Quality evaluations should not be taken for granted

by Gregory Encina Billikopf

Subjective quality-evaluation errors in agriculture, such as discarding good-quality product and packing poor-quality product, can be costly to growers and workers. This study of workers and supervisors in a strawberry-plant packingshed revealed the danger in assuming that those responsible for quality control truly understand what is required. We found that the ability of workers to correctly count plants, and to retain or reject them (and explain why), varied considerably. The results highlight the need for employers to carefully define quality parameters, and then test employees and applicants. When top management does not agree on exactly what constitutes acceptable quality, it is difficult to expect quality-control inspectors and workers to understand. Testing, as a tool, can help growers and producers make better employee selection and placement decisions and can also be used for periodic training.

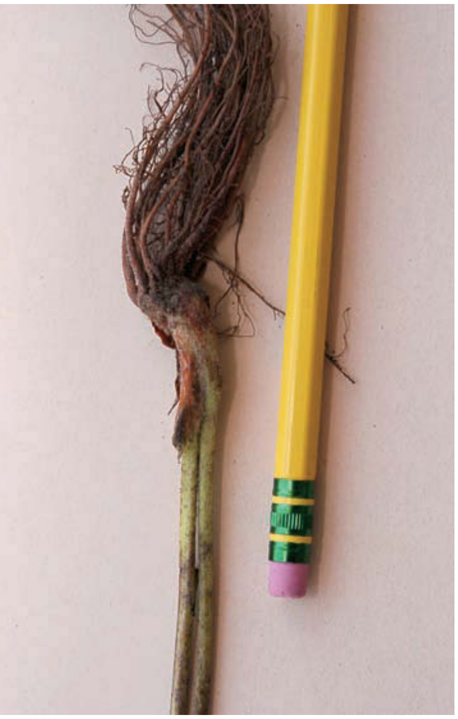
Workers in a California packingshed were tested on several tasks related to sorting and packing strawberry plants. Grower Bob Whitaker (right) and his top manager Areli Toledo banter as they review the the plant ratings for specimens that they did not agree on; Toledo scored highest on the test.

Most agricultural tasks require people to make important subjective decisions of a qualitative nature. For instance, should fruit be picked or left on the tree to reach optimal maturity? Should a cow be milked or moved to a hospital to be treated for mastitis? Does a field need to be irrigated? Should a cucumber on a conveyer belt be packed or discarded? Subjective decisions are made at all hierarchical levels, from farm owner to farmworker.

Over the last 2 decades, the author has carried out a number of informal studies in an attempt to measure "rater reliability" in California and Chile. At one operation in Chile, for example, several managers rated the quality of pruning in a fruit orchard and there was no agreement among them. On another occasion, several respected California viticulturalists were asked to rate the quality of 10 grapevines pruned by different employees. After the score sheets were returned, these raters were asked to go back and redo the evaluation. Often their new scores did not agree with their scores from half an hour before (Billikopf 1994, 2003).

While the consequences for incorrect decisions may vary, such qualitative decision-making is usually a key aspect of farming. But at all operational levels, people in agriculture are usually hired without testing their ability to make qualitative decisions. This casual approach to hiring even extends to research assistants, who are most often interviewed but seldom tested for rater reliability; likewise, inter-rater reliability is rarely checked before results are reported. Such a casual approach to selection compromises the integrity of research results as well as farm profits.

Effective human-resource management offers valuable tools to help improve such critical outcomes. Practical tests (also called "job samples") are an effective and legal way to enhance selection and placement decisions, as well as the training and performance evaluation of present employees (Billikopf 1988, 2003; Federal Register 1978; US Department of Labor 1999).



***Top left*, strawberry-plant workers use a trim tool to cut off plant stems. *Top right*, study participants evaluated 150 numbered samples of strawberry plants, so that their scores could be compared. *Left*, plants suitable for packing should have crowns roughly the size of a pencil, or larger; this root crown is on the small side.**

While the literature mentions the use of testing in agriculture, and even testing that involves the need for test takers to make qualitative decisions (Billikopf 1988, 2003), little has been written on rater reliability in agricultural employment testing (as either a selection, evaluation, placement or testing tool). One exception is Campbell and Madden (1990), in which raters were asked to evaluate the percentage of plant disease incidence in particular samples.

