

JOB-SAMPLE TEST AS A PREDICTOR OF
ON-THE-JOB PERFORMANCE OF VINEYARD PRUNERS

Gregory Encina Billikopf

1987

JOB-SAMPLE TEST AS A PREDICTOR OF
ON-THE-JOB PERFORMANCE OF VINEYARD PRUNERS

A Thesis Presented to the Faculty

of

California State University, Stanislaus

In Partial Fulfillment

Of the Requirements for the Degree

Master of Arts with Special Major in Human Resource Management

By

Gregory Encina Billikopf

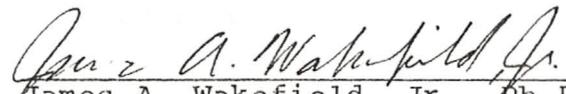
May, 1987

Certification of Approval

I certify that I have read JOB-SAMPLE TEST AS A PREDICTOR OF ON-THE-JOB PERFORMANCE OF VINEYARD PRUNERS by Gregory Encina Billikopf, and that in my opinion this work meets the criteria for approving a thesis submitted in partial fulfillment of requirements for the Master of Arts degree, with Special Major in Human Resource Management, in the Department of Business at California State University, Stanislaus.



David M. Hunt, Ph.D., Chair
Professor of Management



James A. Wakefield, Jr., Ph.D.
Professor of Psychology



A. K. (Gene) Murti, Ph.D.
Professor of Management

Acknowledgements

Buy the truth,
and sell it not;
also wisdom,
and instruction,
and understanding.

(Proverbs 23:23)

The author wishes to express appreciation to the members of his thesis committee, Dr. David M. Hunt, Chairman; Dr. James A. Wakefield, Jr.; and Dr. Gene Murti, for their support and encouragement on this manuscript. Thanks, also, to Dr. Hunt, Dr. Duane Dove, and Professor William Torrens who served as chairmen of my Special Master's program in Human Resource Management, with the assistance of Dr. Wakefield and Dr. Murti.

Special thanks go to the farm managers, foremen, and staff, as well as the approximately three hundred farm workers who participated, without which the study would not have been possible.

The author wishes to thank Dr. John Mamer and Dr. Howard Rosenberg, as well as many others from the University of California Agricultural Extension for encouragement and support. I would also like to thank Dr. Charles C. Hanna,

Table of Contents

Graduate Dean, Ms. Judith Graves of the graduate office, and Ms. Paula Crawford and Mr. J. Carlyle Parker of the library staff for the special help provided. The author also wishes to thank his wife, Linda, children, and parents, David M. Billikopf and Maria Luisa Encina de Billikopf.

To every one of my instructors at California State University, Stanislaus, special thanks for their dedication in teaching and for the information that made this thesis possible: Dr. Dove, Dr. Hendricks, Dr. Hunt, Professor Torrens, Dr. Schoenthaler, Professor J. Shobe, and Dr. Wakefield.

I thank the authors who wrote the excellent texts and journal articles I had the privilege of reading; people such as Anastasi, Chronbach, Ghiselli, and so many others. Finally, I would like to especially thank two instructors that were so effective, so excellent, so exciting, that I spent many sleepless nights thinking of what I had learned from them: Dr. James Wakefield and Dr. Judith J. Hendricks.

Selected Bibliography

Appendices

- A. Bivariate Distribution of scores from 1954-55
- B. Pruning quality Data Collection Instrument
- C. Pruning speed Data Collection Instrument

Table of Contents

Certification of Approval ii

Acknowledgements iii

Table of Contents v

List of Tables vi

Abstract vii

Problem Statement and Significance 1

Literature Review 4

Methodology 16

 Setting 16

 Research Design 18

Results 28

 Rater Reliability 28

 Predictor 28

 Criterion 34

 Validities 39

Conclusions and Suggestions for Future Research . . . 44

Style Manual 47

Selected Bibliography 48

Appendices

 A. Bivariate Distribution of Scores (Farm 1) . . 54

 B. Pruning Quality Data Collection Instrument . 55

 C. Pruning Speed Data Collection Instrument . . 56

List of Tables

1. Farm-wide and Crew Predictor Reliabilities 30

2. Intercorrelations Among Predictive and Concurrent Scores for Farm 3 Subjects 33

3. Farm-wide and Crew Criterion Reliabilities 35

4. Criterion Grape Varieties 39

5. Farm-wide Validity Results 40

6. Farm 1 Validity Coefficients by Crew 42

7. Farm 3B (Concurrent) Validity Coefficients by Crew 43

Abstract

The ability of a performance test to predict on-the-job behavior for piece-rate paid crew workers was investigated. Within an agricultural context, specifically grape pruning, workers were tested under predictive and concurrent criterion-oriented strategies. In a predictive study applicants are tested and test results are later correlated with on-the-job performance. In a concurrent study, test results and on-the-job performance of incumbents are correlated. Validity coefficients (correlations between test results and on-the-job performance) as well as predictor and criterion reliabilities were obtained. The predictor variable was a short pruning test (forty-six minutes) and the criterion variable was pruning speed while on the job. The predictor was measured under open conditions (workers knew they were being tested); the criterion under double-blind conditions (neither researcher nor workers knew when the criterion measure was taken). Four studies involving three farms took place. Three of the four studies resulted in significant validity coefficients (see table on p. viii). The remaining farm, Farm 2, produced non-significant validity results. Farm 2 was the only farm with a non-significant criterion reliability, and it was suggested in the literature

review that very low criterion reliabilities would result in non-valid coefficients. The study, with the small number of farms involved, was not designed to test the differences between predictive and concurrent type tests. Both predictive and concurrent studies showed that tests could be used to anticipate performance in vineyard pruners who are paid on a piece-rate basis. Factors that contributed to the significant validity coefficients included predictor and criterion consistencies (the use of piece rate pay probably contributed to the latter). Nevertheless, individual farmers would do well to conduct their own studies rather than assume that the test will always work. It seems clear that no matter how effective and reliable the test, low or non-existent criterion reliabilities can result in non-valid tests.

Average Farm-wide Validity Coefficients

Crews	Farm			
	1	2	3A	3B
Study	Concurrent	Predictive	Predictive	Concurrent
r (r ²)	.68***(.47)	.14 (.02)	.57**(.32)	.41**(.17)
(n)	(108)	(19)	(23)	(45)

p < .01. *p < .001.

Three farms in the San Joaquin Valley were involved.
The predictor measure consisted of a forty-six minute, open
(i.e., workers knew they were being tested) job-sample
performance test. The criterion was a job-sample performance
measure taken over a sample day, using double-blind
conditions. Neither researcher nor workers knew when
the

**Job-sample Test as a Predictor of
On-the-job Performance of Vineyard Pruners**

Problem Statement and Significance

Selection tests are not new. More than a millenium
B.C., Gideon selected for battle those warriors who brought
water to their mouth and lapped it as dogs do, over those who
bowed down to drink (Judges vii.1-7). The criteria for
selecting many employees today, especially in agriculture,
seems even less discriminating than Gideon's test. Pruners--
and most farm workers--are usually selected on a first-come-
first-hired basis.

(de) The principal objective of this study is to determine if
one can predict work performance of piece-rate paid vineyard
pruners using a concurrent validation strategy. A secondary
objective is to determine the predictive ability of a
predictive-type test. A concurrent test involves correlating
incumbents' performance on a test with on-the-job
performance. A predictive test involves correlating
applicants' test scores with on-the-job performance. Two
sub-problems are to establish predictor and criterion
reliabilities.

Three farms in the San Joaquin Valley were involved. The predictor measure consisted of a forty-six minute, open (i.e., workers knew they were being tested) job-sample performance test. The criterion was a job-sample performance measure, taken over a sample day, using double-blind conditions (i.e., neither researcher nor workers knew when the criterion measure was taken).

