Gregorio Billikopf, University of California (gebillikopf@ucdavis.edu, 209.525-6800)

11

Quality Control

In this chapter we will first look at the dangers of ignoring quality or assuming that our quality control is functioning as it should. Next, we focus on developing an effective quality control program.

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 in a conveyer belt be packed or discarded? Subjective decisions are made at all hierarchical levels, from farm owner to farm worker.

While the consequences for incorrect decisions may vary, such qualitative decision-making is usually a key aspect of farming and quality control.

Over the years I have carried out a number of informal and formal studies in an attempt to measure "rater reliability." 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 viticulturists were asked to rate the quality of 10 grapevines pruned

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.

by different employees. After the score sheets were returned, I discretely asked these raters to go back and redo the evaluation. Generally, their new scores did not agree with their scores from half an hour before. My first formal study involved a California strawberry packing shed in San Joaquin County.

Strawberry Packing Shed

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 to be discarded or retained, a task that normally is carried out in less than a second.

After the sorters have done their job, several levels of quality-control personnel inspect the plants. The two most important quality issues are ensuring that good plants (without defects) are packed and that each bunch contains 100 plants.

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. This extra detail is needed so that sorters can receive performance feedback.

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 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.

A poor-quality pack also has negative economic consequences for the plant buyer, who may cultivate nonviable plants or need to resort them beforehand.

In order to make up for defective plants, some growers ship an extra 10% free. Growers who ship a higher quality pack gain a competitive edge and a positive reputation while saving on plants.

The study examined whether individuals vary in terms of their ability to make consistent (reliable) and valid (correct) evaluative decisions, and if this could be measured through the use of a job sample or practical test.

Testing for accurate evaluations

Through the testing process we set out to determine how accurately sorters 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. 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, the second with retain-versus-reject reasons as well as an understanding for doing so.

For this study six distinct reasons were agreed upon for discarding plants. We ordered them as a combination of both how serious and how uncommon these defects are: (1) cut crown, (2) black roots, (3) inadequate number of healthy roots, (4) thick crowns, (5) thin crowns and (6) lack of root hairs. If a plant had a cut crown and black roots, the recorded reason for rejecting it should be the first in the list, the cut crown.

Sorters, consisting of experienced employees, were shown samples of each discard category and were encouraged to ask questions. Some clearly took better advantage of this opportunity than others.

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 gauge in the hands of multiple raters. Here, instead, we used the Gage R&R to test both rater reliability and rater validity.

Sorters 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 sorters 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 (over 13 mistakes per station, 5.6% error).

There was sufficient overlap in terms of sorters 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 11–1).

Retain versus reject. Thirty-two sorters participated in the retain-versus-reject test. Two separate sets (A and B) consisted of 150 plant samples each. Sorters 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).

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

For each sorter, we obtained: (1) a test score (test results compared to known criterion); (2) a retest score (how sorters 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

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.

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

Position-ID #	Reliability	Reject-reason reliability	Avg. test/retest	With reject reason		
				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.

himself as they evaluated the same plants twice) (Table 11–1).

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 sorters to reject good plants than to pack bad ones.

As test scores increased, generally reliability scores 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 sorters, 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 (Table 11–1)

in terms of their average test/retest scores.

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. To be effective, however, quality-control personnel must be able to do both. The remaining 25 sorters 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 11–2).

Sorters who scored highly in the test/retest also tended to have higher reliability scores. Some of the packing shed 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 had great potential for a super-checker position and were likely to improve with additional training.

As expected, we found high variability among sorters 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 sorters who were already employed and supposedly knowledgeable.

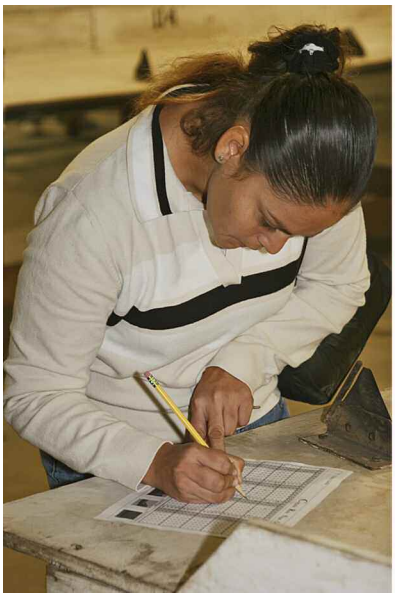
Had we administered the tests to applicants unfamiliar with the industry, we would expect to see even greater variability.

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 sorters who did well in one test did not necessarily do well on another. Consequently, employers should consider the use of tests to make placement as well as selection decisions.