Another example is Mcquillian (2001), who tested medical personnel for how accurately they made decisions regarding medical cases, based on specific guidelines. Much more common is research that studies the reliability of tools or instruments, such as the reliability of a medical survey instrument used for brain injury diagnosis (Desrosiers et al. 1999). Because medical decision-making can have life-and-death implications, it makes sense that much of the work in this field has been conducted in the medical arena.

This study examines whether individuals vary in terms of their ability to make reliable and valid evaluative decisions (that is, their rater reliability), and if this can be measured through the use of a job sample or practical test.

Strawberry-packing study

Data was gathered at a California strawberry-plant packingshed. While the study could have been carried out in any agricultural industry, a task was selected in which workers make multiple quick decisions that can easily be measured against a known standard.

Strawberry plants (for replanting) are harvested in the field and brought to the shed in large, tangled clusters that are separated by workers. Plants are then sorted in terms of a single passing grade (the remaining plants are discarded). Good plants are bunched into groups of 100 units and then packed for shipping nationally and abroad. Sorters are responsible for all the tasks, from untangling the plant clusters to bunching them into 100-plant units. The sorter's most critical job is inspecting each plant and determining if it should be discarded or retained, a task that normally is carried out in less than a second per plant.

After the sorters have done their job, several levels of quality-control personnel inspect the plants. (We define quality control as a system to check that sorters are making correct evaluative decisions.) The two most important quality issues are ensuring that good plants (without defects) are packed and that each bunch contains 100 plants.

While sorters must recognize which plants to retain or reject, quality-control personnel must also be able to understand and describe the reason for rejecting particular plants. This extra detail is needed so that sorters can receive feedback on their performance.

Two salient and costly quality-evaluation errors are (1) discarding good product as not salable and (2) packing poor-quality plants. Discarding good plants is detrimental to both the grower and sorters. The grower loses good plants and all the costs involved in growing them; and the workers, who are often paid on a piece-rate basis, lose good plants they could have packed and earned money from.

A poor-quality pack also has negative economic consequences for the plant buyer, who may cultivate nonviable plants or need to re-sort them beforehand. In order to make up for defective plants, some growers ship an extra 10% free. Growers who ship a

Some of the quality-control personnel did quite poorly in this test, with the super checker doing worse than both the checkers and counters she was supposed to direct.

higher quality pack could gain a competitive edge and positive reputation while saving on plants.

Testing for accurate evaluations

Initial preliminary tests were carried out in December 2004, but the data reported here was collected in September and October 2005. During the initial tests, it became clear that we could not conduct an effective test until top management agreed on what constituted good quality and the reasons for rejecting plants. It took several weeks of negotiation and close work with management to develop a set of known criteria.

Through the testing process we set out to determine how accurately subjects would be able to: (1) count plants per bunch; (2) make reject-versus-retain decisions for each plant; and (3) label the reason for rejecting a plant. To be effective, sorters must make accurate decisions, but not necessarily explain these to someone else. Quality-control personnel, in contrast, must clearly articulate the reason for rejecting plants. Flexibility is required since clients buying the plants can vary in terms of quality pack requirements.

For practical reasons, distinct aspects of the job were tested separately. The first dealt with the accuracy of plant count, and the second with retain-versus-reject reasons. For this study, six distinct reasons were agreed upon for discarding plants. From most serious to least serious, they were: (1) cut crown, (2) black roots, (3) inadequate number of healthy roots, (4) thick crowns, (5) thin crowns and (6) lack of root hairs. For instance, if a plant had a cut crown and black roots, the recorded reason for rejecting it should be the most serious, the cut crown.

Subjects (employees) were shown samples of each discard category and were encouraged to ask questions. Some clearly took better advantage of this opportunity than others.

Subjects, including, *left*, Luz Maria Romero and, *right*, Silvia Araiza, had to make retain-versus-reject decisions for 150 strawberry plants and provide the reason for rejection. Despite the apparent simplicity of the task, few subjects scored well against the known correct answers.

For the retain-versus-reject test, the statistical analysis was adapted from the Gage Repeatability and Reproducibility (Gage R&R) quality evaluation tool. The Gage R&R instrument is often used to test the consistency of a measuring gage in the hands of multiple raters. Here, the instrument being tested was a person rather than a gage.

