

378.794
G43455
S-88-1




GIANNINI·FOUNDATION

OF AGRICULTURAL
ECONOMICS



UNIVERSITY OF
CALIFORNIA



Agricultural Employment Testing: Opportunities for Increased Worker Performance

WAITE MEMORIAL BOOK COLLECTION
DEPARTMENT OF AGRICULTURAL AND APPLIED ECONOMICS
232 CLASSROOM OFFICE BLDG.
1094 BUFORD AVENUE, UNIVERSITY OF MINNESOTA
ST. PAUL, MINNESOTA 55108

Gregory Encina Billikopf



Giannini Foundation Special Report No. 88-1

Division of Agriculture and Natural Resources

PRINTED NOVEMBER 1988

378.794
C43455
S-88-1

Agricultural Employment Testing: Opportunities for Increased Worker Performance

WAITE MEMORIAL BOOK COLLECTION
DEPARTMENT OF AGRICULTURE
232 DEPARTMENT OF APPLIED ECONOMICS
1994 BUFORD AVENUE
ST. PAUL, MINNESOTA 55108

The author is:
Gregory Encina Billikopf
Farm Advisor, Personnel Management
University of California, Cooperative Extension

ACKNOWLEDGMENTS

I wish to express appreciation to Dr. James A. Wakefield, Jr., Psychology Department, California State University, Stanislaus, for suggestions and comments on my master's thesis on the topic of agricultural employment testing, and for guidance in the study of psychological testing; to the farmers, foremen, farm workers, and secretarial applicants who participated in the study; and to Julie Reinertson for review of this manuscript. In addition, the author would like to thank the reviewers of the Giannini Foundation for their helpful comments and suggestions.

TABLE OF CONTENTS

Summary and Highlights	1
Opportunities for Increased Worker Performance	2
Is Agriculture Taking Advantage of Tests?	2
Economic Advantages of Testing	2
Legal Testing Issues	4
Technical Terms Used in Employment Testing Procedures	6
Work-Sample Testing: A Special Case	12
Differences in Worker Performance	12
Agricultural Employment Testing: Case Studies	17
A Content-Oriented Strategy: Agricultural Secretary Selection	17
A Criterion-Oriented Strategy: Tomato Harvest Testing	18
A Criterion-Oriented Strategy: Testing of Vineyard Pruners	19
Concluding Comments	22
Appendix A	23
Appendix B	24
Selected References	25

THE GIANNINI FOUNDATION SPECIAL REPORT

The Giannini Foundation Special Report provides an outlet for items worthy of publication but not fitting into the foundation's regular series. Single copies of this report may be ordered free of charge from Agriculture and Natural Resources Publications, 6701 San Pablo Ave., Oakland, CA 94608. Order by number (Special Report No. 88-1).

Kirby Moulton is serving as Giannini editor. On the editorial board are Michael Caputo, Larry Karp, Keith Knapp, and Quirino Paris. All Giannini publications are peer reviewed. Carole Nuckton is technical editor.

Other publications of the Giannini Foundation and all current publications of foundation members are listed in the *Giannini Reporter* issued annually in the summer.

SUMMARY AND HIGHLIGHTS

[The purposes of this report are to show that workers, even at the lower end of the pay scale, perform at different levels, and that better workers can be selected through the use of properly designed selection systems that include tests. In addition, the report introduces the concept of pre-employment testing as a farm management tool, argues for increased pre-employment testing in agriculture at all levels, and promotes better testing practices by discussing possible legal and managerial pitfalls.]

Terminology applied in the psychological testing field is used for convenience; definitions of the various concepts are given when first introduced.

This report is directed to farmers, many of whose costs are in some way related—directly or indirectly—to the quality of those they hire. Agricultural employees make decisions ranging from choosing which fruit to pick and which to leave on the tree to what crops to plant, what machinery to purchase, and which foremen to hire. The report should also be of interest to personnel managers and to labor management consultants who work closely with farmers. Finally, this report contributes to the field of psychological testing by reporting several applications. While the legal discussion in the report is particularly important to U.S. farmers, the management concepts introduced are applicable beyond the U.S. borders.

The greater the importance of the decision-making outcomes of the person being hired, the greater the effort that should go into the selection process. However, few jobs are so unimportant as to require no effort at all in employee selection. Especially at the

farm worker level, employees are often chosen on a first-come, first-hired basis. Yet there are many potential economic benefits from hiring them instead on the basis of test results. Today's legal climate makes effective employee selection more important, for firing workers is becoming increasingly difficult. Anti-discrimination laws and wrongful discharge litigation have made employers more careful when they hire and fire workers.

The report illustrates employment testing with several case studies. Although workers have innate differences in ability, pay method and other variables also affect how they perform. What applicants claim they can do is not always supported by their performance in a test. Also, when applicants are faced with a test, considerable self selection may take place as some chose to drop out. This self selection process is illustrated by the case study testing for an agricultural secretary position.

Tomato worker performance on the job can be predicted to some extent by brief trial periods of picking. On-the-job performance for piece-rate paid vineyard crew workers was predicted by a 46-minute performance test; the criterion, on-the-job pruning speed. Four field studies on three farms were done. Three showed significant relationships between test results and on-the-job performance.¹ The conclusion is that there is definite potential for the use of performance tests in agriculture. The results are especially important since the passage of the 1986 Immigration Reform and Control Act and other laws that are calling for more business-like farm management.

¹The fourth study supported the notion that low work performance consistency would result in nonsignificant test results.

OPPORTUNITIES FOR INCREASED WORKER PERFORMANCE

Is Agriculture Taking Advantage of Tests?

Employee testing is not new. More than a millennium B.C., Gideon selected for battle those warriors who brought water to their mouth rather than those who bowed down to drink. Around the turn of the century, the modern testing movement was launched by Francis Galton, James McKeen Cattell, and Alfred Binet (Anastasi, 1982). Major developments in testing occurred during the two world wars. However, today most farm workers are not tested but are simply selected on a first-come, first-hired basis.²

Economic Advantages of Testing

Employment tests can contribute to individual firms and to the economy as a whole by directly improving average productivity (Schultz, 1984).³ With the increase in the number of farms being run by farm management firms, it is not unusual to see many highly paid full time workers employed directly or as consultants to the farming enterprise. Dairies and other livestock operations also need year-round employees. However, this report focuses on workers at the lower end of the wage scale (some making less than \$6000 a year) and addresses the question: Is testing economical for these workers? It can be.

The longer the potential employment period of an applicant, the more complex and expensive the

testing can be and still be worthwhile. The more highly paid the workers or the greater the chance of damage to expensive equipment, livestock, or crops, the greater the testing expenditure that can be justified.

An even more important question than whether or not to test is: What types of tests are appropriate to the circumstances? Tests range from a simple interview⁴ to a battery of tests that last all day long. A day-long test for an assistant farm manager/supervisor may involve a written test, an interview, and a job simulation investigating supervisory, farming, and agricultural mechanic skills.

A farm manager who hired the wrong person the last time around was sued for wrongful discharge. To this farmer, no cost is too much to improve the selection process since the average wrongful termination suit is presently costing \$250,000 (\$400,000 in California). Some awards have reached the million dollar mark (McClain, 1987). While in the past, wrongful discharge suits only involved long-term employees, today suits are also filed by short-term ones.

A vineyard manager who cooperated in a testing program commented that he had no idea how time-consuming testing was. This same farmer, however, continued testing on his own the next year and recently asked the author for a statistical analysis

²The only agricultural-related references to testing were found in (1) Ghiselli (1966) who discussed high validity coefficients (0.55) for an arm dexterity test for selection of fruit and vegetable graders and (2) testing of agricultural pilots in Hungary (Lukacsko, 1984).

³Schultz showed that a selection test could result in savings in excess of \$5,000 per-worker-year when (1) the test had a validity coefficient of 0.5; (2) one in ten applicants was hired; and (3) the standard deviation in the value of a worker's production each year was \$6,000.

⁴Interviews are the most commonly used test, but also one of the most invalid tools unless they are well structured.

of the results at the season's end. This required (1) a brief job analysis, (2) detailing the job specifications, and (3) rater training. Figure 1 shows the resulting instrument for quality evaluation developed at this farm. In addition to the author, three foremen and two manager-level personnel were involved. The test development and training time cost about \$1200 (three days times eight hours per day times \$10 per hour average, times five people). A consultant might cost another \$1200. The cost of testing and data gathering cost another \$1200 at most, for a total cost of \$3600. However, all of these costs are not incurred again once the testing program is in place. Furthermore, nearly all the work in designing a test needs to be performed anyway to develop (1) job descriptions, (2) performance appraisal instruments, (3) quality control, (4) supervisory training and (5) worker training. The cost of (1) through (5) should not be fully charged to the selection process. Also, other vineyard managers could most likely use this test with only minor changes for different pruning methods. Where farmers have access to Cooperative Extension labor management farm advisors, the cost could be considerably less.

A labor management rule of thumb for year-round positions is that it is not excessive to spend the equivalent of one employee's pay for one year, for recruitment and testing. Benefits from testing come in the form of reduced costs, increased performance, or both. For sample costs of testing for a secretarial position in a farming operation, see Appendix A.

A test with high validity correlation coefficient (e.g., $r = .70$) can greatly decrease the percentage of job offers to unqualified persons (from 40 percent of all applicants hired, to 7 percent or less) with the use of a test (Anastasi, 1982). More moderate gains are associated with average validity coefficients.

The potential for savings by testing before hir-

ing a foreman, supervisor or farm manager is considerable; not testing can be very costly indeed. One farmer lost \$10,000 in alfalfa because he hired a farm supervisor who claimed he knew when to bale—and did not. Another agriculturalist lost \$70,000 in a hog operation in only three months, from a similar employee selection experience. On another farm, the hired manager planted a vineyard upside down. This wrong decision cost the farmer much more than the cost of the vines; production was delayed for one year. With a piece of farm machinery costing in excess of \$100,000, the person who is hired to operate it can make a big difference. The higher the decision-making responsibilities of the person being hired, the greater the potential costs of hiring the wrong person and the benefits of hiring the right one.