Piece-rate paid pruners have been shown to greatly differ in skills and pruning speed. A test that improves on chance hiring can help farmers reduce pruning costs by selecting fewer, more productive workers. It is possible for a test with high validity (e.g., $r = .70$) to greatly decrease the percentage of job offers to unqualified persons from 40% if all applicants were hired, to 7% or fewer (depending on the position of the test cutoff score) with the use of the test (Anastasi, 1982). More moderate gains can be achieved with average validity coefficients.

Given today's legal climate there is an important place for effective employee selection; firing workers is becoming increasingly difficult. Both anti-discrimination laws and wrongful discharge litigation are making employers more careful when they fire workers.

Work-sample tests have high acceptance by industrial psychologists, applicants and judges, making such a test--if

Literature review

Testing is not new nor is it static. Around the turn of the century, the modern testing movement was launched by the likes of Francis Galton, James McKeen Cattell, and Alfred Binet (Anastasi, 1982). Some of the major developments in testing are recent, however, and got much of their thrust during the two world wars. In the introduction to her fifth edition of Psychological Testing Anastasi describes the fluidity of testing: "... psychological testing today does not stand still long enough to have its picture taken" (p. v).

This literature review includes (1) an overview of personnel selection and validation; (2) a brief discussion of pertinent legal issues; (3) the theories and evidence which support each hypothesis (views for and against testing are provided); (4) potential contributions of this study; and (5) ideas that can be incorporated into research efforts.

Personnel Testing and Validation

Employment tests have a potential to contribute to the economy as well as to individual firms by directly improving average productivity (Schultz, 1984). Schultz showed, for example, how a selection test could result in savings in excess of \$5,000 per-worker-year when (1) test had a validity

coefficient of .5; (2) the selection ratio was .1 (one in ten applicants hired); and (3) the standard deviation in the value of a worker's production each year was \$6,000.

Validity is one of the central points of employment tests. While sometimes one hears of different types of validities (content, construct, criterion), these are merely validation strategies rather than distinct types of validities (Principles, 1980, and Standards, 1985). Criterion-oriented and content-oriented strategies are of particular relevance to this paper. The latter will be discussed in conjunction with legal issues.

Within criterion-related studies there are two approaches, predictive and current. In a predictive study applicants are tested and test results are later compared to some aspect of the job (e.g., performance, absenteeism, turnover, etc.). The concurrent strategy involves testing present workers and at the same time comparing test results to a job measure(s).

Both of these criterion-oriented strategies involve establishment of a correlation coefficient between a predictor (test) and a criterion (job performance). This is usually done with Pearson's "r" (Ghiselli, 1966). If the relationship is not linear then prediction would be much more accurate at one range than at another and would result in an

underestimation of validity (Ghiselli).

The traditional view is that a predictor strategy is superior to a concurrent one for most personnel situations. While the traditional view is generally defended by many (Anastasi, 1982, Ghiselli, 1973, and Standards, 1985), perhaps Guion and Cranny (1982) are the traditional view's most eloquent defenders. The argument often presented is that concurrent validation strategies introduce a larger restriction of range error with corresponding lower validity coefficients for concurrent validities. Many feel that employed workers in a particular job represent a more homogeneous work force than an applicant population, which, in turn, results in a restriction of range and lower validity coefficients.

Schmitt, Gooding, Noe & Kirsch (1984) hold the opposite view; they found higher validity coefficients for concurrent than predictive studies. Schmitt et al. were not able to control for any variables in their metaanalysis, however. Nevertheless, the differences between concurrent and predictive studies have increasingly been minimized by others (Barret, Phillips & Alexander, 1981, and Principles, 1980) especially in the light of a host of corrections that have been developed for restriction of range (e.g., Lee, Miller & Graham, 1982).

Legal Issues and Testing: Validity and Wrongful Discharge

While tests can be misused, tests are often a superior tool for selecting employees without illegal discrimination, and are an improvement over more subjective methods (Barrett et al., 1985, Daniel, 1986, Doverspike, Barret & Alexander, 1985, O'Leary, 1973, Tenopyr, 1981, Whelchel, 1985). More subjective methods (e.g., interview) have not been subject to the same legal problems than objective tests have (Daniel), yet the Guidelines (1978) state that interviews are regulated just as are other types of employee selection tools.

Developing 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 (1) when they were hired, (2) in conversations with foremen or supervisors, or (3) in employee handbooks have given rise to much litigation. These include such terms or phrases as "permanent employee," and "as long as you do good work you will have a job." Some employers who have discharged a "permanent" employee have ended up with a "wrongful discharge" suit. They have been charged with breaking an implied contract of good faith (Billikopf, 1987). In addition to avoiding problems associated with employee discharge, effective employee selection offers legal and

management benefits. Benefits of work-sample tests. Work-sample tests lend themselves to either content or criterion validation or alternative strategies (Robertson & Kandola, 1982, Mount, Muchinsky & Hanser, 1977, Schmidt, Greenthal, Hunter, Berner & Seaton, 1977). Better yet, the Standards (1985) encourage use of multiple sources of evidence to examine the validity of inferences about a test. According to Kleiman and Faley (1985) stringent job-analysis justification is not even required for production criterion.

Job-sample tests also increase the "face validity" (what it seems the test is about) of employment tests (Wernimont & Campbell, 1968). 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 feelings towards the test (Schmidt et al., 1977).

In some cases job-sample tests have reduced adverse impact in employment decisions (Robertson & Kandola, 1982, Schmidt et al., 1977, Whelchel, 1985). The validity of such decisions, unfortunately, has seldom been tested empirically against measures of job performance.

Job sample tests also promote self-selection (Downs, Farr & Colbeck, 1978, Farr, O'Leary & Bartlett, 1973, Robertson & Kandola, 1982). Farr et al., however, found that

self-selection of candidates depended on race. Downs et al. established an inverse correlation between applicant performance (in a work-sample test) and refusal to accept the job.

A final comment on employee selection and legal issues is that of utility. The Guidelines (1978) allow for adverse impact if there is proof of test validity. However, the greater the adverse impact the greater the proof for high validity and/or high utility that is required. Formulas for test utility and utility considerations can be found in Chronbach & Gleser (1965), Hunter & Schmidt (1983), Schmidt & Hunter (1980), and Schultz (1984).

A full discussion of the legal status of employee testing is beyond the scope of this paper. 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 books can be found (Siegel, 1980, and Ramsay, 1981). Further discussions can be seen in Kleiman & Faley (1985) and Bersoff (1981).

The Hypothesis and the Null Hypothesis

Hypothesis. A job-sample test can be used as a predictor of on-the-job performance for a crew of piece-rate paid vineyard pruners.

Argument for testing. Mount et al. (1977) have found that job-sample tests have high reliabilities. High test and high job performance reliabilities should result in significant validity coefficients, too.

Wernimont and Campbell (1968) and Ghiselli (1966) point out that few studies have established the reliability of the criterion measure. Billikopf (1985a, 1985b), however, has established the high reliability of more than a dozen crews of vineyard workers when paid on a piece-rate basis. Grape pruners in each crew performed at quite different rates, and perhaps more important, they did so consistently. Similar worker differences were also reported by Schultz (1984).

Similarity between work-sample tests and on-the-job performance should result in significant validity coefficients, too. In grape pruning, both test and criterion can be measured in terms of pruning speed (or rate). Workers' motivation to perform is probably not powered by the same influences (1) during a selection test (where a job is on the balance); and (2) when extra pay is given for the extra production. Nevertheless, Billikopf (in press) points out that a worker who can do half as well as another when trying his/her best under test conditions is unlikely, no matter what the motivation, to be able to catch up to the faster worker.

The high reliability of both the predictor and criterion, as well as the similarity of the tasks involved (grapevine pruning) shows potential promise as would be shown by a statistically significant validity coefficient.

Null hypothesis. A job-sample test cannot be used as a predictor of on-the-job performance for a crew of piece-rate paid vineyard pruners.

Argument against testing. Researchers see a great future in criterion-oriented validity (e.g.: see Schmidt & Hunter, 1980). However, Robertson and Kandola (1982), and Wernimont & Campbell (1968) found many researchers confuse validity with reliability. For instance Lee et al. (1982) and Mount et al. (1977) feel that having a different predictor and criterion measure is what distinguishes a validity from a reliability coefficient.