For both tests we developed an answer key with the known criterion against which subjects would be compared. There was no null hypothesis to test, but rather the ability of each subject to make quick, accurate decisions.

Subjects tested included the grower/shipper, top manager, super checker, checkers, counters and sorters. While the grower and top manager may communicate quality pack standards, it is the super checker who is responsible for checking the work of the regular checkers and counters. The checkers focus mostly on plant quality, while the counters focus on plant count. There is some overlap between the responsibilities of these two job categories.

Accuracy varied widely

Counting. Twenty-four subjects (22 female and two male) participated in the counting test. A total of 2,919 plants were spread out in uneven bunches at 12 stations (bunches ranged from 200 to 300 plants, with a mean of 243 plants).

One subject recorded 818 plants in a station that only had 222, throwing off her score by a large margin. The remaining participants ranged from a total of 12 mistakes (an average of one mistake per station or 0.4% error) to 163 mistakes (more than 13 mistakes per station or 5.6% error).

There was sufficient overlap in terms of subjects who participated in the counting test and the retain-versus-reject test to note that those who could count accurately were not necessarily the same as those who did well in the reject-versus-retain test, and vice versa (table 1).

Retain versus reject. Thirty-two subjects (29 female and three male) participated in the retain-versus-reject test. Two separate sets (A and B) consisted of

150 plant samples each. Subjects were given 5 seconds per plant to make and annotate their evaluative decisions. Plants were labeled from 1 to 150 (in groups of five plants per station, with 30 stations per set).

Subjects were divided into two groups, half in set A and half in set B. Each subject evaluated the set of samples to which she or he was assigned twice. Only after the first test was completed and the score sheets collected did subjects proceed to the retest (with a new, blank score sheet).

For each subject, we obtained: (1) a test score (test results compared to known criterion); (2) a retest score (how subjects scored against a known criterion when repeating the same test for a second time); (3) an average test-versus-retest score; and (4) a reliability score (for every decision, how consistently did each subject agree with herself or himself as they evaluated the same plants twice) (table 1).

TABLE 1. Job category, number of completed samples and reliability score between the test and retest

Position-ID #	Samples evaluated	Raw count error	Reliability	Test	Retest	Avg. test/ retest
	no.	*no. (%)*	%			
Sorter-1	150		84.00	62.00	55.33	58.67
Sorter-2*	149		67.11	76.51	53.33	64.92
Sorter-3*	150		84.67	62.00	70.67	66.33
Sorter-4*	128		52.54	53.49	81.88	67.69
Sorter-5	150		60.00	88.67	59.33	74.00
Sorter-6	150		86.00	76.00	72.67	74.33
Checker-7	150	33 (1.1)	86.67	80.00	73.33	76.67
Sorter-8	150		79.33	78.67	75.33	77.00
Sorter-9	150		84.00	82.67	73.33	78.00
Checker-10*	147	163 (5.6)	74.15	73.65	83.22	78.44
Sorter-11	150		90.67	80.67	80.67	80.67
Sorter-12	150		76.67	81.33	83.33	82.33
Sorter-13	150	15 (0.5)	90.67	82.00	83.33	82.67
Sorter-14*	148		82.88	85.23	84.35	84.79
Sorter-15	150		84.00	82.00	88.67	85.33
Sorter-16	150		84.00	84.67	86.00	85.33
Sorter-17	150		89.33	84.00	86.67	85.33
Sorter-18	150		86.67	85.33	85.33	85.33
Sorter-19	150		84.72	84.72	87.50	86.11
Counter-20	150	33 (1.1)	92.00	86.67	88.00	87.33
Super checker-21	150	44 (1.5)	84.00	92.67	84.67	88.67
Sorter-22	150		86.67	88.67	88.67	88.67
Counter- 23	150		86.67	88.00	90.67	89.33
Counter-24	150	32 (1.1)	90.67	88.67	90.00	89.33
Counter-25	150	57 (2.0)	90.67	93.33	89.33	91.33
Checker-26	150	12 (0.4)	90.67	92.00	90.67	91.33
Sorter-27	150		91.33	91.33	92.00	91.67
Counter-28*	146	37 (1.3)	94.44	90.41	94.59	92.50
Owner-29	150		91.33	94.00	92.00	93.00
Checker-30	150		94.67	94.00	94.00	94.00
Manager-31	150	20 (0.7)	96.00	96.00	94.67	95.33

* Subjects did not complete the reasons for rejecting plants.