Two studies reported here involve selection of piece-rate paid workers. Some have asked, "Why bother with careful employee selection when I pay my workers piece rate? If they are slow they will just earn less." The reason is that, with testing, farmers can hire fewer and more productive workers. Benefits to the farmer from hiring *fewer*, better workers include (1) reduced paper work, (2) need for fewer supervisors, (3) reduced overhead for costs not associated directly with performance (e.g., vacation, health insurance), and (4) a more stable work force with an increased working season length for those workers who are hired. Benefits to the farmer from hiring more *productive* workers include (1) not having to pay the minimum wage to workers who do not pick enough piece-rate units and (2) a reduced danger of workers setting "bogeys" at very low production levels. (Bogeys occur when workers decide to pick no faster than an agreed-upon pace to prevent working themselves out of a job, protect slow workers from being embarrassed or fired, and/or prevent their employers from lowering the piece rate.)

Legal Testing Issues

In hiring it is illegal, for example (1) to assume that because an applicant is a woman, she cannot load three wire alfalfa bales onto the back of the pick-up or (2) to ask only applicants with an accent if they have a legal right to work in the United States. Farmers need to keep several categories (protected from discrimination by law) in mind, including: age (40 or older), sex, race, color, national origin, handicap, medical condition (cancer-related), and religion. Outright discrimination—or in the language of the courts, *disparate treatment*—involves differential treatment of people falling into these *protected* categories. However, it is legal to refuse employment to unqualified—or less qualified—applicants regardless of their age, sex, national origin or the like.

Adverse Impact

Courts look not only at disparate treatment, but also at *adverse impact*. For instance, requiring a high school diploma for tractor drivers might keep out proportionately more non-White applicants. On the surface there is nothing discriminatory about the practice—or perhaps even about the intent—but the policy could have an adverse impact on non-Whites. Another policy which might cause adverse impact would be to require all applicants to lift 125-pound sacks, regardless of whether they will be hired as calf feeders, pruners, office clerks, or strawberry pickers.

Equal employment opportunity guidelines use the 4/5th rule to establish adverse impact. If a given group (usually sex or race) is selected in ratios that are

less than 4/5th of another group, then there generally is evidence of adverse impact. However the rule is only a guideline. Where small numbers are involved or where the employer goes out of the way to try to recruit persons in under-represented groups, then the rule does not strictly hold.⁵

Employers, by law, are not legally required to hire unqualified workers, regardless of the 4/5th rule. They are, however, expected to show good business reasons for using a hiring procedure. (The *Uniform Guidelines on Employee Selection Procedures*, (Equal Employment Opportunity Commission, 1978) uses the word *valid* almost interchangeably with “good business reasons.”) For instance, an employer can give workers a pruning test and not hire those who do not have the skills required, or who prune too slowly. If it turns out that only the women who applied could do the job, the farmer would not have to hire any of the male applicants. But the employer could assume neither that other men who apply for the job later will not be qualified nor that all women will.

Some jobs are segregated by race or by sex or both. “Occupational concentration” refers to segregation not caused by the employer, but by the applicant’s choice. This phenomenon has long been evident in agriculture. Out of 120 persons who showed interest in the secretarial position, only one was a male. The more than 300 vineyard workers included no more than 10 non-Hispanics.⁶

Many employers hire indiscriminately: They hire everyone who applies. The problem with the indiscriminate-hire approach is that adverse impact

⁵The *Guidelines* (Equal Employment Opportunity Commission, 1978) also allow for adverse impact if there is proof of test validity or of high utility. However, the greater the adverse impact, the greater the proof required. Formulas for test utility and utility considerations can be found in Chronback and Gleser (1965), Hunter and Schmidt (1983), Schmidt and Hunter (1980), and Schultz (1984).

⁶Also of interest is that some farms in this study had both men and women employed in the vineyards, while in at least one vineyard there were *only* male workers. Therefore, there seems to be a difference between farms in hiring policy.

is postponed. Because everyone (including Caucasian males) belongs to at least protected groups (i.e., color, sex), it is likely that some in a protected group will be rejected from promotions or will be terminated, leading to possible litigation.

While tests can be misused, they offer a superior way to select employees without illegal discrimination, and are an improvement over subjective methods such as interviewing (Barrett, Phillips, and Alexander, 1981; Daniel, 1986; Doverspike, Barrett and Alexander, 1985; O'Leary, 1973; Tenopyr, 1981; Whelchel, 1985).

Developments in administrative and case law have made employee termination more difficult. Employers are often told that effective employee selection is the first step in avoiding wrongful discharge litigation. Promises or statements made to workers when they are hired, in conversation with foremen or supervisors, or in employee handbooks, have given rise to much litigation—for example, references such as “permanent employee” or “as long as you do good work you will have a job.” Some employers who have discharged a “permanent” employee have ended with a wrongful discharge suit. They have been charged with breaking an implied contract of good faith (Billikopf, 1987).

Employees are winning in court even though there are *no laws* prohibiting employers from firing workers (except in retaliation for whistle blowing or for discrimination because of sex, race, national origin, and so on). Wrongful discharge cases are becoming more frequent in agriculture.

Other Legal Issues

When workers are injured in an employment test, they can be covered under workers compensation even though they are not on the employer's payroll when the accident occurs.⁷ Employers need to take special precautions to prevent accidents when workers are trying out for hazardous jobs for which they claim competence but may actually lack the necessary skill.

Employment testing falls into a continuum from pre-employment to an on-the-job test. An example of the former is an equipment operator who loads and unloads a tractor from a ramp or a manager who answers questions in an interview; the farmer is not getting a “product” or “real work” out of the applicant. An example of the latter is a farmer who tries out a dairy worker for a day milking cows—a product results. Individual states may well develop different policies regarding payment for this work. Some representatives of the Division of Industrial Relations in California contend that if workers are asked to prune grapes, milk cows, or plow a field, they must be paid for their time, because they are “permitted to work.” Policy in this area remains unclear.

Among the many benefits of pre-employment testing, where applicants are not officially hired until after the selection process, is *not* having to (1) place all applicants on payroll records, (2) fill out the many forms required in employment (including the I-9 employment verification form from the Immigration Service), and (3) experience an increase in unemployment insurance costs.

⁷See, for example, *Laeng vs. Workers Compensation Appeal Board* (California Compensation Case, 1972). The court decided “Where an employer requires an applicant to take a physical agility test on a course designed by the employer as a precondition to employment, and the applicant is injured while taking the test, the applicant is rendering service and exposing himself to a ‘risk of employment,’ and his injury is therefore compensable.” A New York court case (*Smith v. Venezian*) was among numerous cited by the California Supreme Court in making its decision. “[It] is ... our view that where a tryout involves an operation that would be ordinarily viewed as hazardous ... a special employment exists.... A tryout is for the benefit of the employer, as well as the applicant, and if it involves a hazardous job we see no valid reason why the applicant should not be entitled to the protection of the statute.”

Potential areas for compromise in the pay-for-work-done-while-being-tested issue include (1) limiting the total amount of time for tests relating to "borderline work" and (2) paying all applicants for the vines they pruned, the fruit they picked, etc.—if they are selected for the job.

If compromise is not possible then selection tests in some states will be more expensive, but may still be worth the extra cost. Workers who are paid for a one-hour test and then let go, are less likely to sue for "wrongful discharge" than untested workers who are allowed to stay for several weeks and then are terminated. It is likely, however, that farmers will be reluctant to test if they have to write out many extra checks (and do other paper work) for people they do not hire.

Perhaps the best and most thorough discussion of employee discrimination and testing from a legal perspective can be found in Schlei and Grossman (1983). Other references are Siegel (1980) and Ramsay (1981), Kleiman and Faley (1985), and Bersoff (1981).

Technical Terms Used in

Employment Testing Procedures

A careful and properly targeted selection process is said to be "valid." Validity is a *descriptor* for the quality of the selection process. Selection may involve a single tool for obtaining information about applicants (e.g., applicant interview) or utilize a combination of several selection tools (e.g., three tests, an interview, an application blank, and reference checks). All these tools are considered *predictors* of job performance. Selection tools (or predictors) are used by employers to predict a single result (e.g., pruning speed) or multiple results (e.g., productivity, absenteeism, turnover, safety record). The results that are being predicted by the predictors are called the *criterion* or the *criteria*.

Validity describes *what* the selection process measures and *how well* it does so. The "how well" part of the definition refers to the strength of the correlation between the predictor and the criterion and their reliability, i.e., the consistency, of both the predictors and the criteria. The "what" implies that a test can never be "valid" on its own, that it is only made valid or not valid with respect to its utility or predictive ability in a given use. The better that predictors forecast results or criteria, the more valid the selection process is said to be.

Reliability

For a test to be valid, it must be reliable. The more unreliable a test is, the more invalid it will be. Reliability involves the *consistency* of a test in measuring something. For instance, how consistently can a degree brix meter measure sugar content in table grapes? How consistently can a tension meter measure soil water content? Or a scale, the weight of a calf? Reliability refers to the ability of a test to give the same results time after time. Test results differ because of differing *content*, the *time* between tests, and how they are scored by *different raters* (or even the same rater at different times).

Examples of reliability problems that might be encountered include:

1. *Test-retest.* The exact test is given twice on two different occasions. The type of error associated with this procedure is called "time sampling" (Anastasi, 1982). It is likely that a person learns something or forgets something between the first time and the second time a test is given. How much difference there is between one test result and another depends in part on the type of test, and in part on the level of skill possessed by those taking the test.
2. *Alternate-form reliability.* Two different forms of the same test are given. When an alternate form test is

given at two different times there are two sources of error: content and time (Anastasi, 1982). Time error is similar to test-retest just discussed. Content error is when skills or abilities required on one test differ from those on another.⁸

3. *Scorer reliability.* Different persons may score a test differently (Anastasi, 1982).⁹ Objective tests have fewer problems than subjective ones, but even objective tests are not free from scorer reliability.

