What I Wish I Knew About Assessment[1]

Robert M. Guion

A quarter of a century ago I wrote a book on personnel testing. I'm now writing another. I've done two Annual Review chapters, and I've done individual assessments for promotion, transfer, and hiring decisions. I've developed and validated testing programs and exercises for assessment centers. I've had the expert witness experience. I've been invited to speak–here and elsewhere–on assessment topics, presumably because the people who invited me thought I might have something to say. I've thought so, too.

Now I'm not so sure. In retirement, one wants to look back fondly at what one has accomplished. It's unsettling to realize that less has been accomplished in the last quarter of a century than all the activity would suggest.

The last quarter century can be called the EEO era in personnel selection. Shortly before the EEO era began, Marvin Dunnette outlined an idea for individualized assessment and decision making. A little before that, he and I and others wrote articles calling for the prediction of multiple criteria. Such ideas and others of the early 1960s were worth pursuing. We didn't pursue them because we got preoccupied with laws and regulations and court decisions that, in all too many respects, froze the field of personnel assessment at a 1965 level.

I make no apology for that preoccupation. Our society, through its many prejudices, had placed too many obstacles to good jobs in the way of too many well qualified people. Those obstacles *had* to be removed, and professional assessment techniques helped remove the. Nevertheless, I regret that the unfolding of events early in the EEO era fossilized some ideas and procedures that deserved further development. I wish I know how that development would have turned out.

Not all areas of study were so directly preoccupied with EEO issues. The coming of the computer age led scholars to important advances in educational assessment methods and to new areas of research in cognition. I didn't keep up well with the relatively new developments in these other fields, but I think they may have untried relevance for us.

I've settled on a list of twelve things I wish I knew. The list is incomplete, but long enough to stretch your patience. Wishes in my list are grouped under three headings. The first set stems from undeveloped ideas set aside because of the EEO preoccupation. The second set emerges from my ignorance about the work of our intellectual neighbors. The third set–a nontrivial set–has grown out of just plain frustration!

The Underdeveloped, Frozen Ideas

1. In 1961, I suggested a five-step validation sequence that called for the identification and valid measurement of multiple criteria, and the validation of predictors independently against *each* of them (Guion, 1961). I argued that the relative importance of different criteria was not important at the time of validation; it becomes important at the time that administrative decisions have to be made. The relative importance of predicted quality and quantity of performance, for example, may depend both on the relative validities of those predictions and the current state of production in the organization. Marv Dunnette put it well when he said, "Let us cease searching for single or composite measures of job success and proceed to

[1]Presented at the IPMA Assessment Council annual conference, Chicago, June 24, 1991.

undertake research which accepts the word of success dimensionality as it really exists" (Dunnette, 1963b, p. 252). We may both have been guilty at the time of cutting the criterion domain into too many itsy-bitsy pieces, but there *are* different kinds of criteria. Performance quality is surely different from staying vs. quitting.

Consider table 1. Candidate A is an obviously good choice; every prediction is consistently favorable. This is a candidate who is likely to perform well, be trustworthy, and stay long enough to repay the cost of hiring. Candidate B is likely to be trustworthy, and may stick around, but is not likely to boost average performance. Candidate C is likely to perform very well indeed, but with questions about both staying power and trustworthiness. What decision will you make about B or C?

	Candidate		
Probability of:	A	B	C
Above average performance	.80	.20	.70
Quitting in a year	.10	.40	.80
Being fully trustworthy	.90	.85	.35

Table 1.

Wish #1: I wish I knew how to handle predictions of different criteria when the predictions are inconsistent. It shouldn't be too hard to work on. Standard policy-capturing might be used to develop an equation, or perhaps a set of equations, to model the judgments of organizational decision makers; lens model research could be used to validate those judgments. But against what? The lens model uses a single criterion!

Neither Guion nor Dunnette nor anyone else developed the idea in the early years of Title VII. We faced a big enough problem trying to explain to lawyers, enforcement agencies, and confused employers how to validate with even one criterion. If we had been more successful with that, we might have worked on the case of inconsistent multiple criteria. But we didn't, and I wish I knew what we would have learned had we done so.