Tests can be designed so that sorters 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 pre-selection 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.



Subjects 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.

Vineyard managers eye crop load in order to make thinning decisions and manage sugar levels. After doing so, grapes are weighed and comparisons made between managers estimates and actual weights.

Other observations

I have since been involved in formal as well as demonstration studies in numerous types of crops throughout Chile and California (in both plant and animal agriculture). The bottom line is that quality determinations should not be taken for granted.

Establishing a Quality Program

One of the dangers inherent in increased worker productivity—including instances when workers earn piece-rate—is the potential for poor quality, or the neglect of other non-measured desired outcomes.

A systematic approach towards quality improvement involves a number of specific steps:

- calibrate quality control personnel
- refine standards
- establish hierarchy of reasons for discards
- test quality control personnel
- evaluate quality control personnel
- train and test applicants
- establish sample procedure
- create a feedback mechanism
- train supervisors and runners
- establish behavioral consequences

Calibrate Quality Control Personnel

A good first step is to have some of your top management and quality control personnel evaluate samples and make quality determinations. This may include making decisions about such things as: (1) whether or not animals need treatment; (2) which apples need to be picked or left on the tree; (3) determining the grade of cherries in a packing shed (e.g., export quality, national quality, canning, discard); (4) making decisions about thinning so the ideal fruit load will remain; and (5) making correct vine training or pruning decisions. For the rest of the chapter, I will focus mainly on the strawberry packing shed, knowing that you will apply it to your own commodities—either in the field or the shed.

Calibration is the process through which top management, quality control personnel and, eventually, packing shed employees get to see the strawberry plants with the same eyes.

A group of about six to a dozen or so people sit around a table while a single strawberry plant is passed around. The assignment is for each person to examine the plant and make a mental decision. Should this plant be packed or not? The answer to this question mostly revolves around whether such a plant will grow once it is planted by the customer. Supply and demand affects these decisions.

When all have looked at the plant, the facilitator asks that participants show, at the same time, whether they would have packed or rejected the specimen. Everyone needs to simultaneously take a public stand as to not be affected by others in the group. Such as raising a two sided paddle and showing the green side for packing and the red side for discarding.

In an attempt to reduce the natural anxiety of taking such a public posture, it is vital that the facilitator assure everyone that it is normal to see lots of disagreement at first. Each individual is asked to explain and defend her stance. Conversation is one of the most vital activities in the calibration process.

With every plant that is discussed some individuals will yield to their colleagues' opinions while others will prevail. The decision needs to be a concensus.

At one packing shed everyone but one individual gave the green light for packing. The whole group erupted into a jovial laughter when this happened. After the round table conversation, however, all were convinced that the plant needed to be rejected.

It is vital to build in plenty of time for dialogue and debate. This procedure is repeated five or six times with new plants.

After everyone gets the general idea the facilitator can now move to a more rapid approach, setting 30 plants out on the table (each with a number) and having the participants fill out a pre-printed rating sheet with a decision

about whether or not each sample should be packed or discarded.

Raters find an open place by the table and begin with whatever plant number is available and then all rotate together either clockwise or counterclockwise throughout the exercise. It should be clear to participants, the first time, that the first plant they will rate may be number 1, or 9, or 23, etc. This is sometimes a source of rater error as some individuals begin with plant number 1 in the form regardless of what plant they rate first.

Similar conversations take place about these plants as with those which were passed around singly. As each sample is discussed and debated, it is important for these to be passed around so individuals can see them again and observe what others are saying. It is through this process that more precise standards are arrived at.

Some commodities and tasks are much harder to evaluate than others. But even with those that seem fairly simple and straight forward there seldom is much agreement at first, underscoring the importance of this process. For instance, at one avocado packing shed—despite the apparent ease of the decision-making process—there was only 22% agreement in the first round.

This procedure with 30 plants is carried out several times until people seem to have a general understanding of the process.

Refine Standards

Through the process we have outlined, rich conversations arise about the precise standards that need to be used in making quality-related decisions. People realize they have been operating without very precise parameters for a long time.

As clarity emerges, it is a good idea to take photos of plants that clearly show the difference between those that merit packing vs. being discarded. These photos can be posted or made into a video for employees and applicants.

Some types of plant damage are less obvious. Borderline cases are the most difficult and yet of vital importance. In one of most illuminating talks I have attended in my University career the speaker explained that farmers prepare the land, plant, irrigate and harvest (to name a few cultural practices) in order to get the benefit from the crop they have so carefully worked for. The cost of these procedures, including labor, go to pay for the crop. If the farmer can save even a small percentage of additional plants, these are pure profit.