Careful analysis of what constitutes a good response to an employment interview question—or a good job in a performance test—helps raters or judges improve their ratings. It is also important that all judges are using the same rules and time limits to rate. If one foreman allows more time or gives different instructions to applicants taking a mechanical test than another foreman giving the same test, chances are that applicants' scores will vary depending on the foreman giving the test. Even if only one person does the rating, there can be rater error from one test to another. Some of the same factors that can help improve consistency among raters can help improve consistency for a single rater, too. Not only is the reliability of the test (or predictor) important, but also the reliability of the job performance measure (the criterion).¹⁰

Reliability of the predictor and the criteria is usually measured in terms of a simple Pearson's "r" correlation coefficient. Many low-cost calculators are available today that will quickly compute the coefficient :

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

where r is the correlation coefficient (here, the reliability coefficient); x is one variable, e.g., Test 1 results; y is another variable, e.g., Test 2 results; and n is the number of pairs involved. The value of r ranges from -1 (a perfect inverse relationship), through zero (no relationship indicated), to +1 (a perfect positive relationship). For predictors or criteria to be "reliable," they should show a large, positive or negative correlation coefficient, i.e., close to +1 or to -1.

Correlation analysis is one of the principal statistical tools used in employee selection testing. The weakness of correlation analysis (e.g., Leedy, 1985; Little and Hills, 1978) is that it is often improperly used to claim a cause-and-effect relationship. The purpose of using correlations here is not to show causality, but rather, to show *association* or *closeness* of the relationship.¹¹

⁸If, for instance, the selection procedure for hiring a vineyard production manager involves one test with questions on eutypa and mildew and another test with questions on phyloxera and grape leaf skeletonizer, it is possible that an applicant might do very well on one test and poorly on the other.

⁹For instance, opinions differ among the five dressage judges in the equestrian events at the Olympic Games; opinions differ among three veterinarians as to the cause of a disease in a young bull; and foremen differ in their judgment of a worker's pruning quality.

¹⁰Wernimont and Campbell (1968) and Ghiselli (1966) point out that few studies have established the reliability of the criterion measure. Uniform crew working speeds result in an unreliable criterion measure. An excellent test is not a substitute for criterion reliability (Ghiselli, 1966, Green, 1981). Green says: "Most performance measures are much less reliable than the tests they are validating. An unreliable criterion is just as limiting as an unreliable test" (p. 1006).

¹¹If the relationship is not linear, prediction would be more accurate at some ranges than others and could result in an underestimation of validity (Ghiselli, 1966).

Validity

It is possible for an instrument to be consistent yet useless in predicting success on the job. That is, a test can be reliable but not valid. For example, a tree pruning test might be very reliable in predicting the fastest workers in a peach orchard crew. A farmer could conceivably use such a test, as consistent and reliable as it is, to also try to predict quality of work. The relation between speed and quality, however, might be very weak or even nonexistent. This test, while reliable, is invalid—for the purpose of predicting pruning quality.¹²

There are various strategies to establishing validity. Two of these are explained in this report and are illustrated using case studies: (1) criterion-oriented and (2) content-oriented. Another type of strategy is known as construct-oriented. This involves testing various psychological constructs such as motivation, intelligence, and personality.

Regardless of which strategy a farmer uses to establish validity of the selection system, a thorough job analysis is essential. A job analysis entails collecting information about the job through worker and foreman interviews, surveys, and observation. A job analysis can, in turn, be made into a list of job speci-

fications or specific requirements.¹³

The job analysis is important because it is the data base for the job specifications. If something meaningful is missing from the job analysis, a significant factor might go untested in the selection process.

A farmer must not feel limited to thinking of a job the way it has been done in the past, but rather, should feel free to add skills that might be needed in the future and discard those that might not. Finally, not every item listed in the job specifications is of equal importance and items should be weighted accordingly (as in Figure 1).

Criterion-Oriented Strategies

A criterion-oriented strategy is one in which a statistical inference is made between the test (predictor) and the results (criteria). Again, this is done through the Pearson correlation coefficient. In this case "r" is referred to as a validity coefficient. Instead of x representing one set of test results and y the other in calculating the reliability coefficient for a predictor, here the validity coefficient x is the test, say Test 1, and y, the criterion, say, Criterion 1.¹⁴ Thus, the criterion-oriented strategy is a statistical approach. In principle, as long as the predictor is not outrightly

¹²There are other examples. A farmer may test melon picker applicants for speed and hire the faster workers. Then they are placed in crews paid by the hour. The motivation of these workers to perform in the test and to perform on the job might be very different. In fact, when employees work in a crew there are many social forces that tend to keep them working together at the same speed. A farmer might also retest workers for speed every year as part of their performance appraisal. Such a test does not measure how many melons an hour are actually picked by a worker on the job, but rather, it measures how many melons an hour a worker is capable of picking.

This does not mean that a test could not be used to hire crew workers who are paid by the hour. Certainly such a test could be used to eliminate applicants who cannot keep up with the main group. Also, it is possible that if all workers are selected at the beginning of the season through a melon-picking test, such a carefully-selected crew will move at a faster pace. Problems would arise if the farmer is adding newly tested-and-hired workers to an already existing hourly-paid crew consisting of fast and slow workers. Some farmers argue that crew workers paid by the hour work no faster than the slowest in the crew.

¹³For instance, if a calf-feeder must be able to lift 50 lb. grain sacks and 100 lb. cement sacks for construction of a new barn, the job specification would say: "ability to lift 100 lb. sacks." If a job analysis shows that a secretary must type letters and also must type reports, a job specification would restate such requirements as simply: "ability to type."

Figure 1. Pruning Quality Data Collection Instrument
(This instrument was used as a part of the testing process for the predictive test on Farm 3.)

PRUNING QUALITY (CALIDAD DE LA PODA)							
	SELECTION OF FRUITING WOOD Selección de la madera frutal	GOOD = 0 - 1	3	2	.4 x		
		FAIR = 2 - 3	1	0			
		POOR = 4 - 5					
	PLACEMENT OF SPURS Colocación de los pitones (dagas)	BUENO = 0 - 2	3	2	.3 x		
		REGULAR = 3 - 4	1	0			
		MALO = 5 - 6					
	NUMBER OF SPURS Número de pitones (dagas)	GOOD = 0 - 2	3	2	.2 x		
		FAIR = 3 - 4	1	0			
		POOR = 5 - 6					
	LENGTH OF SPURS Largo de los pitones (dagas)	BUENO = 0 - 2	3	2	.2 x		
		REGULAR = 3 - 4	1	0			
		MALO = 5 - 6					
	CLOSENESS OF CUTS Corte a nivel con la madera vieja	GOOD = 0 - 2	3	2	.2 x		
		FAIR = 3 - 4	1	0			
		POOR = 5 - 6					
	ANGLE OF CUT ON SPUR Angulo del corte del pitón (daga)	BUENO = 0 - 2	3	2	.1 x		
		REGULAR = 3 - 4	1	0			
		MALO = 5 - 6					
	DISTANCE OF CUT FROM LAST BUD Distancia del corte a la última yema	GOOD = 0 - 2	3	2	.1 x		
		FAIR = 3 - 4	1	0			
		POOR = 5 - 6					
	REMOVAL OF SUCKERS Eliminación de chupones	BUENO = 0 - 2	3	2	.1 x		
		REGULAR = 3 - 4	1	0			
		MALO = 5 - 6					
			MISTAKE TOLERANCE		SCORE	FACTOR	TOTAL:

¹⁴Researchers see a great future in criterion-oriented validity (e.g., see Schmidt and Hunter, 1980). However, Robertson and Kandola (1982), and Wernimont and Campbell (1968) found that many researchers confuse validity with reliability.

Lee, Miller, and Graham (1982) and Mount, Munchinsky, and Hanser (1977) feel that having a different predictor and criterion measure is what distinguishes a validity from a reliability coefficient. Mount, Munchinsky, and Hanser used an open job-sample test as a predictor and a different, but more complex, open job-sample test as the criterion. Such a comparison ignores worker motivation on the job, but could be considered a validity study measuring the capacity of a pre-employment test to predict *success in training*. However, it can only be considered a test of validity if actual performance on the job is being correlated against the pre-employment test. Ebel (1977) argues: "Ability to do ... work is a necessary, but not sufficient condition for success ... [and] the success of a person ... on a job depends to a considerable extent on the efforts of the person" (p. 60).

Schmitt et al. (1984) found few studies that used production as the criterion. More often, subjective performance ratings are used, resulting in lower validity coefficients because of the lack of reliability in the criterion scores.

Daniel (1986) said that "the best opportunities to improve selection exist in organizations in which one or more readily indentifiable, quantifiable characteristics affect organizational performance" (p. 6).

Green (1981) suggests that 50 or more cases are required to establish some credence for a validity coefficient. A sample of 100 or more data pairs is needed to establish a solid base for a study (Ghiselli, 1966; Green). Schmidt and Hunter (1980) call for a much larger sample. But Schmitt et al. (1984) had good results with small sample sizes. Finally, Ramos (1981) found that offering test instructions in Spanish—to those who preferred it—resulted in "small but significant" test score improvements.

discriminatory, it does not matter what the factor is, if it predicts performance. For instance, if women prove to be better milkers, the factor *women* would be illegal and could not be used; a farmer cannot reject men on the basis that women make better milkers. But if a dexterity test is a good predictor of milking ability, and more women passed that test than men, then more women could be hired for that job *based on the test*.

Two ways of carrying out a criterion-oriented strategy are (1) a predictive study and (2) a concurrent study. In a predictive study, all applicants are tested, but normally they are hired without the benefit of the test results. Test results are *not* given out to supervisors as they could contaminate the data (by influencing the supervisors). After workers have been on the job for a period of time, test results are correlated with some measure of on-the-job performance, i.e., the criteria. The better the test predicts success on the job, the more valid it is. This method works well when a farmer will be hiring many new workers.