The average test/retest scores ranged from a high of 95.3% (excellent by any standard) to a low of 58.7%. Had the low-scoring subject indiscriminately accepted all plants for packing without rejecting any, she would have scored better (60%). In fact, it was much more common for subjects to reject good plants than to pack bad ones. Campbell and Madden (1990) also found that experienced raters tended to overestimate plant disease incidence. Our results are similar to those of the medical decision-making study (Mcquillian 2001), in terms of finding a large variation between the best and worst rater in the group.

As test scores increased, reliability scores generally increased as well. Low reliability scores (i.e., assigning different quality scores to the same plants) mean that a subject does not see quality issues consistently. It is possible for individuals to have high reliability scores, yet do poorly in the test/retest. Such individuals may have a reliable eye for quality, but be calibrated to a different north.

We told prospective study participants that they must be able to read and write, but nonetheless had one subject who could not fill out the score sheet. Perhaps this individual felt trapped into making a face-saving move, or else wanted the hourly wage that the grower paid to study participants.

Of the remaining 31 subjects, six turned in partial results. They recorded retain-versus-reject decisions, but not reject reasons. These six ranged from the second lowest score to the fourth highest of all participants in terms of their average test/retest scores (table 1).

Identifying discard reasons. As long as sorters understand quality parameters, it is not essential that they (1) can explain it, or (2) can read or write. In contrast, quality-control personnel must be able to do both. The remaining 25 subjects (23 female, two male) completed the final portion of the study, where the reasons for rejecting plants were incorporated into retain-versus-reject decisions. Average test/retest scores ranged from a low of 40% to a high of 92% (table 2).

Subjects who scored highly in the test/retest also tended to have higher reliability scores. Some of the quality-control personnel did quite poorly in this test, with the super checker doing worse than both the checkers and counters she was supposed to direct. Several checkers and counters showed great potential for a super-checker position and were likely to improve with additional training.

As expected, we found high variability among subjects in terms of consistently being able to count plants, make retain-versus-reject decisions and determine the reason for rejecting plants. This variability existed among subjects who were already employed and supposedly knowledgeable. Had we administered the tests to applicants unfamiliar with the industry, we would expect to see even greater variability.

Job samples for testing employees

Our tests involved straightforward, objective issues (such as counting), as well as more subjective questions (such as whether a strawberry plant has sufficient root hairs). We found that subjects who did well in one test did not necessarily do well on another. Consequently, employers should also consider the use of tests to make placement decisions.

Selection procedures for particular tasks vary widely in terms of how valid they are. Validity, in the context of employment testing and placement, deals with how well an instrument or test predicts on-the-job behavior. Intelligence and personality tests are of limited value for predicting job performance, while job samples are highly valid predictors.

A job sample involves asking subjects to perform portions of the actual job duties. Examples may include picking oranges, pruning a deciduous orchard, driving a tractor or treating a calf. Agriculture lends itself well to job sample testing. Farm employers can set up several stations with different job duties to test and evaluate (Billikopf 2003).

It is important to test for as many different types of job tasks as the person will perform on the job. Such practical tests can be easily submitted to content oriented validity and in some instances may also be validated through a criterion-oriented approach (Anastasi 1982; Billikopf 1988, 2003; Federal Register 1978; US Department of Labor 1999). Tests can be designed so that subjects need not be able to read or write. The individualized nature of these tests can make them more time-consuming, however.

The most common error in the reject-versus-retain test was discarding good plants. A combination of preselection testing and careful placement, as well as the use of testing as a performance evaluation and training tool, should reduce material waste and at the same time increase worker wages by a considerable percentage (such as when workers get paid for plants they were previously discarding). Without testing, management mistakes could lead to, for example, placing a super checker in a position of responsibility (such as training and evaluating) over more-skilled individuals.

The objective of this study was to warn researchers involved in subjective evaluations, as well as farm employers whose personnel must evaluate quality on-the-job, that quality determinations should not be taken for granted. Even though the study was carried out in a specific industry, almost every agricultural industry should pay more careful attention to quality.