Mount et al. used an open job-sample test as a predictor and a different but more complex, still open, job-sample test as the criterion. Such a study ignores worker motivation on the job. 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).

Further, in concurrent-oriented studies, workers do not have the same motivation to do well in a test than in a

predictive study (Guion & Cranny, 1982; Principles, 1980). Guion & Cranny feel that these differences in motivation introduce random error in the predictor. Concurrent studies have restriction of range problems which often result in a non-significant validity coefficient (Guion & Cranny, 1982). This traditional view, however, is being questioned by others (see discussion on predictive vs. concurrent studies, above).

Many assume that piece-rate paid workers are motivated to do their best during work. Paying crew members a piece-rate wage does not guarantee expression of individual differences, however. Billikopf (1985a) found that in certain circumstances workers under piece-rate will work no faster than an agreed-upon pace, or bogey. The forces that induce workers to do their individual best are particularly subordinate to group cohesiveness when workers perform in a crew (see Billikopf, for a discussion on bogeys in agriculture). While uniform working speed is especially evident when workers are paid on an hourly basis, uniform speeds have also been found when paid on a piece-rate basis (Billikopf).

Uniform working speeds result in an unreliable criterion measure. An excellent test is not a substitute for criterion unreliability (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). Results showing the null hypothesis would tend to confirm this view.

Importance and Contribution of the Study

Agriculture. The only agricultural-related references to testing were found in (1) Ghiselli (1966) who discussed high validity coefficients (.55) for an arm dexterity test for selection of fruit and vegetable graders; and (2) testing of agricultural pilots in Hungary (Lukacsko, 1984).

Although there is a danger in work-sample tests if work methods change (Robertson & Kandola, 1982) work-sample tests are often considered superior tests for employee selection (Mount et al., 1977, Schmitt et al., 1984, Whelchel, 1985).

A work-sample test provides for more "face validity," and could be validated both with content and criterion-related strategies, thus minimizing risk of litigation (e.g., O'Leary, 1973). Farmers could come closer to utilizing employment tests and increase the average performance of the workers they employ.

Personnel and industrial psychology. Besides the potential contributions to agricultural employee selection, there is an opportunity to do research that will also contribute something to the field of personnel management/industrial psychology.

Wernimont & Campbell (1968) point out that unfortunately few studies have established the reliability of the criterion measure. A study that uses two (or more) samples of the criterion measure--as well as two predictor measures--could contribute something to personnel psychology.

Schmitt et al. (1984) found few studies that used production as the criterion. Often performance ratings are used instead and result in lower validity coefficients because of the unpredictability of the criterion scores (Schmitt).

Daniel (1986) said that "the best opportunities to improve selection exist in organizations in which one or more readily identifiable, quantifiable characteristics affect organizational performance" (p. 6). Vine pruning is both identifiable, quantifiable, and--under most piece-rate paid situations--pay is totally dependent on such quantification.

Other considerations from literature

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 the study (Ghiselli, 1966, Green). Schmidt & Hunter (1980) have called for much greater numbers but Schmitt et al. (1984) did not find the same weakness in 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. Where appropriate, researchers should make use of these suggestions.

Setting

The research was conducted in two vineyards in California. The first vineyard was a 10-acre vineyard in the Central Valley and the second was a 20-acre vineyard in the Central Valley. The vineyards were selected because they were representative of the vineyard industry in California. The vineyards were selected because they were representative of the vineyard industry in California. The vineyards were selected because they were representative of the vineyard industry in California.

The analysis was conducted

Methodology

This was a criterion-oriented (concurrent and predictive types) test validation study where predictive ability of a work-sample test (predictor) under open conditions was correlated against double-blind measures of on-the-job performance (criterion) of vineyard workers. The principal goal was to establish the extent of correlation between predictor and criterion (validity coefficient). Two subproblems were to establish measures of (1) predictor reliability and (2) criterion reliability.

Setting

Location. The principal study was conducted on a San Joaquin Valley vineyard (Farm 1). Additional data was collected from two other vineyards (Farm 2 and Farm 3) in case the principal farm involved did not follow through with the study. These additional farms are also located in the San Joaquin Valley. Farms were selected because they (1) have cooperated with the researcher in past studies, (2) employed large number of workers in the past, and (3) paid workers on a piece-rate basis.

Unit of analysis. The unit of analysis was grapevines

pruned. Worker productivity was measured in terms of number of vines pruned per worker during the predictor (test). Productivity during the criterion was measured in terms of number of vines pruned per hour per worker.

Sample. Approximate numbers of workers participating in the study were 115 for Farm 1; 45 for Farm 2; and 67 (concurrent study) and 116 (predictive study) for Farm 3. Pruners worked in crews of varying size (usually 15 to 40 pruners per crew). Name of each farm worker was obtained from the worker or from the farmer (or agent of the farmer) in order to identify and pair predictor and criterion data for individual workers and carry out regression analysis. Workers were paid on a piece-rate basis. While workers' pay is directly proportional to the number of vines pruned, quality of production is only as high--or low--as supervisors normally demand.

Grapevines included only those varieties that are cordon pruned (e.g., French Colombard, Chenin Blanc). Cordon pruning was defined as a bilateral arm pruning system (as compared to the more unusual quadrilateral cordon pruning). Other viticultural conditions were to be consistent for a given farm (a) within the predictor and (b) within the criterion but, (c) not necessarily consistent between both. Such conditions include vine age, vine vigor, spacing between

rows, spacing within the row, missing plants, grafting, and vine variety. Any inconsistencies in viticultural conditions between predictor and criterion--other than pruning method--work to give the study greater external validity while inconsistencies within reduce reliability of predictor and/or criterion. For the purposes of this study most vines within the same farm that have the same variety, age, and spacing were considered to be of sufficient similarity to constitute uniform viticultural conditions.

Duration. Data was collected during the 1986-1987 winter season. Grapes are deciduous and are pruned during the dormant stage. Winter was defined as December 1, 1986, to March 30, 1987.

Research Design

Principal statistical tool. Linear regression analysis was the principal statistical tool for this study. The weakness often attributed to regression analysis (e.g.: Leedy, 1985; Little & Hills, 1978) is that it is often used--improperly--to show a cause and effect relationship. The purpose of using regression analysis in this study was not to show causality, but rather, to show correlation or closeness of the relationship (Little & Hills). All the correlations in this study were established through Pearson's product-moment correlation "r" as follows (see Note 1):

Work-sample test procedure

Workers received instructions in Spanish and/or English to avoid language misunderstandings and to create some interest in the study. Work-sample consisted of cutting back of the vines and placing the clothes pin under the vine to the center of the row. Pruners were assigned to several rows. Each row was cut by the pruner.

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

During the predictor measure, workers started pruning at the first vine from a starting point. Workers were not to

Where:

- r = correlation coefficient
- x = one variable
- y = another variable
- n = number of pairs involved

Predictor measure. Predictor data was collected from work-sample pruning Test 1 and Test 2. Each of these tests lasted forty-six minutes. Through these tests two sets of data were collected. Total number of vines pruned per person, for each test, was computed by adding completely pruned vine totals plus a possible partially completed vine--rounded to the nearest 1/4 of a vine as estimated by the researcher.

Work-sample test procedure The clothes pin was placed at the furthest vine pruned (or partially pruned). The clothes pin was to be placed at the side of the vine furthest away from the pruned vines. This procedure concluded Test 1.

Workers received instructions in Spanish and/or English to avoid language misunderstandings and to create more interest in the study. Work-sample consisted of cutting wood off the vines and removing the brush from under the vine row to the center of the row. Pruners were assigned individual rows. Each row was matched to worker name or code for identification. During the predictor measure, workers started pruning at the first shot from a starting pistol. Workers were not to proceed from one vine to another without first clearing the brush from under the vine toward the center of the row. When the second shot was sounded, forty-six minutes after the first, workers were to clip a clothes pin by the furthest vine pruned (or partially pruned). The clothes pin was to be placed at the side of the vine furthest away from the pruned vines. This procedure concluded Test 1.