The second or concurrent approach involves testing those already holding a certain job to see how they do on a test. Performance data and supervisor ratings for incumbents may be collected before the test is given, eliminating contamination problems. If the test proves to be valid, that is, if there is a high correlation between the test results and the performance data, the farmer can use the test with some confidence with new hires.

The traditional view is that a predictor strategy is superior to a concurrent one for most personnel situations (Anastasi, 1982; Ghiselli, 1973; American

Psychological Association, *Standards*, 1985; and Guion and Cranny, 1982). The argument often presented is that concurrent validation strategies may introduce a large *restriction of range* error—with corresponding lower validity coefficients (Guion and Cranny). The problem arises because workers in a particular job tend to be a more homogeneous work force than an applicant population, resulting in a restriction of range problem and lower validity coefficients. This on-the-job homogeneity may occur when poor performers drop out or are terminated and best workers are promoted (see Figure 2).¹⁵

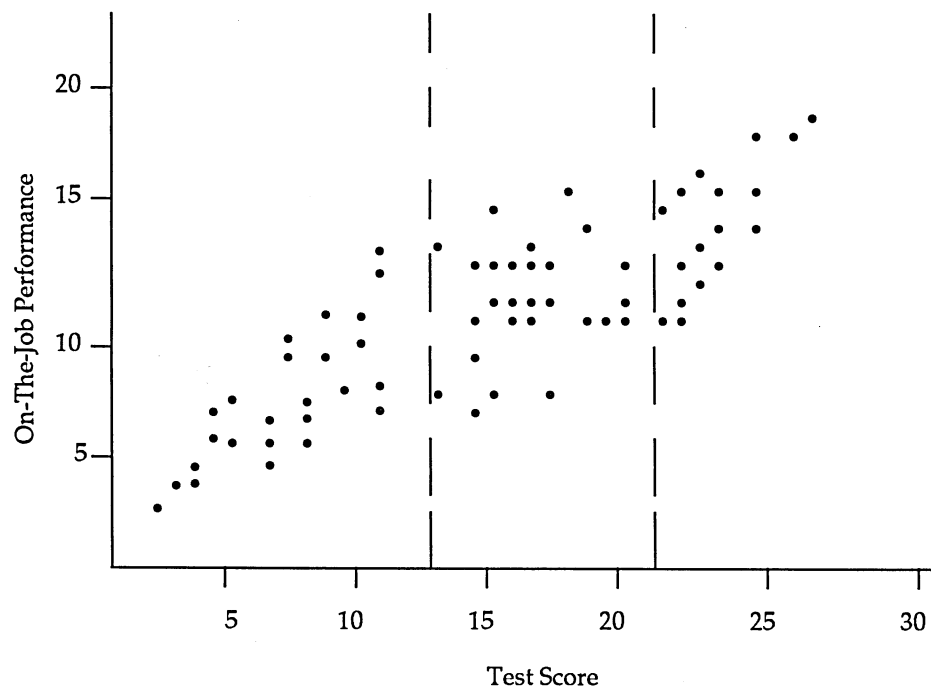
Another problem with concurrent-oriented studies is that workers may not have as great a motivation to do well in a test as in a predictive study (Guion and Cranny, 1982; *Principles*, 1980). Guion and Cranny feel that these differences in motivation introduce random error in the predictor.

Content-Oriented Strategies

A content-oriented strategy is one that emphasizes a comparison between the content of the job and the content of a test. The assumption is that an applicant's being tested in a job-related area, contributes to validity. Thus, it makes sense for a herdsman who performs artificial insemination (AI) to be tested in AI, for a farm clerk-typist to be given a typing test, and so on. The danger with the construction of a content-oriented test is that people tend to be tested in those areas that are easiest to test. For instance, if a job necessitates driving equipment, identifying plant diseases, irrigating, and doing some farm supervision, a farmer might test someone mostly on pathol-

¹⁵Schmitt, Gooding, Noe, and Kirsch (1984) hold the opposite view. They found higher validity coefficients for concurrent than for predictive studies. (However, they were not able to control any variables in their meta-analysis.) The differences between concurrent and predictive studies have increasingly been minimized by others (Barrett, Phillips, and Alexander, 1981; Division of Industrial and Organizational Psychology, *Principles*, 1980), especially due to a host of corrections that have been developed for restriction of range (e.g., Lee, Miller, and Graham, 1982).

Figure 2. The Restriction of Range Problem; On the Job Performance (in terms of vines pruned, lugs picked, etc.) versus Test Scores



When all the data points are considered in this bivariate distribution between test and on-the-job performance scores, the validity coefficient is high ($r=0.84$). However, if a farmer had set a minimum score of 13 for selecting applicants, the coefficient drops to $r=0.68$. Then, if the best workers are promoted to other positions and only the middle range is considered, the coefficient would drop to $r=0.14$. This illustrates the restriction of range problem.

ogy identification, and only a little on driving equipment. To the extent that the areas not tested are important parts of the job, the process is likely to be invalid. A content-oriented process, then, will be more valid when persons are tested in a greater variety of areas. The argument for using the test, here, is mostly a logical one rather than a statistical one.

Ideally, a test or battery of tests might include two or more strategies. The better the job is understood, the better the chances that a selection process can be designed to account for all the factors that determine whether an applicant will be a successful worker.

Summary

Any tool that attempts to measure an applicant's knowledge, skill, ability, education, or even personality can be evaluated by how consistent it is (reliable) and by how well it predicts results (validity)—such as worker performance, theft, or length of employment. There are several ways to establish the validity of a selection instrument. The more valid a selection process, the better the chances a farmer has of hiring the right person for the job—and of staying out of trouble with the law—or winning a case in court when challenged.

Work-Sample Testing: A Special Case

Personnel tests can be classified as those measuring ability, motivation, personality, performance, and even honesty. Tests can also be classified as (1) pencil and paper tests, (2) situational (e.g., What would you do if ...?), (3) job simulation, (4) in-basket exercises, and (5) work-sample (often termed "job" sample) tests. Not all of these are considered equally valuable and many combinations are possible. The case studies reported here involve at least one work-sample test or one job simulation test because these are particularly well accepted.

Work-sample tests are versatile in that they may be validated through content or criterion strategies; stringent job analysis justification is not required for a production criterion, and multiple sources of evidence may be useful to examine the validity of inferences derived from a test. These tests are often considered superior for use in employee selection, as long as work methods don't change (Kleiman and Faley, 1985; Mount, Muchinsky, and Hanser, 1977; Robertson and Kandola, 1982; Schmitt et al., 1984; Standards, 1985; Whelchel, 1985).

Work-sample tests also increase the "face validity" (i.e., what it seems the test is about) of employment tests (Wernimont and Campbell, 1968).¹⁶ Those who take the tests, and judges in courts of law, can see the connection between the test and the job and are more likely to develop favorable opinions about the test (O'Leary, 1973; Schmitt et al., 1977).

In some cases job-sample tests have reduced adverse impact in employment decisions (Robertson and Kandola, 1982; Schmitt et al., 1977; Whelchel,

1985). The validity of such decisions, unfortunately, has seldom been tested empirically against measures of job performance.

Work-sample tests also promote self selection (Downs, Farr, and Colbeck, 1978; Farr, O'Leary and Bartlett, 1973; Robertson and Kandola, 1982). Self selection occurs when applicants realize that they are *not* really qualified for a job, or that a job does *not* suit their economic, emotional, social, or other needs. Such applicants "select themselves" out of contention for a position.

Mount, Muchinsky, and Hanser (1977) found that work-sample tests have high reliabilities. Comparison between work-sample tests and on-the-job performance also resulted in significant validity coefficients. Often, predictor (test results) and criterion (on-the-job performance) can be measured in similar ways (e.g., pruning speed, number of buckets picked, typing speed and quality).

Worker motivation is probably different during a selection test (where a job is on the balance) than it is on the job. Nevertheless, workers who do only half as well as others when trying their best under test conditions are unlikely, no matter what the motivation (e.g., extra pay for extra production), to catch up to the faster workers on the job. Farmers are more likely to adopt work-sample tests over other tests because they are easy to understand, simple to administer, and reasonably effective.

Differences in Worker Performance

Most farmers are comfortable with the notion of adapting varieties of plants and breeds of animals

¹⁶Face validity refers to what a test appears to measure on the surface. For instance, a farmer wanting to test for a herdsman's knowledge of math would do much better to use test problems that involve cows, barns, and Dairy Herd Improvement Association records than, say, examples that use marbles and baseball caps. At times, however, the purpose of a test is less apparent. For instance, it might be easy for an applicant to lie in a personality or honesty test that clearly has proper and improper responses, but more difficult when the answers are clouded by a hidden purpose.

to different uses. When it comes to workers' differences, however, some agriculturalists neglect the great variability in people and their performance. For virtually every task there would seem to be workers who can perform better than others. Workers who excel in one area, however, might not compare so well with others in a different task.

Schultz (1984) reported that at times the best worker can be four times as productive as the worst worker. This kind of difference—if consistent—can greatly increase the utility of a test to select the more productive workers. Productivity is a result of both worker *ability* (the “can do”) and *motivation* (the “will do”); the greater the differences in applicant abilities, the more effective a test can be in identifying these differences. Motivation—the “will do”—is equally important. A key motivator in most employment situations is, of course, pay. The method of payment may also influence motivation.

Paying farm workers by the piece may tend to bring out worker differences, while paying them hourly may camouflage them. One setting in which farmers can observe individual differences in productivity is in vineyard pruning. In most grape growing operations each worker is assigned one row to prune. When workers are paid on an hourly basis, they tend to finish their respective rows at almost the same time. Before moving on to a new row, workers often help others to finish theirs. While slow workers feel pressure not to be left far behind the main group, fast workers are also pressured into not leaving the main group far behind.