TABLE 2. Reliability, average test/retest score, test score and retest score for retain-versus-reject test*

				With reject reason		
Position-ID #	Reliability	Reject-reason reliability	Avg. test/retest	Avg. test/retest	Test	Retest
				%		
Sorter-1	84.00	38.67	58.67	40.00	42.00	38.00
Sorter-11	90.67	68.00	80.67	58.00	54.00	62.00
Sorter-5	60.00	60.00	74.00	61.67	64.00	59.33
Sorter-6	86.00	63.33	74.33	62.67	63.33	62.00
Sorter-9	84.00	62.67	78.00	63.67	68.00	59.33
Sorter-8	79.33	65.33	77.00	64.00	64.67	63.33
Sorter-12	76.67	57.33	82.33	64.33	57.33	71.33
Super checker-21	84.00	59.33	88.67	67.33	68.67	66.00
Checker-7	86.67	70.00	76.67	68.67	70.00	67.33
Sorter-15	84.00	65.33	85.33	69.67	65.33	74.00
Counter-23	86.67	78.00	89.33	72.67	72.00	73.33
Sorter-18	86.67	70.00	85.33	74.00	77.33	70.67
Sorter-16	84.00	76.00	85.33	74.33	76.00	72.67
Sorter-22	86.67	74.67	88.67	74.67	73.33	76.00
Sorter-27	91.33	82.67	91.67	75.33	74.67	76.00
Sorter-17	89.33	82.00	85.33	75.67	72.67	78.67
Sorter-19	84.72	69.44	86.11	77.43	71.53	83.33
Sorter-13	90.67	86.00	82.67	78.33	78.67	78.00
Counter-24	90.67	78.67	89.33	78.67	78.67	78.67
Counter-20	92.00	80.67	87.33	80.67	80.00	81.33
Checker-26	90.67	80.00	91.33	82.00	78.67	85.33
Counter-25	90.67	80.00	91.33	82.67	84.67	80.67
Checker-30	94.67	86.00	94.00	84.67	84.00	85.33
Owner-29	91.33	84.00	93.00	89.67	88.67	90.67
Manager-31	96.00	92.00	95.33	92.00	92.00	92.00

*Test and retest scores are measures of how well subjects did when contrasted against known correct answers.

G.E. Billikopf is Area Labor Management Farm Advisor, UC Cooperative Extension, Stanislaus County. Readers may request an Excel spreadsheet for calculating reliability and test scores, or obtain additional information from the author at gebillikopf@ucdavis.edu.

References

Anastasi A. 1982. *Psychological Testing* (5th ed.). MacMillan. 784 p.

Billikopf GE. 1988. Agricultural Employment Testing: Opportunities for Increased Worker Performance. UC ANR, Giannini Found Spec Rep No 88-1. www.cnr.berkeley.edu/ucce50/ag-labor/7research/giannini.htm.

Billikopf GE. 1994. Agricultural Labor Management: Cultivating Personnel Productivity. UC Agricultural Extension, Stanislaus County.

Billikopf GE. 2003. Agricultural Labor Management: Cultivating Personnel Productivity (2nd ed.). UC Agricultural Issues Center. ANR Pub 3417. www.cnr.berkeley.edu/ucce50/ag-labor/7labor/AgLabor.pdf.

Campbell CL, Madden LV. 1990. *Introduction to Plant Disease Epidemiology.* New York: Wiley-Interscience. 523 p.

Desrosiers J, Mercier L, Rochette A. 1999. Test-retest and inter-rater reliability of the French version of the Ontario Society of Occupational Therapy (OSOT) Perceptual Evaluation. Can J Occup Therapy 66(3):134–9 (in French).

Federal Register. 1978. Uniform Guidelines on Employee Selection Procedure. Equal Employment Opportunity Commission. Section 60-3, 43 FR 38295. www.dol.gov/esa/regs/fedreg/final/2004004090.pdf.

Mcquillian S. 2001. Practice Variations: Inter-Rater Reliability Testing for Utilization Management Staff. Managed Care. www.managedcaremag.com/archives/0106/0106.peer_rater.pdf

US Department of Labor. 1999. Testing and Assessment: An Employer's Guide to Good Practices. Employment and Training Administration, Washington, DC. www.cnr.berkeley.edu/ucce50/ag-labor/7labor/test_validity.pdf.