Immediately after placing the clothes pin workers skipped to the closest completely unpruned vine to the clothes pin. Starting with the new set of vines, workers continued to move brush from under each pruned vine before moving on to the next vine. When the third shot was sounded, 45 minutes after the second shot (90 minutes after the first shot), workers clipped a second clothes pin by the then

furthest vine pruned (or partially pruned). The clothes pin was placed at the side of the vine which was furthest away from the starting point. This procedure concluded Test 2. Workers were to immediately (a) stop working until they received further instructions, or (b) move to a pruning section outside of Test 1 and Test 2 pruning areas.

Workers in predictive tests on Farm 2 and Farm 3 were asked to carefully follow the procedures above. On Farm 3, on the concurrent study workers were allowed to prune vines on one side and then on the other, or to leave brush under the vine. On Farm 1, workers followed the procedures stated above but were allowed to partially clean the brush under the vines. Workers who arrived after the first gun shot were not used for Test 1. Foremen and the researchers reminded workers to follow instructions during the test, as needed. Ability to follow instructions was not part of the testing process.

Careful measurement of the predictor is important. Any factor that reduces predictor (or criterion) reliability introduces a source of error variance and is likely to reduce validity.

Rater reliability

One such factor that can limit predictor reliability is rater error. A sampling of twenty-four row sections

(completed plus partially completed vines) was used to establish rater reliability. Rater consistency was determined only on one farm. Reliability was calculated using Pearson's product-moment correlation "r" where:

- r = the reliability of the rater
- x = vine count of individual row as determined from first count
- y = vine count of same row as done for "x" counted a second time
- n = number of row pairs involved

Predictor reliability

Predictor reliability was established using data from all farm workers for whom there was test data for Test 1 and Test 2 within a particular farm. Predictor reliability was established through a test re-test reliability coefficient using Pearson's "r" where:

- r = the reliability of the work-sample tests
- x = individual worker raw score for Test 1
- y = individual worker raw score for Test 2
- n = number of workers completing Test 1 and Test 2

they were initially hired. After the test, all workers on Farm 2 (1) whose quality level was initially acceptable and (2) who pruned fast enough to make minimum wage were hired. On

Within crew predictor reliabilities

Where there was more than one crew in the concurrent studies (i.e., in Farm 1 and Farm 3), reliability coefficients were calculated for each crew to compare individual crew reliabilities against each other and against the average for all crews. Theoretically, crews with higher predictor reliability and higher criterion reliability should obtain higher validity coefficients. Nevertheless, this part of the study will only work as a pre-test for further research as little credence can be given to comparisons of coefficients among crews where workers are not randomly assigned to crews.

Predictor appropriateness

One limitation of using a concurrent-type test where workers already hold the job rather than test workers applying for a job is the assumption that workers will try and do their best on the test. Voluntary participation of workers already employed was used for concurrent studies on Farm 1 and Farm 3 to avoid having workers participate who might not be motivated to do their best. Workers on predictive studies on Farm 2 and Farm 3 were tested before they were formally hired. After the test all workers on Farm 2 (1) whose quality level was minimally acceptable and (2) who pruned fast enough to make minimum wage were hired. On

Farm 3, workers' performance was carefully evaluated and workers successfully completing quality and quantity tests were invited to come to work for the farm.

Criterion measure. Criterion data was obtained from each farm's payroll records. These records normally include pruning date, and for each worker, the number of hours worked as well as the number of vines pruned for that time period. The criterion measure was obtained by reading the appropriate column representing average (arithmetic mean) vines pruned per hour for each worker. If such a column did not exist, the mean vines per hour for each individual worker was determined as follows:

$$\bar{x} \text{ vines/hour for a worker} = \frac{\text{total vines pruned for the day by worker}}{\text{total hours worked for the day by worker}}$$

Criterion reliability

Criterion reliability data was collected by sampling from two randomly selected days per crew within the pruning season. These were labeled Criterion 1 and Criterion 2. Each Criterion represented a different randomly selected variety. Criterion 1 was generally the earliest (chronological) date of the two samples. Randomly selected days were chosen from Tuesdays, Wednesdays, and Thursdays to

avoid any beginning-of-the-week or end-of-the-week effects if any exist. In addition, no dates were picked for a complete working week after the predictor test was given, to avoid criterion contamination caused by possible excitement about the test itself. If a short day was picked (caused by rain, for instance) another day was selected instead until a day was selected where workers pruned for at least six hours on the average. The same two criterion days were selected for as many crews within a farm as possible.

Sample criterion reliability was established using a Pearson's "r" criterion reliability coefficient (where there was more than one crew on a farm, criterion reliability measures were determined for each crew) where:

r = the reliability of the criterion

x = individual worker mean pruning score for Criterion 1

y = individual worker mean pruning score for Criterion 2

n = number of workers for whom Criterion 1 and Criterion 2 data was available

Validity. Four validity coefficients were established by correlating each of Test 1 and Test 2 against each of Criterion 1 and Criterion 2 (Test 1 v. Criterion 1; Test 1 v. Criterion 2; Test 2 v. Criterion 1; Test 2 v. Criterion 2).

Each of these pairs was correlated through Pearson's "r"

where: r = validity coefficient

x = Test 1 or Test 2

y = Criterion 1 or Criterion 2

n = number of cases for which respective pairs of data were available

Perhaps more important than the "r" value for these four validity coefficients is " r^2 ". The squared validity coefficient indicates the percentage of variance shared by the two variables (Wakefield & Goad, 1982). For each of the four test and criterion combinations a " r^2 " value was calculated by squaring the validity coefficient (r).

Means of obtaining the data. Predictor data was obtained by the researcher during--and right after--the predictor test. The researcher met with managers of each of the three farms before data-collection and they agreed to provide the data and allow for the study to be conducted. The researcher obtained all the needed data from each of the participating farms.

Interpretation of the data. Data from each of the regression analyses was collected and statistical significance determined at the $p < .05$; $p < .01$; and $p < .001$. A finding of statistical significance will

regardless of test results. Borderline workers were given several days to improve. Farm 1 involved a concurrent study

Results

and a predictive one. In the predictive one, few workers were hired. The results obtained include a discussion of the (1) rater reliability; (2) the predictor; (3) the criterion; and (4) validity coefficients. Besides the hard data, some pertinent anecdotal information is also included.

Rater Reliability

Rater reliability was established on farm 2, with 24 data-point pairs. Reliability was very high ($r = .999$). Such a high reliability was not difficult to obtain, as it involved counting total number of plants as well as partial plants. Weak vines or half vines were partially discounted and dead vines eliminated. Sometimes there were several dead plants in a row and these were totally discounted. A few plants could more significantly change the results in a short test period than in the criterion period (e.g., total plants pruned in day divided by 8 hours). An improvement on the rating reliability procedure would have been to include multiple raters for the vine count.

Predictor

Farm 1 involved a concurrent study, as expected. Farm 2 was a predictive study in which workers were told they would be hired on the basis of the test but most workers were hired

regardless of test results. Borderline workers were given several days to improve. Farm 3 involved a concurrent study and a predictive one. In the predictive one, few workers were selected compared to the high number of applicants. Some borderline cases were selected and given a chance to improve for a few days.

Predictor reliabilities were high (see Table 1). Farm-wide reliabilities ranged from (r) .79 to .86. The highest individual crew predictor reliability took place on Farm 1 (r = .96) and the lowest in the concurrent study on Farm 3 (r = 0.52, ns). The latter was the only non-significant predictor reliability found.

High reliabilities simply mean that workers performed similarly between both tests (Test 1 and Test 2). High test reliabilities for both concurrent and predictive studies do not necessarily tell us whether workers were equally motivated to do their best in both concurrent and predictive tests. Logically, one might predict that predictive test conditions would be more motivating to workers than concurrent test conditions. In the predictive study, a worker's performance on the test can mean obtaining--or losing--a job. On Farm 3, a larger mean test score for workers on the concurrent test over the predictive one has at least three plausible explanations.