When paid by the vine, pruners will usually spread considerably throughout the field. When workers come to the end of their rows, they start a

new one. The group-cohesiveness that fosters homogeneous work speed among hourly-paid workers is tempered by the desire to increase personal earnings under piece-rate pay.

Reduced differences in speeds when workers are paid hourly may result in an unreliable criterion measure. Any personnel practice that tends to increase manifestations of individual differences, conversely, is likely to increase the criterion reliability. In the first case study, differences between piece-rate paid and hourly paid crew workers are explored.

Hourly versus Piece-Rate Paid Vineyard Pruners

Labor pruning-rate data for nine crews from seven farms in the California San Joaquin Valley were examined. All data—vines pruned per person-day and number of hours worked per person-day—were collected by farmers and their foremen. Data for four days were collected for each of 10 workers (randomly selected) per crew for eight of the crews. Data from three days were collected from 13 workers (complete crew) for the ninth crew. (No data were collected for day 4 for this crew.) Six crews were paid by the vine and three by the hour.

The data were analyzed for criterion reliability—individual worker performance between two different work days—using a sample Pearson's “r” reliability coefficient. In addition, Analysis of Variance (ANOVA) was used to test differences in the variation (1) between workers and (2) between days (vineyard, weather, or other differences between days). ANOVA determines if the differences observed (between workers or between days) are statistically significant.¹⁷

The hypothesis: Criterion reliability for crew

¹⁷Variance is relative, so the greater the differences in working conditions from one day to the next, the greater the real differences that must exist between individual workers for the test to be statistically significant.

workers is dependent on pay method. Piece-rate paid workers should have larger coefficient values, while hourly paid workers should not. The ANOVAs should show statistically significant variance among workers when paid by the piece, but not when paid by the hour.

Results and Discussion

The analysis confirmed visual observations: Workers paid by the hour tend to cling together, while those paid by the vine tend to spread out, with some working much faster than others. Nevertheless, there were exceptions and differences among crews not wholly explained by pay method.

Analysis of the data (Table 1) shows that all six piece-rate crews had greater variance (by ANOVA, randomized complete block design) among workers than among days worked. The three hourly-paid crews, on the other hand, showed more variance in workers' output from day to day than among workers on any given day, offering support for the hypothesis. In five of the piece-rate crews the variance was significant beyond the 99 percent confidence level. Workers pruned at differing rates—and did so rather consistently from day to day. However, among piece-rate crew 6, the test was not statistically significant.¹⁸ Meanwhile, there were no significant individual performance differences found among hourly crews by ANOVA.

None of the correlation coefficients between two days for the hourly-paid workers exhibited a strong relationship; all of the piece-rate coefficients were strongly positive. Differing distributions of production (i.e., numbers of vines pruned) within crews paid by the vine and those paid by the hour are shown in Figure 3. These differences suggest that

workers in hourly-paid crews either are selected for their similar abilities or, more likely, tend to restrict their potential output.

Thus, farmers and other employers should carefully consider pay and other factors that motivate workers to do their best. While this is probably true for virtually all personnel management situations, it is especially true for agricultural field crew conditions where workers can easily see how their speed compares to that of others.

A Limitation on Case 1 Results

While normally a worker who finishes a row will help another, workers were asked by their employers not to help each other during the study period. This departure from the usual pattern was necessary to enable the grower to count the number of vines pruned by each worker on each day of the study. Unfortunately, the criterion could have been contaminated by this change in the normal procedure.

It would have, of course, been preferable to have a farmer implement the work-on-your-own system for a while before collecting any data. But this would have required too great a commitment among the participating farmers, for one of the benefits of paying by the hour is not having to count the number of vines pruned per worker each day. Equivalent problems did not exist with the piece-rate crews; growers collected the same output data they normally need for payroll purposes.

Other Motivational Factors Besides Pay Method

Another case study examined individual pruning rates for two grapevine pruning crews (Crew 1

¹⁸One piece-rate crew had several workers with the same last name; another had a husband and wife who worked at the same speed. These features help to explain reduced differences between workers found for these crews.

Table 1. Test of Statistically Significant Differences
Between Hourly and Piece-Rate Paid Crews
of Grape Pruners

	ANOVA		Criterion Reliability ^a
	<i>Hourly-Paid Crews</i>		
Crew No.	F-statistic		Pearson's r
1	.04	457.23***	-0.47
8	.82	225.66***	0.17
9	1.47	127.19***	0.06
	<i>Piece-Rate Paid Crews</i>		
2	22.48***	22.11***	0.92
3	6.09***	1.40	0.73
4	21.76***	3.07*	0.96
5	11.98***	1.84	0.82
6	2.05	0.91	0.79
7	13.75***	5.31**	0.69

a. Days for correlations were randomly selected as day 2 and day 4 for all crews except crew 2; day 1 and day 3 were used for crew 2 because of missing plots on the other days.

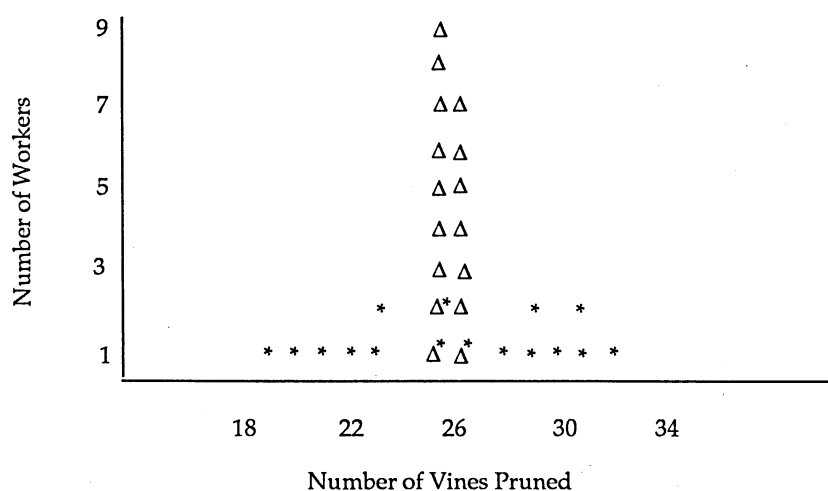
* = the probability is less than 5 % that the difference observed occurs by mere chance; ** = less than 1 %; *** = less than 0.1 %.

and Crew 2) within the same farming operation in the San Joaquin Valley; both crews were paid on a piece-rate basis.

Crew 1 had 18 workers. Each pruner's work-rate was recorded as average number of vines pruned per hour on each of four consecutive days. Table 2 reveals that there is much variability between workers, and, perhaps more significantly, workers are consistently different. On day 1 the slowest worker pruned an average of 33.4 vines/hour; the fastest, 90.1. Faster workers on day 1 tended to be more productive on days 2-4, also. Analysis of the data rejects the null hypothesis that such observed variability among workers was merely by chance. Crew 1, then, looks like a typical crew under piece rate, with workers showing consistent differences from day to day.

Crew 2 had 17 workers whose pruning rates ranged from 44.8 to 67.2 vines per hour on day 1 (Table 3). Statistical analysis of Crew 2 worker productivity also rejects the null hypothesis that worker differences occurred by chance. However, within a large subset of Crew 2, the grower's payroll records show that seven out of ten workers (among workers 1 through 10) pruned exactly the same number of vines on day one; so did the remaining three. On day two, another seven among these 10 workers pruned at the same rate. On the third day, four workers out of 10 pruned at the same rate. Considering only subset A, we would not reject the null hypothesis. That is, there were no statistically significant differences

Figure 3. Distribution of hourly and piece-rate paid crews scores (distributions adjusted and superimposed to have the same mean)



Average vines pruned per hour by 16 piece-rate workers (*) and 16 hourly workers (Δ) in one day.

Table 2. Average speed (vines/hour for 18 workers in Crew 1 on each of four consecutive days).

Worker	Day			
	1	2	3	4
1	43.9	47.7	43.4	53.7
2	37.8	38.1	49.0	41.5
3	33.4	35.8	48.6	39.1
4	46.3	34.4	54.8	46.9
5	46.0	45.0	39.2	43.9
6	48.1	42.9	42.0	47.4
7	58.8	45.8	67.6	58.8
8	76.0	58.8	74.0	72.8
9	41.8	36.7	59.2	44.2
10	44.2	47.9	43.7	53.9
11	46.1	44.1	52.4	46.1
12	44.5	38.8	43.4	41.6
13	90.1	58.8	69.2	70.6
14	44.8	37.4	54.4	51.5
15	59.1	44.8	72.0	52.5
16	47.6	37.6	70.2	51.1
17	34.9	38.5	38.2	37.8
18	76.5	58.8	68.6	70.6

Table 3. Average speed (vines/hour for 7 workers in Crew 2 on each of four consecutive days).

Worker	Day			
	1	2	3	4
1	67.2	58.8	64.8	54.1
2	67.2	58.8	64.8	65.9
3	58.3	58.8	64.8	60.0
4	67.2	64.2	47.2	54.0
5	58.3	52.5	53.4	47.9
6	67.2	58.8	66.5	60.0
7	67.2	69.8	60.5	65.9
8	67.2	58.8	53.3	54.0
9	67.2	58.8	64.8	54.1
10	58.3	58.8	53.5	65.9
11	56.3	49.3	47.5	44.9
12	52.1	63.9	65.4	54.0
13	44.8	39.8	41.5	42.4
14	60.0	76.5	64.7	70.8
15	54.3	63.4	56.9	42.0
16	47.3	52.5	53.5	42.0
17	52.5	56.5	44.7	43.2

¹⁹In many settings, but especially in agriculture, where variations in conditions are pronounced, farmers ask employees to first work on an hourly basis so that a piece rate can be set. Employers derive a piece rate from the production rate experienced in the initial "trial period." The higher the trial production rate, the lower the piece rate is set. Workers get a more favorable piece rate if they don't work too fast during this period. Once the piece rate is firm, they tend to produce up to their ability. A slow down, while being paid by the piece is sometimes used to reestablish a higher rate.

among the workers in subset A. These 10 workers, especially during the first part of the week, did not behave as typical piece-rate paid workers.