2. Dunnette also presented a model for individualized selection (Dunnette, 1963a). The model recognized two unfortunate facts, too often ignored. One is the one-model-fits-all assumption in traditional validation that everyone is just like everyone else on everything except the predictors and the criteria being studied–and that therefore the same predictors and regression equations should be used for all varieties of candidates. The extreme alternative is an assumption of complete uniqueness; the unfortunate fact is that we have no research model for validating with an *n* of 1. Dunnette's very intelligent compromise was that we should establish homogeneous groups of predictors, of people, of work situations, of job behaviors, and of organizationally-relevant consequences of those behaviors in those situations. That is, he proposed a massive tree of moderators such that predictors used for one group of people,

to predict one kind of behaviors, might prove different from those used for other categories of people and behavior. To make a quick if obvious case, should we use the same kinds of assessments for applicants whose backgrounds suggest that they could *learn* to do the job that we use for those with real experience doing it? How about aptitude tests for the former and work samples for the latter?

With gross over-simplification, figure 1 gives the flavor of the Dunnette model. Suppose we have a policy against hiring people who ignore no smoking signs, are dirty and unkempt, spit on the floor, or talk abusively to others in the waiting room. If a candidate doesn't fit that policy, he or she is quickly rejected. Otherwise, an interview is conducted. In this example, the interview seeks only two kinds of information: What is the candidate's general history of prior accomplishment–that is, of getting this done and done well–and is the candidate experienced in doing the job at hand? If the candidate has many impressive accomplishments in work, school, or community life, but has no experience in this specific job, we might hire the person anyway, for this or some other position; we would not let a real gem get away. If such a candidate *does* have relevant experience, or if another person has a modest record of accomplishment plus real job experience, we might administer a work sample test. A high score is a sure further reason for hiring the person of many accomplishments, but we might insist on a *really* high score for the one with more modest prior achievements. Poor work sample performance would tilt the decision toward rejection. With someone with no record of achievement, even a lot of relevant job experience might not outweigh the poverty of the record; such a candidate might be rejected. People with little or no achievement history and no relevant job experience could be testing to assess both aptitude and work motivation. If their scores are good, we hire them; with poor scores, we reject them.