Table 1

Farm-wide and Crew Predictor Reliabilities

Crews	Farm			
	1	2	3A	3B
Study	Concurrent	Predictive	Predictive	Concurrent
A (n)	.81*** (21)	.88*** (25)	.85*** (39)	.85*** (21)
B (n)	.96*** (17)	.75*** (18)	.91*** (43)	.52 (12)
C (n)	.91*** (23)	--	.95** (6)	.88*** (19)
D (n)	.74*** (17)	--	.93*** a (9)	--
E (n)	.74*** (19)	--	.57* (17)	--
F (n)	.83*** (14)	--	--	--
FARM-WIDE (n)	.86*** (111)	.84*** (43)	.84*** (105)	.79*** (52)
Crit 1 \bar{X} (SDn-1)	20.48 (3.36)	13.96 (4.42)	14.35 (3.46)	21.83 (5.46)
Crit 2 \bar{X} (SDn-1)	21.21 (3.75)	14.52 (4.20)	14.50 (3.49)	22.04 (5.93)

*p < .05. **p < .01. ***p < .001. aThis reliability not included in summary--or validity analysis--as test 1 period not 46 minutes long.

pruned with low quality. The concurrent study involved those workers recently hired because their speed and quality was acceptable as well as the regular workers from previous years.

First, the predictive study involved a quality test and a speed test. Originally, only workers who passed the quality test would be allowed to go on to the speed test. This did not occur, and workers were rated both on their quality (Appendix B--pruning quality form) and speed (Appendix C--pruning speed form). Many workers did not know enough about cordon pruning to know they were not doing a good quality job while others must have been very aware of the quality rating. Workers on the concurrent test knew what the minimum acceptable quality was and were able to prune as fast as they could without doing a bad quality job.

Second, the researcher did not demand the same precision in the concurrent test as in the predictive one at Farm 3. In the predictive test workers were not allowed to move on to the next vine until the one they were working on was completely finished. Workers were permitted to prune without removing the brush from under the vines in the concurrent test; also, some workers would prune one side of the vine and come back and prune the other and therefore work more efficiently. The researcher had to estimate the total number of vines pruned in about 7 cases (no difference in correlations with or without these 7 cases).

Third, workers in the predictive study included new workers, many who were not hired because they were so slow or

pruned with low quality. The concurrent study involved those workers recently hired because their speed and quality was acceptable as well as the regular workers from previous years.

Intercorrelations between the predictive and concurrent tests were low but significant (Table 2). Beside the problems stated, too few numbers were involved in the intercorrelation to make any definite conclusions about how predictive and concurrent studies compare. Anecdotal observation, however, can shed some light on the question.

In all three farms, in both concurrent and predictive settings workers seemed very competitive and several tried to get a head start (workers who made any cut before the test started, however, were told to skip that vine and start with the next). Some pruners told others to slow down, often using the fear of a speed-up (were farmers would reduce the piece rate) as a reason for their colleagues to slow down. Such comments as "No me dejes atrás, no te apures tanto" (Don't leave me behind, don't hurry so much), and "¿Quieres podar por menos?" (Do you want to prune for less?) were common. Notwithstanding the calls to slow down, most workers seemed motivated to do their best in both the concurrent and predictive studies.

It was anticipated that workers might be concerned about

predominantly given in Spanish except to a few who preferred speed-ups in the concurrent studies, but not in the predictive ones. In the predictive study on Farm 2, one worker stopped working to help another. This was very surprising and could perhaps be explained by groups of workers who prefer--or need to because of transportation--to work together. In such circumstances, getting the job when a friend did not was useless.

Table 2 Distribution of scores, which closely fit normality, on Farm

Intercorrelations Among Predictive and Concurrent Scores For Farm 3 Subjects

	Concurrent Tests:		Predictive Tests:	
	1	2	1	1
r	.53**	.53**	.44*	.52**
(n)	(24)	(28)	(27)	(29)

*p < .05. **p < .01.

Except for the data obtained in the concurrent test on Farm 3, guidelines for working from one vine to the next were carefully kept (see above, "Work-sample test procedure," pp. 20-21). Cleaning under the vines on Farm 1 was done but not as carefully as it could have. Test instructions were

predominantly given in Spanish except to a few who preferred English or did not understand Spanish.

Finally, review of the distribution of scores on Farm 3 showed normal curves for the predictive tests while positively skewed curves for the concurrent tests (scores, in terms of vines pruned per forty-six minute test period ranged from 3 to 24 and from 5 to 26 for the predictive test, and ranged from 12 to 38 and 10 to 40 in the concurrent test). The distribution of scores, while closer to normality on Farm 1 (concurrent study) was also slightly skewed to the right (scores ranged from 12 to 28 and 14 to 30). Comparison of scores between farms is not possible as vine age, spacing, and vine conditions were not similar. Details of viticultural differences are not provided, in order to keep farm identities confidential.

Criterion

Farm-wide criterion reliabilities were significant (Table 3), except for Farm 2. Several possible explanations are offered for such nonsignificance on Farm 2. First, there is no relation, or even a negative one, between what workers will do one day and another even under some piece rate conditions. Some credence is given to this argument when several persons work at the same speed and there is more variance between days than among people. Some examples of

Table 3

Farm-wide and Crew Criterion Reliabilities

Test	Farm			
	1	2	3A	3B
CREWS	Concurrent	Predictive	Predictive	Concurrent
A (n)	.85*** (18)	--	--	.74*** (19)
B (n)	.82*** (17)	--	--	.82** (11)
C (n)	.75*** (19)	--	--	.73** (14)
D (n)	.65** (16)	--	--	--
E (n)	.92*** (23)	--	--	--
F (n)	.75** (13)	--	--	--
FARM-WIDE (n)	.76*** (106)	-.44 (16)	.51* (20)	.57*** (44)
Crit 1 \bar{X} (SDn-1)	27.46 (4.83)	34.73 (5.78)	30.48 (7.29)	30.96 (6.68)
Crit 2 \bar{X} (SDn-1)	32.59 (6.47)	22.68 (5.70)	31.12 (8.01)	28.77 (7.19)

*p < .05. **p < .01. ***p < .001. Samples too small for crew criterion reliabilities on predictive studies.

this phenomenon were present in Farm 1 and Farm 3. Some husband-wife or friend teams worked at the same speed. In one case a friend team helped each other on the same row when one would get behind (this pair was left out of all the correlations). In one crew on Farm 1 several pruners worked at the same speed for several days in a row, including one of the two randomly selected criterion days but not the other. Relatively high criterion reliabilities for that crew (Farm 1, Crew B, 11 of 17 workers performed at the same speed, Criterion 1, $r = .82$, $p < .001$) could be explained by presence of slower workers not in the same-speed group. While sometimes workers might slow down to work together to avoid speed-ups, this does not seem to be the case here as workers who performed at the same speed were faster than the rest of the crew.

A second explanation revolves around total number of hours worked and/or exactness in the vine count. When workers are paid by the vine some farmers might be less exact about documenting start and finish times than when paying by the hour. Managers from Farm 1 and Farm 3 mentioned that workers start at the same time in the morning and while leaving time is not always the same, foremen do a good job of writing leaving time. Farm 2's manager said that workers are permitted to straggle in (as much as 1/2 hour) or leave early

without notice as foremen are more involved in checking quality of work than in determining start or finish time. Farm 2's manager estimated that for some workers this could amount to as much as 1 1/2 hours off from the official working time per day. Along these lines, there are other reasons why criterion measures are not as precise as they should (these reasons were not reported as having occurred on Farm 2): Some foremen do not count total number of vines but will credit a worker with an unfinished row, resulting in an artificially high vine count. Conversely, a worker might prune quite a few vines credited to a previous day so the total vine count is lower than what was actually pruned that day.