Implications

The first and most important implication from both of these case studies is that workers have differing abilities. Farm operators can make use of such differences if they understand how to select and motivate employees. However, pay incentives are not the only influence on worker performance. Possible explanations of the lack of statistical differences among subset A of Crew 2 range from a deliberate work slowdown (a bo-gey) to try to induce the grower to increase the pay per vine,¹⁹ to a general desire to stay close for social contact while working. The manager of the vineyard noted that there were several related workers in the crew.

Summary

Differences in workers with respect to pruning productivity are very real and important. Of course, paying piece rate does not always guarantee that worker differences will be brought out. Testing is one way to determine these differences and take advantage of them when hiring.

AGRICULTURAL EMPLOYMENT TESTING: CASE STUDIES

A Content-Oriented Strategy: Agricultural Secretary Selection

A secretarial job was analyzed and job specifications were laid out. In developing the testing strategy, particular attention was paid to testing for skills that would be needed on a day-to-day basis on the job. That is, a content-oriented validation strategy was followed. A short employment advertisement specifying qualifications, including typing speed—60 words per minute (wpm) minimum—and artistic ability, was run twice in the largest and only local daily paper. Other recruitment efforts were made at a local college.

The correlation between what applicants said they could do and how they performed on a test was measured.²⁰ One-hundred eight complete applications were received, plus additional inquiries, resumes, and incomplete applications. Most of the 108 were invited to demonstrate their artistic ability. The test consisted of (1) writing "Agriculture is California's Future" using dry transfer letters, (2) designing a flyer/poster, and (3) free hand drawing an object. Only about 60 of the applicants showed up for the art test. A few of these did not stay after the procedure was explained. Others left before completing the exam.

The quality of the art work, which varied enormously, was evaluated by three raters. The 25 applicants who performed at a satisfactory or better level were scheduled to be tested for typing speed and for spelling and punctuation.

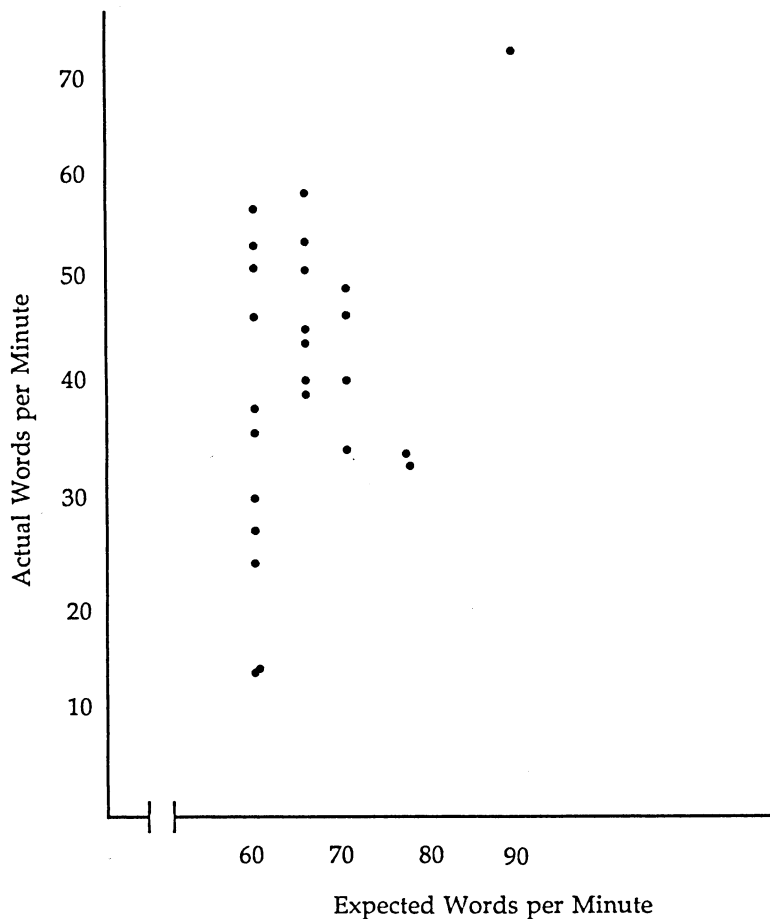
Actual typing speeds ranged from 15 wpm to 76 wpm. The 11 (of 25) applicants claiming typing speeds of 60 wpm on the application were rather evenly dispersed from 15 wpm to almost 60 wpm (see Figure 4). The range between expected and actual typing speeds narrowed for those who claimed they could type faster than 60 wpm. The average *claimed* typing speed was 65 wpm while the *actual* average speed in the test was about 44 wpm.

In the spelling and punctuation test, applicants were provided a dictionary and were asked to re-type a letter and make any necessary corrections. This portion of the exam allowed sufficient time for applicants to finish, review their work, and re-type if necessary. Applicants ranged from those who found and corrected almost every mistake in the original letter and re-typed it without changing the meaning to those who missed many of the mistakes and took correctly spelled words and misspelled them. Eight persons qualified for a final interview; three of these showed the most potential; one was selected unanimously by a five person panel.

This content-oriented study has "face validity" in that the test was directly related to the performance required on the job. It brought out the differences in more than 100 applicants, all of who claimed to have some degree of artistic ability and secretarial skills. Had applications been taken on face value and top candidates interviewed, it is likely that a much less qualified candidate would have been selected. The applicant who was hired, never would have been

²⁰Some applicants reported their typing speed even though it was below 60 wpm, while most indicated at least a 60 wpm speed, perhaps realizing they would probably never even get an interview if they did not. For those who claimed they could type better than 60 wpm there was generally a better correspondence between expected and actual skills because they had less incentive to exaggerate. Slower typists did not want to report speeds over the minimum required, especially since they might be accountable for typing that fast if they were selected for the job.

Figure 4. Typing Speeds



interviewed in a usual selection process as she had much less secretarial job experience than many of the other applicants.

All along self-selection was occurring, with applicants dropping out at different stages of the selection process, especially at the beginning. Future studies on the effect of minimum standards on skill self-reporting would be valuable.²¹

A Criterion-Oriented Strategy:

Tomato Harvest Testing

The purpose of this portion of the study was to determine whether a work-sample test—when workers know they are being observed—can be used to predict actual tomato harvest worker performance (when they do not know they are being observed). Workers change in performance (usually improvement) when observed and/or special attention is paid

to them is known as the Hawthorne effect. In this test and in the next (with grape pruners) test results did predict performance well, despite the occurrence of some Hawthorne effect during the test.

The test was with hand-harvest of green tomatoes on a San Joaquin Valley farm, summer 1986. Farm workers pick into two buckets which they carry to bins where they receive one chip; they are paid by the number of chips they collect per day.

Methods

A concurrent criterion-oriented test was conducted between 9 a.m. and 12:45 p.m. Trial one was for one-half hour; after a short break, trial two lasted another one-half hour, followed by another short break and a regular two and one-half hour work period during which workers did not know they were being observed. The point was to measure the correlation between the trial periods and the regular work period. Workers were informed that this was an experiment; participation was voluntary. More than 100 workers participated. The beginning and end of each trial period were signaled by a shot from a starting gun; workers recorded their names and the

²¹An applicant for a ranchhand position claimed to know how to handle horses, mend fences, and have other skills related to the job. He was hired on his self-proclaimed proficiencies. His lack of these skills became readily apparent on the job. When asked about the discrepancy between his claims and abilities, he replied, "I needed a job so badly, I would have said I could fly a plane—if that's what was needed."

number of chips collected on a card at the end of each trial. During their regular work period workers did not know their performance was being tested—until the final 15 minutes when again their names and number of chips collected were recorded on cards.

Results

The test-retest correlation coefficient between trial one and trial two on 97 pairs of observations was $r=0.73$, indicating statistical reliability. From 65 pairs of useable observations, the validity coefficient between trial one results and the regular work period was $r=0.44$; between trial two and the regular work period, $r=0.57$; and between trials one plus trial two and the regular period, $r=0.55$. These results indicate a strong positive correlation between the tests and actual work, indicating validity of the tests.

The range of chips collected during trials one and two was from three to 12; the range for the regular work period was from eight to 41. Thus, again, substantial differences between workers were observed, showing the potential value of using testing procedures when hiring.

Limitations

There were some problems with the tomato-picking test. For one thing, precision control on starting-stopping times was difficult when so many workers were involved. Worker self-recording also afforded less than desirable accuracy. These limitations were corrected on the next test, using vineyard pruners.

A Criterion-Oriented Strategy:

Testing of Vineyard Pruners

The purpose of this portion of the study was to determine if a work-sample test—when workers knew they were being tested—can be used to predict on-the-job performance of piece-rate paid crew vineyard pruners. The test was conducted on three San Joaquin Valley, California, farms, selected because they (1) paid on a piece-rate basis, (2) employed 40 or more workers, and (3) had previously cooperated with the author in other research efforts. Both concurrent and predictive studies were done. Recall that a *concurrent* study uses incumbent workers and a *predictive* study uses job applicants. There were about 115 workers who participated in a concurrent study on Farm 1 and 45 in a predictive study on Farm 2. Farm 3 as divided into two groups: 116 applicants participated in a predictive study on Farm 3A; 67 workers on Farm 3B were in a concurrent test. Workers received instructions in Spanish and/or English.