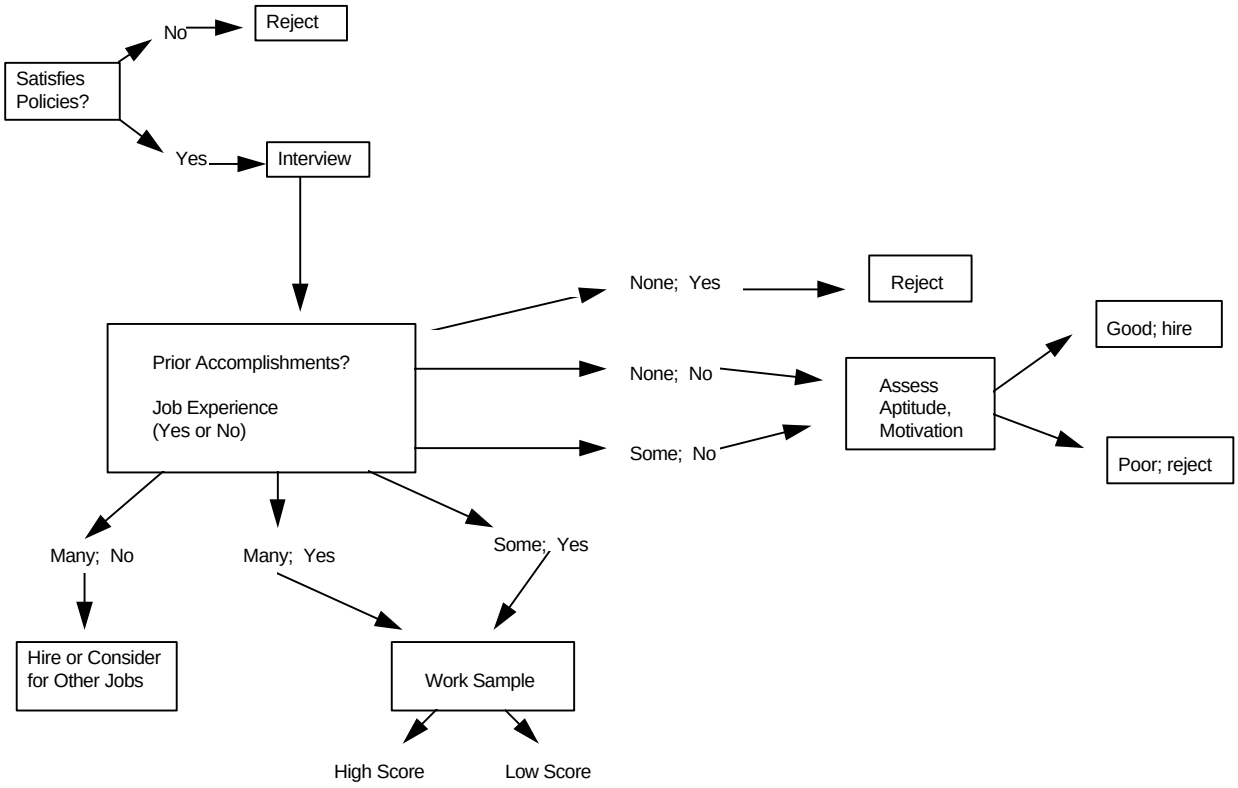

Figure 1

We were never very happy with treating everyone alike. After all, the whole concept of selection is based on the fact of individual differences. Dunnette's model might have been taken seriously had we not been sidetracked in treating demographic variables–protected classes–as potential moderators. They weren't, so we said there are no moderators.

Wish #2: I wish I knew how to individualize assessments in making employment decisions, in a way something like Dunnette's model. I suspect that there are psychological, nondemographic moderators–perhaps in attitudes, background experiences, or temperament–that would make the model work if the search for them were given some serious thought. I wish I knew a good statistical model for evaluating the components of the process. Maybe some sort of path analysis or other structural model might work, but my impression is that everyone in such an analysis must follow a common path.

I think the model requires a lot of background research. The taxonomic requirements alone will take years to develop. When the various categories are satisfactorily in place, many bivariate studies will be needed, followed by meta-analyses, to form the required linkages.

3. According to an old axiom in personnel work, performance is a function of ability and motivation. Maybe motivation is provided by good management, but many of us have believed, and I still believe, that the interests and personality traits people bring with them in applying for a job can to some degree add something to the prediction of performance. It has not been very widely demonstrated, however.

Wish #3: I wish I knew what personality variables to assess for what kinds of performance problems, how they might best be assessed, and what their assessment would add to the accuracy of predictions of performance or other job-related behavior.

Dick Gottier and I reviewed research on personality test validity in 1965. We did not, as so many have assumed, conclude that personality tests are not good; indeed, we found evidence that "in *some* personality measures can offer helpful predictions" (Guion & Gottier, 1965, p. 159, emphasis in original). What we *did* conclude was that there was no *generalizable* evidence of the value of personality tests in employee selection. Personality testing had virtually stopped until quite recently, so there is still not much generalizable evidence that they work well.

4. In the 1950s, the Air Force did many studies of air crews and how they might best be formed. Group processes and structure were extensively studied in experimental social psychology and some private employment settings. Some of the relevant research centered on the initial formation of teams, some of it on selecting new members for existing teams.

The decline of such research cannot be properly blamed on the EEO era, but EEO concerns were surely relevant. Consider a 1952 study by Van Zelst (1952). One group of carpenters formed work teams making their own choices; other teams were formed by assignment. Results were better for groups formed by sociometric choices. But what would have happened in the post 1965 ear, with a mixture of ethnic groups and some women in the pool of carpenters? How integrated would such self-formed groups be? We don't really know, but we can be forgiven for thinking that such groups would be homogeneous in more than ability.

Another practice common in the 1950s was individual assessment of candidates for managerial positions. An assessor typically visited the work site to find out something about it

and to decide what kind of people would "fit in" or "be a team player." Now we are a bit queasy about this notion of "fitting in," and we don't often think about the matter.

Wish #4: I wish I knew how to select teams, or replacement members of existing teams.I wish I knew how to talk about hiring people who will "fit in" without worrying that they phrase itself may mask a desire to discriminate on grounds unrelated to performance.