Third, Farm 2 involved such a small number of subjects ($n = 16$) that were present between Criterion 1 and Criterion 2 measurements that there could have been large room for error. A similar observation about low subject numbers in general could be made of both Farm 3 studies.

Criterion dates originally were going to be the same farm-wide. Nevertheless, this was not possible as not all crews worked on the same variety on the same day. In one farm, for instance, one variety was considered easier than another and so worker crews took turns pruning easier varieties. The dates were all picked at random, however,

following the provisions set out in the methodology section. In addition, if a day was selected in which workers had to drive to move from one block to another; or if the farmer mentioned that a certain part of the vineyard was different than that of the other crews, these dates were rejected and others chosen at random. Once the dates were picked the researcher saw the criterion data for the first time. Farm 2 was particularly limited in the number of available days that met all the pre-set conditions as the employer decided early in the season not to keep his crews and decided to work with a farm labor contractor.

Data for Farm 3 was directly obtained in terms of vines/hour while data for Farm 1 and 2 was calculated from total daily vine count and number of hours worked as shown in the methodology section.

There were three cordon pruned grape varieties involved (Table 4). French Colombard was used as the predictor variety for all three farms. Criterion varieties included French Colombard (F.C.), Barbera (B), and Chenin Blanc (C.B.). There were other varieties available, but these were chosen at random out of a selection of cordon pruned varieties. Criterion 1 and Criterion 2 were more closely related to varieties than to dates.

Table 5
 Table 4

Farm-wide Validity Results

Criterion Varieties

Study:	Farm			
	1	2	3A	3B
Test	Concurrent	Predictive	Predictive	Concurrent
Criterion 1	F.C.	F.C.	C.B.	C.B.
Criterion 2	B	B	F.C.	F.C.

F.C. = French Colombard. C.B. = Chenin Blanc. B = Barbera

Validities

Validity coefficients ranged from $-.13$ (n.s.) to $.73$ ($p < .001$) on farm-wide results (Table 5). The only farm that showed no levels of significance either farm-wide or in individual crews was Farm 2. This finding does not seem to contradict the notion that very unreliable criterion measures would make a test--no matter how reliable--invalid. The only other farm-wide results that were extremely low were found in the low Criterion 2 results of Farm 3B (concurrent study).

Table 5

Review of individual crew validity coefficients for

Farm-wide Validity Results

Farm 3B (Table 7).

Table 7. (for Farm 3B) does Farm predict any crew validity

	1	2	3A	3B
Study:	Concurrent	Predictive	Predictive	Concurrent
Test 1				
CR 1 (r2)	.73***(.53)	.35(.13)	.41*(.17)	.60***(.36)
(n)	(110)	(21)	(26)	(43)
CR 2 (r2)	.72***(.52)	.11(.01)	.66**(.44)	.14(.02)
(n)	(108)	(18)	(20)	(45)
Test 2				
CR 1 (r2)	.67***(.45)	.23(.05)	.52**(.27)	.59***(.35)
(n)	(108)	(20)	(27)	(47)
CR 2 (r2)	.61***(.37)	-.13(.02)	.67***(.45)	.31*(.10)
(n)	(106)	(17)	(21)	(47)

*p < .05. **p < .01. ***p < .001.

Review of individual crew validity coefficients for concurrent studies are presented for Farm 1 (Table 6) and Farm 3B (Table 7).

Table 7 (for Farm 3) does not present any crew validity coefficients as low as the farm-wide Criterion 2 validity coefficients. It is possible that Criterion 2 involved vineyard blocks of different difficulty levels. This could explain how validity coefficients were higher (although not necessarily significant) for each crew than for the whole farm.

On the other hand, Farm 1 (Table 6) does present some-- even if few--non significant crew validity coefficients. This is especially true for Crew F. It is possible that there was some error due to such small number of crew members participating in Crew F. Another possibility is that since workers were not randomly assigned to crews that some crews had faster workers than others. Reliabilities for predictor (.83) and criterion (.75) in Crew F were not particularly lower than those of other crews. It does not seem that low validities for Crew F can be explained in terms of the predictor or criterion reliabilities. From this study it was not possible to be sure why Crew F had low validity coefficients.

Table 6

Farm 1 Validity Coefficients by Crew

Crew	Test	1		2	
	CR	1	2	1	2
A (r2) (n)	.79***(.62) (19)	.88***(.77) (20)	.86***(.74) (19)	.83***(.69) (20)	
B (r2) (n)	.78***(.61) (17)	.83***(.68) (17)	.85***(.72) (17)	.82***(.67) (17)	
C (r2) (n)	.80***(.64) (23)	.73***(.54) (19)	.69***(.47) (23)	.76***(.57) (19)	
D (r2) (n)	.62*(.39) (16)	.59*(.35) (16)	.75***(.56) (16)	.51*(.26) (16)	
E (r2) (n)	.66***(.43) (22)	.67***(.45) (22)	.59**(.34) (20)	.40(.16) (20)	
F (r2) (n)	.35(.12) (13)	.62*(.38) (14)	.07(.004) (13)	.39(.15) (14)	
Farm-wide(r2) (n)	.73***(.53) (110)	.72***(.52) (108)	.67***(.45) (108)	.61***(.37) (106)	

*p < .05. **p < .01. ***p < .001.

Table 7

Farm 3B (Concurrent) Validity Coefficients by Crew

Crew	CR	The Test		2	
		1	2	1	2
A (r2)		.85***(.72)	.63**(.40)	.72**(.52)	.60**(.35)
(n)		(18)	(19)	(17)	(18)
B (r2)		.36(.13)	.53(.29)	.51(.26)	.86***(.74)
(n)		(10)	(10)	(14)	(12)
C (r2)		.39(.15)	.46(.21)	.54*(.30)	.54*(.29)
(n)		(15)	(16)	(16)	(17)
Farm-wide(r2)		.60***(.36)	.14(.02)	.59***(.35)	.31*(.10)
(n)		(43)	(45)	(47)	(47)

*p < .05. **p < .01. ***p < .001.

These results are important because (1) more effective work can be hired; and (2) the potential for increased employment testing in production agriculture has been expanded (from the harvest to grape planting).

Limitations

It also seems certain that employer's cannot see the test and assume that it will always work. Farm 3's test might not be totally invalid. Predictions and criterion

reliabilities are important. No inferences can be made either, about the relative effectiveness of concurrent and predictive tests.

Conclusions and Suggestions for Future Research

The principal question being raised was: Can a concurrent-type criterion-oriented test predict on-the-job performance of vineyard workers? Additional supportive data was gathered on a predictive-type criterion oriented test. Having established predictor and criterion reliabilities it is possible to turn to the validity analyses.

Farm 1, overall, had the highest validity coefficients. There seems little doubt that a concurrent type test can predict performance and that a predictive test can predict performance. Farm 3 results showed that there is some generizability to the use of concurrent tests in vineyards, and that predictive tests can also be valid predictors of work performance for vineyard workers.

These results are important because (1) more effective workers can be hired; and (2) the potential for use of employment testing in production agriculture has been expanded (from tomato harvest to grape pruning).

Limitations

It also seems certain that employers cannot use the test and assume that it will always work. Farm 2's test turned out to be totally invalid. Predictor and criterion

reliabilities are important. No inferences can be made, either, about the relative effectiveness of concurrent and predictive tests. In both predictive tests the number of persons lost was so high that there probably was more restriction of range than in the concurrent studies.

Farmers who desire to test workers can do so but responsibility for establishing validity at the individual farm firm still remains. One of the greatest limitations to the generalizability of the study is that it deals with (1) crew work; and (2) incentive-paid workers. Further, this is a static rather than a longitudinal study.

Future Research

Further research can take many possible directions: First, what factors tend to increase predictor reliability? What is the shortest test length that will be an effective predictor of speed? Is the first test or the second test more effective in predicting results? If the first test is more effective, is the presence of the second test important? If so, can the second test be shorter? If the second test is more effective in predicting results can the first test be shortened? Are tests in concurrent studies measuring the same factors as those in predictive studies? How is quality of pruning affected by different performance tests?