Although workers' pay is directly proportional to the number of vines pruned, quality of production depends on supervisors' requirements. Only grapevines that are cordon pruned were included (e.g., French Colombard, Chenin blanc, Barbera).²² Other viticultural conditions, e.g., vine age, vine vigor, spacing between rows, spacing within the row, missing plants, grafting, and vine variety, were consistent for a given farm (1) *within* the predictor and (2) *within* the criterion but, not necessarily consistent *between* (1) and (2). Inconsistencies in viticultural conditions *between* predictor and criterion—other than pruning

²²Cordon pruning is defined here as a bilateral arm-pruning system (as contrasted with the more unusual quadrilateral pruning).

method—give the study greater possible external validity; whereas inconsistencies between predictors or between criteria would reduce reliability. Data were collected during the 1986-87 winter season when the vines are in the dormant stage.

Predictor Measure

Predictor data were collected on the three farms from two work-sample pruning periods of 46 minutes each—Test 1 and Test 2—during which workers *knew* they were being tested and that they needed to prune as fast as possible and still maintain quality.²³

Predictor correlation coefficients were positive and high (see Table 4), ranging from $r = 0.79$ to 0.86 . These high coefficients mean that workers performed consistently between Test 1 and Test 2. There was a large range of scores within each group. For example, in the predictive study on Farm 3, Test 2, workers pruned within a range of 3 to 24 vines; in the concurrent study on Farm 1, Test 1, the range was 12 to 28 vines.

Criterion Measure

Criterion data were obtained from each farm's payroll records on two randomly selected days (Criterion 1 and Criterion 2, respectively) on two randomly selected grape varieties. These criteria measures were taken after the pruning season and thus were free from the Hawthorne effect.

Correlation coefficients were taken between Criteria 1 and 2 and are also reported in Table 4. Except for Farm 2, the criteria reliabilities were large

and positive. On Farm 2 there were only 16 data pairs available to use, i.e., there were only 16 of those who had taken the predictor test who were employed and available for the criterion check. Also, and probably more important, Farm 2 manager was apparently not careful about documenting exact working hours since his workers were being paid by the piece. In contrast, the other farms were very careful to document exact starting and finishing times for workers. Whether or not partially finished vines were counted added to the discrepancies in scores from day to day on Farm 2.

Validity

Four validity coefficients were measured by correlating Test 1 and Test 2 against both Criterion 1 and Criterion 2.²⁴ Results appear in Table 5. These correlation coefficients between predictors and criteria ranged from -0.13 on Farm 2 between Test 2 and Criterion 2 to 0.73 on Farm 1 between Test 1 and Criterion 1. Farm 2 shows low correlation between test results and the criteria, while the other three groups show significant relationships.²⁵

The Criterion 2 results for Farm 3B (concurrent study) also showed low correlation. However, validities for individual crews on Farm 3B were not as low as the farm-wide results, ranging from 0.46 to 0.63 . (For crew level reliability and validity coefficients, see the Appendix B.) A possible explanation for the difference between farm- and crew-level results is that Criterion 2 involved vineyard blocks of differing levels of pruning difficulty.

²³Pruning quality pre-tests were conducted in the predictive studies by the farm managers. (Farm 3 used the pruning quality data collection instrument in Figure 1.) But quality results are not reported here; no statistical significance was found between speed and quality.

²⁴Rater reliability was established on one farm, using 24 data-point pairs.

²⁵This finding of no statistical significance for Farm 2 results corroborates the notion that very unreliable criterion measures (as reported above for Farm 2) would make a test—no matter how reliable—invalid.

Table 4. Farm-level Predictor and Criterion Reliabilities for Vineyard Pruners

Study	farm 1 Concurrent	farm 2 Predictive	farm 3A Predictive	farm 3B Concurrent
<i>Predictor</i>				
Correlation number	0.86 (111)	0.84 (43)	0.84 (105)	0.79 (52)
Test 1, mean standard dev.	20.48 (3.36)	13.96 (4.42)	14.35 (3.46)	21.83 (5.46)
Test 2, mean standard dev.	21.21 (3.75)	14.52 (4.20)	14.50 (3.49)	22.04 (5.93)
<i>Criterion</i>				
Correlation number	0.76 (106)	-0.44 (16)	0.51 (20)	0.57 (44)
Crit. 1, mean standard dev.	27.46 (4.83)	34.73 (5.78)	30.48 (7.29)	30.96 (6.68)
Crit. 2, mean standard dev.	32.59 (6.47)	22.68 (5.70)	31.12 (8.01)	28.77 (7.19)

Table 5. Farm-Level Validity Results for Vineyard Pruners

Study	farm 1 Concurrent	farm 2 Predictive	farm 3A Predictive	farm 3B Concurrent
<i>Test 1</i>				
Criterion 1 number	0.73 (110)	0.35 (21)	0.41 (26)	0.60 (43)
Criterion 2 number	0.72 (108)	0.11 (18)	0.66 (20)	0.14 (45)
<i>Test 2</i>				
Criterion 1 number	0.67 (108)	0.23 (20)	0.52 (27)	0.59 (47)
Criterion 2 number	0.61 (106)	-0.13 (17)	0.67 (21)	0.31 (47)

Conclusions

There is little doubt that both concurrent and predictive type tests can predict performance. It also seems certain that employers cannot assume that a test will always work, for Farm 2's test was not valid.

No conclusions can be drawn from this study about the relative effectiveness of concurrent and predictive tests. Traditionally, restriction of range is a greater problem for concurrent than for predictive studies, but in this agricultural setting, the opposite was true. In both predictive tests, most persons tested were not employed and were no longer available for the criterion test, leading to restriction of range problems. And concurrent studies in agriculture may have less problem with restriction of range than nonagricultural studies, for in agriculture there is less room for upward mobility by good workers, while those who do not perform well may be kept on the payroll (for example, when they are part of an employed family group).

CONCLUDING COMMENTS

Testing applicants before employment in agriculture may become a more prevalent practice in the future. Farm workers' skills vary substantially. The new immigration law may mean a reduced labor supply and higher wages. The casual nature of the agricultural work force may change in the years to come. Farmers are becoming more concerned about labor management decisions, organizational structure, supervision, personnel policies, job analysis, wage structure, incentive pay, performance evaluation, discipline, and farm safety. Agricultural testing programs are complementary to these changing conditions.

Tests—along with other selection tools—can bring out the differences in applicant abilities for specific jobs. In the long run, a better selection process can help farmers hire workers who will be more productive, have fewer absences, have fewer accidents, and stay longer with the organization.

Data for worker productivity in tomato picking and vineyard pruning show that some workers consistently outperform others in the same crew. Often, the better workers can perform twice as well as the worst within a given crew or work group; sometimes individual workers are even four to five times as productive as others. If managers can hire more productive workers they will probably need fewer workers. Interviews, reference checks, applications, and resumes alone often do not bring out these differences.

Some applicants self-select themselves out of the running when they feel they are not qualified. But other applicants will go through the process and try to get the job no matter how unqualified they are. Therefore, when looking towards improving the productivity and viability of agriculture, testing (as well as a better overall selection process) has much to contribute to the farming economy.

APPENDIX A:

Sample Costs of Testing for a Secretarial Position at a Farming Operation

If 40 applicants apply for a secretarial position at a farming operation and if the employer's wage costs for one year for this position (including benefits) are \$21,120, and if the secretary spends 50 percent of the time typing, the following figures apply:

Training and preparation for testing staff (2 persons at \$160 each)	\$320
Manager cost (\$160/day for 6 days)	960
(Manager cost includes time for selecting the instrument, planning, ordering test, and scoring it for all applicants.)	
Staff cost for test administration (40 tests, 1 hour/person at \$10/hour)	400
(Staff cost includes time to give and score the test for all applicants.)	
Test cost (2 sets, \$32 per 25 tests, + \$36 for manual, keys, and practice copies)	<u>100</u>
Total cost for administering a typing test	\$1780

Since the average time spent typing at this job is 50 percent, half the person's wages, or \$10,560, is for typing. If testing resulted in selecting a person whose typing speed was 60 words per minute (wpm) rather than 40 wpm, testing would have increased efficiency by 50 percent. (That is, an untested secretary on average would be 33 percent less efficient.) If the average turnover for a secretary is four years, then the total savings for this farmer would be \$12,160:

\$10,560 - \$1780 (cost of testing)	\$8,780
\$10,560 - 0.33 • \$10,560 (cost of not testing)	<u>-\$7075</u>
<u>Savings for first year:</u>	\$1,705
\$10,560 • 4 years - \$1705 (cost of testing)	\$40,460
\$7,075 • 4 years (cost of not testing)	<u>-\$28,300</u>
<u>Savings over 4 years</u>	\$12,160
<u>Average savings per year</u>	\$3,040

Savings could be even greater, for these figures refer only to typing speed and do not consider the time involved in correcting mistakes or re-typing. Additional savings might be enjoyed by also testing the applicant for skills used during the non-typing portion of the time.

Appendix Table B.1.

Farmwide and Crew Predictor Reliabilities

Farm:	1	2	3A	3B
	concurrent	predictive		concurrent
Crew A	0.81	0.88	0.85	0.85
(n)	(21)	(25)	(39)	(21)
Crew B	0.96	0.75	0.91	0.52
(n)	(17)	(18)	(43)	(12)
Crew C	0.91		0.95	0.88
(n)	(23)		(6)	(19)
Crew D	0.74		0.93 ^a	
(n)	(17)		(9)	
Crew E	0.74		0.57	
(n)	(19)		(17)	
Crew F	0.83			
(n)	(14)			
Farmwide	0.86	0.84	0.84	0.79
(n)	(111)	(43)	(105)	(52)
Crit. 1, mean	20.48	13.96	14.35	
21.83				
Standard dev.	(3.36)	(4.42)	(3.46)	
(5.46)				
Crit. 2, mean	21.21	14.52	14.50	
22.04				
Standard dev.	(3.75)	(4.20)	(3.49)	
(5.93)				

*This reliability not included in the summary or validity analysis because the test 1 period was not 46 minutes long.