The Ideas from the Neighbors

5. Consider the phrase, standardized testing. It has an egalitarian ring to it. Everyone is assessed in a standard way, with no hint of favoritism for anyone. Everyone was given the same items, with the same time limits, the same size type, the same response options, the same scoring key, and the same norm tables or expectancy charts for interpreting the scores.

Now we have item response theory and computerized adaptive testing. Different people get different items and even different numbers of items, and they aren't even parallel items. A score is described, not with a percentile rank or a probability of success but with an estimate of ability called theta.

Wish #5: I wish I knew what a "standardized test" is. I'm prepared to say that the testing procedure, and certainly the score interpretation, is far more standardized with IRT than in more traditional testing because it is the logic of the IRT algorithm of measurement that is standard. I am not prepared, however, to say that measurement by some other standard algorithm (for example, a scheme for letting people choose the items they answer) would have the same egalitarian flavor. So, at a very basic level, I have to admit that there is something about the idea of test standardization that has eluded me–and continues to elude me. I have the uncomfortable feeling that its importance is more substantive than semantic.

6. I distinguish between the use of the work *measurement* and the word *assessment*. Assessment is the more global term, including measurement and a whole lot more. I refer to measurement when the procedure has enough unidimensionality that people can sensibly be ordered along a single scale. I refer to assessment when the procedure assesses a kind of general fitness for some purpose, when a position on a fitness scale can be reached by a variety of combinations of component traits.

Over the last two or three decades, psychometric research has enabled us to do far better measurement than we typically do. We can use IRT, generalizability theory, confirmatory and simultaneous factor analyses, and follow the LISRELites to their promised land.

Wish #6: I wish I knew whether excellence in measurement is really much better than excellence in non-metric assessment. In part I'm wondering whether very precise measurement of one or two well-defined traits is better than carefully developed methods of global assessment. Stated differently: I wish I knew whether maximizing measurement validity is a better strategy for personnel decisions than maximizing job relevance. If you are among those who still equate validity and job-relatedness, consider the use of an examination with a very high level of construct validity as a measure of inductive reasoning for a hamburger-flipping job in a fast food palace. Validity in this sense is no assurance of job relevance. My concern is that, in seeking only psychometric elegance, we may be failing to study and to understand real job requirements. Sophisticated psychometric methods do indeed help us do a better job of measuring, but do they help us do a better job of selecting employees? I wish I knew for sure.

7. Let's back up to the EEO problems and moderators. It has been established that demographic variables are unlikely moderators of employment test validities. Nevertheless, I'm struck by what seems increasingly to be an approaching bimodality in minority performance on both sides of the prediction equation. In the last quarter-century, many minority people have moved into middle class living with all it implies–better housing, education, jobs, and general experiences–and greater comfort with middle class American culture. Many others seem permanently enmeshed in poverty. Does a person's economic status tell us anything useful about the validity of that person's test score? That is, is the score a better description of ability for people with "advantages" and weaker for those called "disadvantaged"? Yes. Although various environmental scales failed to serve well as moderators, we know that test scores tend to be somewhat related to socioeconomic status. But why?

Studies of differential item functioning have considered one possible aspect of the answer, but they are usually based on the demographic categories. Maybe we can do better. Cross cultural research has grown in the last couple of decades. More serious attention has been given to the equivalence of measurement instruments in different languages and for people of different cultures. There is reason to question whether the meaning of taking a test is constant across cultures.

Wish #7: I wish I knew something about test-taking strategies I would like to know if people with a middle class background, people who have gone through reasonably good schools and have taken it seriously, approach a test differently from those from a poverty culture whose educational back ground is weak or abbreviated. Are they more likely to guess and move on or to leave an item blank for later consideration? Are they more or less likely to change responses? Do they spend more time on individual items? Do they tend to rule out certain distractors before guessing among the others? Is it not at least conceivable that test scores on the average are better for people who follow certain strategies than for people who use other strategies? And is it not at least conceivable that, in some ethnic or cultural or economic groups, nonoptimal test strategies are characteristic? I've heard people talk about test-taking strategies for years, but I've seen little data. Traditional studies of strategy have had to rely on retrospective comments. I believe that computer technology can help us do much better studies of the actual test-taking process. The results might be very important in a variety of ways.