Second, what factors increase criterion reliability?

Does increased criterion reliability result in better crew performance? What is the effect of amount of pay on criterion reliability? How is criterion reliability affected by pay method (e.g., hourly vs. piece rate). How reliable is worker performance from year to year? What causes workers to sometimes work together at the same speed when they are being paid on a piece-rate basis?

Third, how well can a quality test predict quality of work when workers are paid a quality bonus and when they are not paid such a bonus? What type of validity coefficients result from testing workers who do not work in physical proximity?

Many of these questions cannot be answered without random assignment of workers and use of larger total number of subjects, while other questions can be answered through field studies.

Style Manual

The style manual used for this manuscript was the Publication Manual of the American Psychological Association (3rd ed.), revised in 1984.

Selected Bibliography

American Psychological Association. Handbook for educational and psychological testing. Washington, DC: Author, 1974.

Ministry of Education. Psychological testing. New York: Macmillan, 1974.

Ministry of Education. Statistical methods for educational and psychological testing. London: Her Majesty's Stationery Office, 1974.

Ministry of Education. Statistical methods for educational and psychological testing. London: Her Majesty's Stationery Office, 1974.

Ministry of Education. Statistical methods for educational and psychological testing. London: Her Majesty's Stationery Office, 1974.

Ministry of Education. Statistical methods for educational and psychological testing. London: Her Majesty's Stationery Office, 1974.

Ministry of Education. Statistical methods for educational and psychological testing. London: Her Majesty's Stationery Office, 1974.

Billikopf, G. E. (in press). Testing as a predictor of worker performance in tomato harvest. California Agriculture.

Chronbach, L. J., & Gleser G. C. (1965). Psychological tests and personnel decisions (Notes). University of Illinois Press.

Daniel, S. (1986). Science, system, or myth: Alternative 1. Formula used in Radio Shack EC-4004 owner's manual, p. 51. For similar formula see Little and Hills, 1978, pp. 167-194, p. 175.

Division of Industrial and Organizational Psychology. (1980). Principles for the validation and use of personnel selection procedures (2nd ed.). Berkeley, CA: American Psychological Association.

Overstaple, D., Barrett, G. V., & Alexander, R. A. (1985). The feasibility of traditional validation procedures for demonstrating job relatedness. Law & Psychology Review, 9, 35-44. Personnel Psychological Abstracts, 1985, 2114.

Selected Bibliography

American Psychological Association. Standards for educational and psychological testing, (1985). Washington, DC: Author.

Anastasi, A. (1982). Psychological testing (5th ed.). New York: Macmillan.

Barret, G. V., Phillips, J. S., & Alexander, R. A. (1981). Concurrent and predictive validity designs: A critical reanalysis. Journal of Applied Psychology, 66(1), 1-6.

Bersoff, D. N. (1981). Testing and the law. American Psychologist, 36(10), 1047-1056.

Billikopf, G. E. (1985a). Response to incentive pay among vineyard workers. California Agriculture, 39(7-8), 13-14.

Billikopf, G. E. (1985b). Pruning quality and output in response to pay method in vineyards. Unpublished manuscript.

Billikopf, G. E. (1987, January). How to fire without getting burnt. California Farmer, pp. 24-25E.

- Billikopf, G. E. (in press). Testing as a predictor of worker performance in tomato harvest. California Agriculture.
- Chronbach, L. J., & Gleser G. C. (1965). Psychological tests and personnel decisions (2nd ed.). University of Illinois Press.
- Daniel, C. (1986). Science, system, or hunch: Alternative approaches to improving employee selection. Public Personnel Management, 15(1), 1-10.
- Division of Industrial and Organizational Psychology. (1980). Principles for the validation and use of personnel selection procedures (2nd ed.). Berkeley, CA: American Psychological Association.
- Doverspike, D., Barrett, G. V., & Alexander, R. A. (1985). The feasibility of traditional validation procedures for demonstrating job relatedness. Law & Psychology Review, 9, 35-44. (From Psychological Abstracts, 1986, 73(4), Abstract No. 10587)
- Downs, S., Farr, R. M., & Colbeck, L. (1978). Self appraisal: A convergence of selection and guidance. Journal of Occupational Psychology, 51, 271-278.
- Ebel, R. L. (1977). Comments on some problems of employment testing. Personnel Psychology, 30(1), 55-63.
- Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. Federal Register, 43(166), 38290-38315.
- Equal Employment Opportunity Commission, Office of Personnel Management, Department of Justice, Department of Labor, & Department of the Treasury. (1979). Adoption of questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. Federal Register, 44(43), 11996-12009.
- Holloman, H. & Mills, F. J. (1978). Agricultural experimentation: Design and analysis. New York: John Wiley and Sons.

- Equal Employment Opportunity Commission, Office of Personnel Management, Department of Justice, Department of the Treasury, & Department of Labor Office of Federal Contract Compliance Programs. (1980). Adoption of additional questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. Federal Register, 45(87), 29530-29531.
- Farr, J. L., O'Leary, B. S., & Bartlett, C. J. (1973). Effect of work sample test upon self-selection and turnover of job applicants. Journal of Applied Psychology, 58(2), 283-285.
- Ghiselli, E. E. (1966). The validity of occupational aptitude tests. New York: John Wiley & Sons.
- Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. Personnel Psychology, 26(4), 461-477.
- Green, B. F. (1981). A primer of testing. American Psychologist, 36(10), 1001-1011.
- Guion, R. M. & Cranny, C. J. (1982). A note on concurrent and predictive validity designs: A critical reanalysis. Journal of Applied Psychology, 67(2), 239-244.
- Kleiman, L. S., & Faley, R. H. (1985). The implications of professional and legal guidelines for court decisions involving criterion-related validity: A review and analysis. Personnel Psychology, 38(4), 803-833.
- Lee, R., Miller, K. J., & Graham, W. K. (1982). Corrections for restriction of range and attenuation in criterion-related validation studies. Journal of Applied Psychology, 67(5), 637-639.
- Leedy, D. L. (1985). Practical research: Planning and design (3rd ed.). New York: Macmillan.
- Little, T. M. & Hills, F. J. (1978). Agricultural experimentation: Design and analysis. New York: John Wiley and Sons.