Appendix Table B.3.
Farm 1 Validity Coefficients by Crew

Test:	1	2	1	2
Criterion:	1	2	1	2
Crew A	0.79	.88	.86	.83
(n)	(19)	(20)	(19)	(20)
Crew B	0.78	0.83	0.85	0.82
(n)	(17)	(17)	(17)	(17)
Crew C	0.80	0.73	0.69	0.76
(n)	(23)	(19)	(23)	(19)
Crew D	0.62	0.59	0.75	0.51
(n)	(16)	(16)	(16)	(16)
Crew E	0.66	0.67	0.59	0.40
(n)	(22)	(22)	(20)	(20)
Crew F	0.35	0.62	0.07	0.39
(n)	(13)	(14)	(13)	(14)
Farmwide	0.73	0.72	0.67	0.61
(n)	(110)	(108)	(108)	
(106)				

Appendix Table B.2.

Farmwide and Crew Criterion Reliabilities

Farm:	1	2	3A	3B
	concurrent	predictive		concurrent
Crew A	0.85			0.74
(n)	(18)			(19)
Crew B	0.82			0.82
(n)	(17)			(11)
Crew C	0.75			0.73
(n)	(19)			(14)
Crew D	0.65			
(n)	(16)			
Crew E	0.92			
(n)	(23)			
Crew F	0.75			
(n)	(13)			
Farmwide	0.76	-0.44	0.51	0.57
(n)	(106)	(16)	(20)	(44)
Crit. 1, mean	27.46	34.73	30.48	30.96
standard dev.	(4.83)	(5.78)	(7.29)	(6.68)
Crit. 2, mean	32.59	22.68	31.12	28.77
standard dev.	(6.47)	(5.70)	(8.01)	(7.19)

Appendix Table B.4.
Farm 3B (Concurrent) Validity Coefficients by Crew

Test	1	2	1	2
Criterion:	1	2	1	2
Crew A	0.85	0.630	.72	0.60
(n)	(18)	(19)	(17)	(18)
Crew B	0.36	0.53	0.51	0.86
(n)	(10)	(10)	(14)	(12)
Crew C	0.39	0.46	.54	0.54
(n)	(15)	(16)	(16)	(17)
Farmwide	0.60	0.14	0.59	0.31
(n)	(43)	(45)	(47)	(47)

SELECTED REFERENCES

- American Psychological Association. *Standards for Educational and Psychological Testing*, 1985.
- Anastasi, A. *Psychological Testing*, 5th ed. New York: Macmillan, 1982.
- Barrett, G. V., J. S. Phillips, and R. A. Alexander. "Concurrent and Predictive Validity Designs: A Critical Reanalysis." *Journal of Applied Psychology* 66(1981): 1-6.
- Bersoff, D. N. "Testing and the Law." *American Psychologist* 36(1981): 1047-1056.
- Billikopf, G. E. "Response to Incentive Pay Among Vineyard Workers." *California Agriculture* 39(1985): 13-14.
- Billikopf, G. E. "How to Fire Without Getting Burnt." *California Farmer* (January 1987): 24-25E.
- Billikopf, G. E. "Testing to Predict Tomato Harvest Worker Performance." *California Agriculture*, 41(1987): 16-17.
- California Compensation Case*, Vol. 37(1972): 185-194.
- Chronbach, L. J. and G. C. Gleser. *Psychological Tests and Personnel Decisions*, 2nd ed. University of Illinois Press, 1965.
- Daniel, C. "Science, System, or Hunch: Alternative Approaches to Improving Employee Selection." *Personnel Management* 15(1986): 1-10.
- Division of Industrial and Organizational Psychology. *Principles for the Validation and Use of Personnel Selection Procedures*, 2nd ed. American Psychological Association, Berkeley, CA, 1980.
- Doverspike, D., G. V. Barrett, and R. A. Alexander. "The Feasibility of Traditional Validation Procedures for Demonstrating Job Relatedness." *Law and Psychology Review* 9(1985): 35-44. (From *Psychological Abstracts* 73(1986), Abstract No. 10587).
- Downs, S., R. M. Farr, and L. Colbeck. "Self Appraisal: A Convergence of Selection and Guidance." *Journal of Occupational Psychology* 51(1978): 271-278.
- Ebel, R. L. "Comments on Some Problems of Employment Testing." *Personnel Psychology* 30(1977): 55-63.
- Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor and Department of Justice. "Uniform Guidelines on Employee Selection Procedures." *Federal Register* 43(1978): 38290-38315.
- Equal Employment Opportunity Commission, Office of Personnel Management, Department of Justice, Department of Labor and Department of the Treasury. "Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures." *Federal Register* 44(1979): 11996-12009.
- Equal Employment Opportunity Commission, Office of Personnel Management, Department of Justice, Department of the Treasury, and Department of Labor, Office of Federal Contract Compliance Programs. "Adoption of Additional Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures." *Federal Register* 45(1980): 29530-29531.
- Farr, J. L., B. S. O'Leary, and C. J. Bartlett. "Effect of Work Sample Test Upon Self-selection and Turnover of Job Applicants." *Journal of Applied Psychology* 58(1973): 283-285.
- Ghiselli, E. E. *The Validity of Occupational Aptitude Tests*. New York: John Wiley and Sons, 1966.
- Ghiselli, E. E. "The Validity of Aptitude Tests in Personnel Selection." *Personnel Psychology* 26(1973): 461-477.
- Green, B. F. "A Primer of Testing." *American Psychologist* 36(1981): 1001-1011.
- Guion, R. M. and C. J. Cranny. "A Note on Concurrent and Predictive Validity Designs: A Critical Reanalysis." *Journal of Applied Psychology* 67(1982): 239-244.
- Kleiman, L. S., and R. H. Faley. "The Implications of Professional and Legal Guidelines for Court Decisions Involving Criterion-Related Validity: A Review and Analysis." *Personnel Psychology* 38(1985): 803-833.
- Lee, R., K. J. Miller, and W. K. Graham. "Corrections for Restriction of Range and Attenuation in Criterion Related Validation Studies." *Journal of Applied Psychology* 67(1982): 637-639.
- Leedy, D. L. *Practical Research: Planning and Design*, 3rd ed. New York: Macmillan, 1985.
- Little, T. M. and F. J. Hills. *Agricultural Experimentation: Design and Analysis*. New York: John Wiley and Sons, 1978.
- Lukacsko, Z. "Mezogazdasagi Repuloge Pvezetok Kivalasztasi Rendszerenek Kezdeti Eredmenyei." (Initial results on the selection of agricultural air pilots.) *Magyar Pszichologiai Szemle* 41(1984): 129-139. (From *Psychological Abstracts* 72(1985), Abstract No. 16033.)
- McClain, M. E. *Employment Termination Law: A Practical Guide for Employers*. CEB, California Education of the Bar, 1987.
- Mount, M. K., P. M. Muchinsky, and L. M. Hanser. "The Predictive Validity of a Work Sample: A Laboratory Study." *Personnel Psychology* 30(1977): 637-645.

- O'Leary, L. R. "Fair Employment, Sound Psychometric Practice, and Reality: A Dilemma and a Partial Solution." *American Psychologist* 28(1973): 147-150.
- Ramos, R. E. "Employment Battery Performance of Hispanic Applicants as a Function of English or Spanish Test Instructions." *Journal of Applied Psychology* 66(1981): 291-295.
- Ramsay, R. T. *Management's Guide to Effective Employment Testing: What's Legal, Valid and Fair*. Chicago: Dartnell, 1981.
- Robertson, I. T. and R. S. Kandola. "Work Sample Tests: Validity, Adverse Impact and Applicant Reaction." *Journal of Occupational Psychology* 55(1982): 171-183.
- Tenopyr, M. L. "The Realities of Employment Testing." *American Psychologist* 36(1981): 1120-1127.
- Schlei, B. L. and P. Grossman. *Employment Discrimination Law*, 2nd ed. The Bureau of National Affairs, 1983.
- Schmidt, F. L., A. C. Greenthal, J. E. Hunter, J. G. Berner, and F. W. Seaton. "Job Samples vs. Paper and Pencil Trades and Technical Tests: Adverse Impact and Examinee Attitudes." *Personnel Psychology* 30(1977): 187-197.
- Schmidt, F. L. and J. E. Hunter. "The Future of Criterion-Related Validity." *Personnel Psychology* 33(1980): 41-60.
- Schmitt, N, R. Z. Gooding, R. A. Noe, and M. Kirsch. "Meta-analyses of Validity Studies Published Between 1964 and 1982 and the Investigation of Study Characteristics." *Personnel Psychology* 37(1984): 407-422.
- Schultz, C. B. "Saving Millions Through Judicious Selection of Employees." Special Issue: Techniques and Challenges. *Public Personnel Management* 13(1984): 409-415.
- Siegel, J. *Personnel Testing Under EEO*. New York: Amacom, 1980.
- Wakefield, J. A. and N. A. Goad. *Psychological Differences: Causes, Consequences, and Uses in Education and Guidance*. San Diego: Edits, 1982.
- Wernimont, P. F. and J. P. Campbell. "Signs, Samples, and Criteria." *Journal of Applied Psychology* 52(1968): 372-376.
- Whelchel, B. D. "Use of Performance Tests to Select Craft Apprentices." *Personnel Journal* 64(1985): 65-69.

In accordance with applicable State and Federal laws and University policy, the University of California does not discriminate in any of its policies, procedures, or practices on the basis of race, color, national origin, religion, sex, sexual orientation, handicap, age, veterans status, medical condition (as defined in Section 12926 of the California Government Code), ancestry, or marital status; nor does the University discriminate on the basis of citizenship, within the limits imposed by law or University policy. In conformance with applicable law and University policy, the University of California is an affirmative action/equal opportunity employer. Inquiries regarding the University's equal opportunity policies may be directed to the Vice Chancellor of Academic Affairs—Affirmative Action Officer and Title IX Coordinator, 521 Mrak Hall, (916) 752-2070. Speech and hearing impaired persons may dial 752-7320 (TDD).