8. For jobs where lots of people are hired, traditional methods of validation work quite well. Increasingly, however, people work in small organizations with few on any given job. Positions are often unique in important ways. Do we have anything to offer employers in such settings? I think so, and we all have some suggestions. One with little exploration, however, is a systematic program in which standard procedures are used to identify (a) position responsibilities, (b) traits to be assessed for designated responsibilities, and (c) prescribed ways to asses those traits. In such a system, one would evaluate, not inferences from test scores, but the program as a whole.

Comprehensive programs are evaluated by quasi-experimental research. Here are two very simple designs. The top one has been used several times in validating integrity tests. In a chain store, for example, the test may given in some experimental group of locations and not in other, somewhat similar control locations. After time, the criterion performance (usually inventory shrinkage) is measured. If there is less shrinkage in the stores where the test is used, the selection system is considered valid.

This really is a rather weak design. Not much can be said about causes. Is the test valid, or is shrinkage reduced only because the test shows corporate concern about employee theft, or were the two sets of locations poorly matched? Design 2 is still weak, but it's a little bit better because it calls for measuring shrinkage before doing any testing in any of the locations. If the control locations show little or no change, and the experimental locations show substantial improvement, problem with Design 1 are reduced.

Wish #8: I wish I knew more about program design and evaluation Cook and Campbell (1979) described many different quasi-experimental designs and some of them are stronger than others; most are stronger than these. Program evaluation methods have been around a long time; their popularity has both waxed and waned without my ever being involved in the debates. I don't know much about their virtues or weaknesses, and I'm not at all sure how to use them for evaluating a selection decision program. Indeed, I'm not sure what to include in an integrated assessment-and-decision program. I'm not alone. For most of us, program evaluation hasn't seemed especially relevant to personnel selection. We'd better take another look.

Frustrations

9. Many of use believe that simulations are better than multiple-choice aptitude tests. When candidates cannot have had direct job experience, realistic simulations may not be feasible. In one situation I have been following, a great deal of research time and money has gone into a project to develop simulated training and subsequent work samples–certainly far more than would have been required for a multiple-choice aptitude test.

I've also been serving on a panel to make recommendations to a federal agency. The rest of the panel is quite convinced that the traditional multiple-choice test is a bad idea and is recommending "new"–and untried–approaches to assessment. The new ideas being suggested are indeed intriguing, and I wish the agency well.

But, *wish #9: I wish I knew whether simulations and related content-oriented assessments work as well as we think they do, and if so, whether they work as well as the more traditional assessments of aptitudes or whether they add incrementally to aptitude assessment.*We do need to work diligently to improve what we can do, to develop new constructs more relevant to work performance, and to develop alternative and preferable improved ways to measure them. But we must avoid the trap of assuming that new is necessarily improved, and I'm getting terribly frustrated by the increasingly common attitude that enthusiasm for the new–whether the once new idea of assessment centers on the currently new notions of individual differences in cognition–are necessarily better without confirming comparative evidence.

10. It has been very frustrating to work with colleagues who consider cognitive abilities the only ones worth assessing. Sensory abilities, motor skills, and physical condition–these are almost universally ignored or, if considered, considered and assessed superficially. Except for the fitness center and the jogging path, muscles are simply no long in style. The image of a job requiring the big he-man with bulging muscles is pretty much limited to work with a jack-hammer. For many years, industrial engineers and miscellaneous inventors worked to find ways to simplify jobs, making it possible for wimps to do the work that formerly required Charles Atlas types. The Civil Rights Act accelerated the trend by making it useful to find work aids to help women do jobs that previously required high levels of strength. The Act for Disabled Americans will call for further innovations in accommodating to a variety of physical problems.

Indeed, it is their variety that concerns me. The physical demands of jobs are not merely muscular. Susceptibility to allergic reactions or to physical stress or even physical susceptibility to effects of emotional stress are physical problems associated with work.

Wish #10: I wish I knew more about these kinds of physical demands of jobs. They are important not merely because of the laws and the potential for litigation under various legal theories; they are important because some effects may cause long term problems for employees–problems more long lasting than a sore muscle and requiring treatment more sophisticated and expensive than liniment.

