through the practice of norm-referenced testing; but in judging clinical performance it seems imperative to adopt a criterion-referenced method. It has been difficult in other examinations to gain agreement on criteria, and there is no reason to suppose it will be any less so with SP-based procedures. Nonetheless, if we are to be faithful to the charge placed upon us by society to certify adequacy of clinical performance, not merely the rank among performers, then we can no longer evade the responsibility for finding a method that will allow us to do so.

If cooperative inter-institutional efforts are to be mounted successfully, there remains at least one additional issue to be addressed: the techniques and logistics for creating a shared pool of standardized patients. Some initial work on these questions has been carried out at both the University of Massachusetts and Southern Illinois University working with other regional institutions. The former developed a cadre of standardized patients that were transported to the other sites for testing sessions; the latter developed a set of cases and standard training procedure for simulators that were shared with another medical school so that they could give a common examination. Each of these procedures appeared to work well for the limited objective of the experiment. But if there is to be broader sharing, it is essential to find convincing answers to several questions that remain.

For example, there must be more persuasive evidence than now exists that the portrayal of a given case by several standardized patients trained by different trainers, either at the same site or at different sites, results in comparable SP performance. Further, if a single SP is to be used repeatedly for a single case, is that individual's performance stable over time? Without documentation of comparability and stability, the reliability of this testing procedure will be subject to serious question. An encouraging recent development is Robyn Tamblyn's work, which not only suggests that this goal can be achieved but also offers leads to methods that might further improve comparability.

While perhaps not as pressing as these concerns, yet still important, is that of practice effect. With most other testing methods, examinees who have had the benefit of prior experience are generally able to perform better. It seems reasonable to suppose that the same thing might be true with standardized patients, despite the fact that they are simply intended to be an accurate representation of the reality that students encounter regularly in both hospital and ambulatory settings.

Of particular concern to the National Board of Medical Examiners and the Educational Commission for Foreign Medical Graduates, who are committed to the implementation of SP-based certifying examinations, as well as to other certifying bodies that may embark upon such efforts, is the issue of costs and logistics. It is not clear whether present cost estimates, derived from the always expensive developmental phase of any program, can be used as reasonably accurate projections of what might be required to mount large-scale operations for a national or international constituency. And even if they are, is a cost in that order of magnitude a justifiable expense for certifying or licensing examinations? If there is a way to reduce that expense, whatever it may be, it is worth exploring. The most promising possibility at this point appears to be the sequential examination strategy, using a coarse screen to identify all who are clearly acceptable and reserving the fine screen for those who fall in the gray zone of doubt. Research on this technique is badly needed, for it has major implications in the ultimate implementation of new strategies for testing.

Up to this point I have attempted to be dispassionate, setting forth what appears to be a reasonable representation of our state of knowledge about the assessment of clinical skills/competence/performance. But let me close with a set of personal views, which some might regard as no more than biases.

First is the sense of urgency I feel about getting on with the task. We have for too long been willing to base our judgments about readiness to engage in professional practice by determining whether individuals could demonstrate that they had acquired a body of knowledge that reference groups (most commonly academicians) believed essential for that function. It would be pointless to question the importance of knowledge, despite its transitory character. More important is that we demonstrate decisively through our testing procedures that knowledge alone will not be enough to succeed either in passing the examinations or in performing as a physician. Each contemporary refinement in competence testing has been aimed at drawing closer to that goal, but not until the more recent studies of SP-based examinations have we had something that approached the reality of encounters with patients and their families, in all the ambiguity that reality entails.

Given the experiential and psychometric data now available, it seems not merely desirable but essential to widen the adoption of such methods and to incorporate them, as quickly as answers to remaining problems can be found, in the high-risk examinations that qualify candidates for independent general or special practice.

Lastly, in these assessments it is time to abandon the comfortable camouflage of normative procedures and adopt criterion-referenced testing. Ranking candidates, with arbitrary cut-off points that reflect distinctions far more than differences, is neither good education nor good medicine.

It will not be easy to convince conservative medical faculties, reasonably comfortable with the current conventions that allow clinical impressions to substitute for systematic accumulation of behavioral evidence, that change is in order. Neither will it be possible to convince them with data alone. But without data, passionate arguments are bound to falter for, as one keen observer pointed out many years ago, where data are sparse opinions are plentiful. And that would seem to describe the status of clinical skills/competence/performance assessment in many parts of the globe. I can only hope that the research in medical education community, the change agents who are here today, will in this matter ultimately deserve the words with which Adlai Stevenson described Eleanor Roosevelt: "She would rather light a candle than curse the darkness, and her glow has warmed the world." I wish you well in this worthy enterprise.