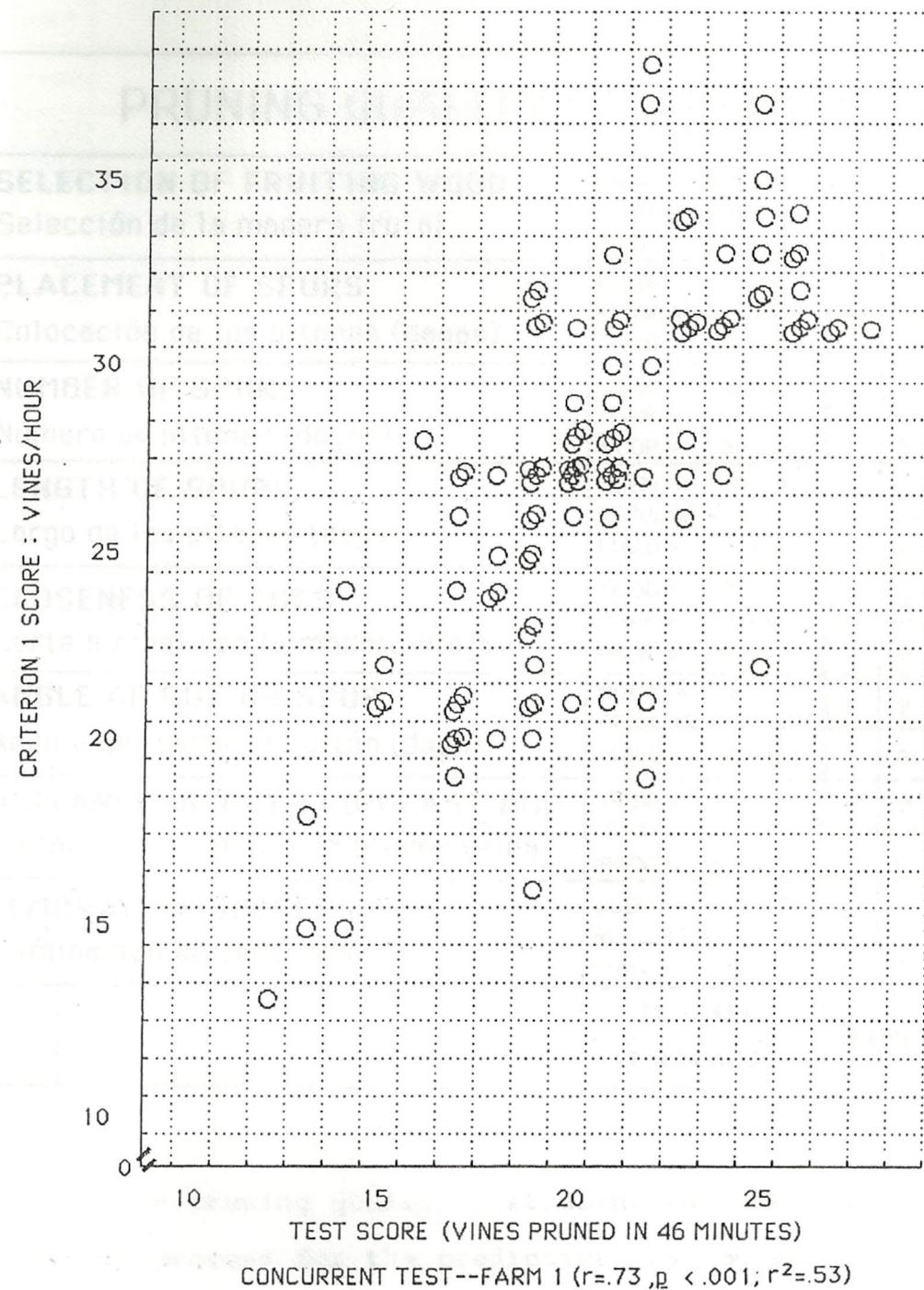
Schwitt, N., Gouling, R. Z., Kow, K. A., & Rieck, M. (1984). Metaanalysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37(3), 407-427.

- Lukacsko, Z. (1984). Mezogazdasagi repuloge pvezetok kivalasztasi rendszerenek kezdeti eredmenyei [Initial results concerning the system according to which agricultural air pilots are selected]. Magyar Pszichologiai Szemle, 41(2), 129-139. (From Psychological Abstracts, 1985, 72, Abstract No. 16033)
- Mount, M. K., Muchinsky, P. M., & Hanser, Lawrence, M. (1977). The predictive validity of a work sample: A laboratory study. Personnel Psychology, 30(4), 637-645.
- O'Leary, L. R. (1973). Fair employment, sound psychometric practice, and reality: A dilemma and a partial solution. American Psychologist, 28(2), 147-150.
- Ramos, R. E. (1981). Employment battery performance of Hispanic applicants as a function of English or Spanish test instructions. Journal of Applied Psychology, 66(3), 291-295.
- Robertson, I. T., & Kandola, R. S. (1982) Work sample tests: validity, adverse impact and applicant reaction. Journal of Occupational Psychology, 55(3), 171-183.
- Ramsay, R. T. (1981). Management's guide to effective employment testing: What's legal, valid and fair, Chicago: Dartnell.
- Tenopyr, M. L. (1981). The realities of employment testing. American Psychologist, 36(10), 1120-1127.
- Schlei, B. L., & Grossman, P. (1983). Employment discrimination law (2nd ed.). Washington: The Bureau of National Affairs.
- Schmidt, F. L., Greenthal, A. C., Hunter, J. E., Berner, J. G., & Seaton, F. W. (1977). Job samples vs. paper and pencil trades and technical tests: Adverse impact and examinee attitudes. Personnel Psychology, 30, 187-197.
- Schmidt, F. L., & Hunter, J. E. (1980). The future of criterion-related validity. Personnel Psychology, 33(1), 41-60.
- Schmitt, N, Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37(3), 407-422.

- Schultz, C. B. (1984). Saving millions through judicious selection of employees. Special Issue: Assessment techniques and challenges. Public Personnel Management, 13(4), 409-415.
- Siegel, J. (1980). Personnel Testing Under EEO. New York: Amacom.
- Wakefield, J. A. & Goad, N. A. (1982). Psychological differences: Causes, consequences, and uses in education and guidance. San Diego: Edits.
- Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. Journal of Applied Psychology, 52(5), 372-376.
- Whelchel, B. D. (1985). Use of performance tests to select craft apprentices. Personnel Journal, 64(7), 65-69.

Appendices

Appendix A--Bivariate Distribution of Scores (Farm 1)



Bivariate distribution of Test 1 and Criterion 1 (for Farm 1) shows linear relationship of the correlation.

Appendix B--Pruning Quality Data Collection Instrument

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			
PLACEMENT OF SPURS Colocación de los pitones (dagas)	BUENO = 0-2	3	2	3 x		
	REGULAR = 3-4	1	0			
NUMBER OF SPURS Número de pitones (dagas)	GOOD= 0-2	3	2	2 x		
	FAIR = 3-4	1	0			
LENGTH OF SPURS Largo de los pitones (dagas)	BUENO = 0-2	3	2	2 x		
	REGULAR = 3-4	1	0			
CLOSENESS OF CUTS Corte a nivel con la madera vieja	GOOD= 0-2	3	2	2 x		
	FAIR = 3-4	1	0			
ANGLE OF CUT ON SPUR Angulo del corte del pitón (daga)	BUENO = 0-2	3	2	1 x		
	REGULAR = 3-4	1	0			
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			
REMOVAL OF SUCKERS Eliminación de chupones	BUENO = 0-2	3	2	1 x		
	REGULAR = 3-4	1	0			
		MISTAKE TOLERANCE		SCORE	FACTOR	TOTAL:

This pruning quality instrument was used as part of the testing process for the predictive test on Farm 3.

This pruning quality data collection instrument was used to collect predictor pruning quality on all farms.

Appendix C--Pruning Speed Data Collection Instrument

UNIVERSITY OF CALIFORNIA							AGRICULTURAL EXTENSION								
ROW #															
# VINES															
FARM CODE	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
	2 2	2 2	2 2	2 2	2 2	2 2	2 2	2 2	2 2	2 2	2 2	2 2	2 2	2 2	2 2
	3 3	3 3	3 3	3 3	3 3	3 3	3 3	3 3	3 3	3 3	3 3	3 3	3 3	3 3	3 3
	4 4	4 4	4 4	4 4	4 4	4 4	4 4	4 4	4 4	4 4	4 4	4 4	4 4	4 4	4 4
DATE	5 5	5 5	5 5	5 5	5 5	5 5	5 5	5 5	5 5	5 5	5 5	5 5	5 5	5 5	5 5
	6 6	6 6	6 6	6 6	6 6	6 6	6 6	6 6	6 6	6 6	6 6	6 6	6 6	6 6	6 6
	7 7	7 7	7 7	7 7	7 7	7 7	7 7	7 7	7 7	7 7	7 7	7 7	7 7	7 7	7 7
	8 8	8 8	8 8	8 8	8 8	8 8	8 8	8 8	8 8	8 8	8 8	8 8	8 8	8 8	8 8
	9 9	9 9	9 9	9 9	9 9	9 9	9 9	9 9	9 9	9 9	9 9	9 9	9 9	9 9	9 9
VARIETY	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4	1/4
	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2	1/2
	3/4	3/4	3/4	3/4	3/4	3/4	3/4	3/4	3/4	3/4	3/4	3/4	3/4	3/4	3/4
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SPACING															
VINE AGE															
TEST #															
NAME															
AGE															
ETHNIC	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
SEX	M F	M F	M F	M F	M F	M F	M F	M F	M F	M F	M F	M F	M F	M F	M F

This pruning speed data collection instrument was used to collect predictor pruning speed on all farms.