11. A major frustration for me these days is the almost universal and axiomatic use of cutting scores. I'm not referring to the kind of cutting score marking the lower limit of ranked scores on an eligibility list. I'm referring to the kind of cut score above which anyone who comes can be hired and below which no one will be–the kind that changes a continuous score distribution to a dichotomy. A major part of my frustration is with the reason most often given for setting cut scores: "My managers just can't handle anything more complicated than a pass-or-fail score."

Fervent wish #11: I wish I knew why and when we stopped assuming that decision makers had any brains. If they are as unteachable as many psychologists assume, we certainly won't get very far with some of the other items on my wish list. For example, we couldn't hope to teach them to use a specified policy for combining inconsistent predictions and would have to stay with the same tired old overall criterion.

12. Most of all, I wish I knew how to handle the people who don't seem to realize that assessments will be made, that they will be better if competently made, and that decisions have to be made about the way in which opportunities are to be allocated in society. You will surely recognize the frustrations behind this item in my list. We face daily a society and its government that, on the one hand, insists that tests are inherently invalid, biased, and violations of civil rights. We face a bill in Congress that is fundamentally directed to tests as instruments of employment decision making but has no corresponding concerns about interviews. We get nonsense from adversarial lawyers who seek to solve no problem but to win their kinds of cases. Consider the recent Harvard Law Review paper by Kelman (1991). He argues that what he called general ability tests are necessarily invalid, even statistically, because they do not correlate well with pay, "the best presumptive measure of worker productivity in a market system" (Kelman, 1991, p. 1208). Even for a "perfectly valid" test that is "racially unbiased in the Cleary sense," "it is still not obvious that those with higher test scores are entitled to the jobs that by hypothesis they are better able to perform" (Kelman, 1991, p. 1243, italics in original). We get reports like the recent one from the "National Commission on Testing and Public Policy" that complains of the overuse of multiple choice examinations without a systematic evaluations of alternatives to them. (Kelman, at least, offered an alternative to testing: hire those who apply to see how well they can eventually perform– with, of course, curbs on the right of employers to fire those deemed poor performs.)

Somehow, we need to get the questions asked in the right order. The first question is not whether to test, or how to test, but how society's opportunities and rewards are to be allocated. We have generally said that nepotism is not an acceptable basis, we disparage employers who chose on the basis of applicant appearance, and we pass laws against the allocation of rewards on the basis of sex, race, or ethic identification. We oppose quotas; we promote affirmative action; unfortunately, we don't articulate the difference very well. We aren't totally sure, but we think we don't like random allocation, not even of the first come, first hired variety.

We like to say, at least some of us do, that we think opportunity should be allocated on the basis of merit. *Only if we agree on that do we reach the issues of testing* And *then* we must ask how merit is best identified. One way to assess merit is to test. Another way is to interview. Still another is Kelman's open-ended probationary period.

We can stipulate that, if merit is to be assessed by testing, the testing must be done competently, and we can probably agree on at least some of the requirements of competence. But there are still questions. There are many problems with testing, and those of us who are here probably know those problems better than the critics do. But when we've considered the problems, we still have one remaining question: "What will do the job better?" How can we get this sequence of questions to characterize the deliberations of those concerned with public policy? How can we get such deliberations to be directed toward solving a specifiable problem rather than to trumpet previously held adversarial positions? I wish I knew.

References

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.

Dunnette, M. D. (1063a). A modified model for test validation and selection research. *Journal of Applied Psychology*, *47*, 317-323.

Dunnette, M. D. (1963b). A note on *the* criterion. *Journal of Applied Psychology*, *47*, 251-254.

Guion, R. M. (1061). Criterion measurement and personnel judgments. *Personnel Psychology*, *14*, 141-149.

Guion, R. M., & Gottier, R. F. (1065). Validity of personality measures in personnel selection. *Personnel Psychology*, *18*, 135-164.

Kelman, M. (1991). Concepts of discrimination in "general ability: job testing. *Harvard Law Review*, *104*, 1158-1247.

National Commission on Testing and Public Policy. (1990). *From Gatekeeper to gateway: Transforming testing in America* Chestnut Hill, MA: Author, Boston College.

Van Zelst, R. H. (1952). Validation of a sociometric regrouping procedure. *Journal of Abnormal and Social Psychology*, *47*, 299-301.