Let me also say, of course, that it is not just these borderline plants that give people difficulty Some are not able to see even the most obvious problems while others tell me that even the plants 'talk' to them.

I am simply not able, as an example, to read a dog. I lean on my wife who

Calibration consists of having management and quality control personnel come to an agreement on subjective decisions, such as plants that should be packed or discarded. When finished recording their decisions, a conversation will take place about each plant where there was any disagreement.

has great skill. When I go out on farm visits and there are dogs present, I do not get out of the car until the farmer indicates it is safe to do so. On the other hand, when it comes to horses, I can hold my own. At the cross-country phase of a combined training equestrian event, I noticed, long before the horse arrived at a jump, that something was wrong. I told the person standing next to me, "that horse will not make it over the jump." The horse crashed through the obstacle and the rider spent weeks in the hospital.

What makes quality control somewhat challenging is that standards often change with different varieties and depending on a number of other factors—including market influences.

In some instances, farm managers may wish to invite trusted, proven, outside experts to help make quality determinations.

Establish Hierarchy of Reasons for Discards

So far, participants have mainly focused on making pack vs. discard decisions. In the process, the precise reasons for discarding have become clearer. Now the time has come for participants to also explicitly note the reason for discarding.

For instance, we may consider such defects as too few roots, lack of root hairs, cut crowns, crowns that are too thick or too thin, and mother plants. Your list may be different. Because each of these defects means we might want to discard the plant—and for ease in statistical analysis—it helps to create a hierarchy of plant damage.

One approach is to generally order the types of damage from those which are less frequently present to more common ones. Another approach might be to begin with the most serious issues

Individuals may be tested for their ability to make subjective decisions. While the manager points to each flower, the applicant will indicate his understanding of flower maturity. This greenhouse grower sells flower seeds.

SIDEBAR 11–1

Apple Packingshed Quality Control[1]

International fruit markets often demand high quality packing criteria that need to be carefully communicated to packingshed employees. Crew workers can also improve quality as they better understand quality parameters, and thus increase the number of packs per bin.

If management disagrees on which fruit meets high quality criteria, how can packingshed employees be expected to fare any better? This research report summarizes work carried out with apple sheds in Chile. Our objective was to measure and improve reliability as well as accuracy of the decisions of apple packingshed personnel—from management to packing staff.

METHODS

We chose apple packingsheds, where employees have to make many decisions quickly. Apples are harvested and brought to the packingsheds in bins and are subsequently sorted and packed according to quality standards. Packing decisions can be quite complex. This study attempted to identify those individuals who were able to make accurate packing decisions given a specific packing norm. Women are prevalently employed in this task, and were paid by the hour. The study involved several steps, including 1) definition of criteria, 2) verification of criteria, 3) personnel training, and 4) personnel testing.

Definition of criteria. The packingshed client or administrative team determined a given norm, in terms of what types of defects would be permitted in packed fruit.

Verification of criteria. Packing shed and quality control management participated in an exercise in which they had to evaluate small samples of apples (25 to 50 fruit per sample), in terms of whether or not each apple should be packed, taking into account the pre-determined criteria. Each member of the team was asked to evaluate the apples independently. After that, a conversation was facilitated in which differences in opinion where fomented.

The goal was to identify at least two management or quality control individuals who had a good eye (consistently obtained at least a 92% accuracy score). Two costly errors packing errors include: (1) discarding good quality fruit; and (2) packing bad quality fruit.

The average accuracy score for the three packingshed teams, each in turn composed of three individuals (thus, 9 individuals in all) increased from an 85% to a 95% accuracy through this process. It should be pointed out, however, that individuals who did not improve substantially were eliminated from the teams.

Worker training. Packingshed line workers received training in two steps: (1) a detailed explanation of the types of fruit damage and their causes; and (2) hands-on exercises in which participants would study, evaluate, and receive feedback on decisions made.

Formal testing. In each of the packing sheds, samples (100 to 150 apples each) were numbered and spread over several tables. Participants were given a sheet of paper and pencil in order to note their opinion as to whether each apple in the sample should be packed. When finished, they turned in their completed sheet to one of the researchers, and were given a new blank sheet in which to evaluate the next sample.

There were two samples per packing shed, and each subject was expected to evaluate each sample twice, resulting in four tests per individual. Subjects were given about 20 seconds per fruit, but these times were reduced as they felt comfortable with the testing process.

RESULTS

We found great variability in terms of people's abilities to make correct decisions. Subjects ranged from 95% to 68% accuracy

SIDEBAR 11–1 (CONTINUED)

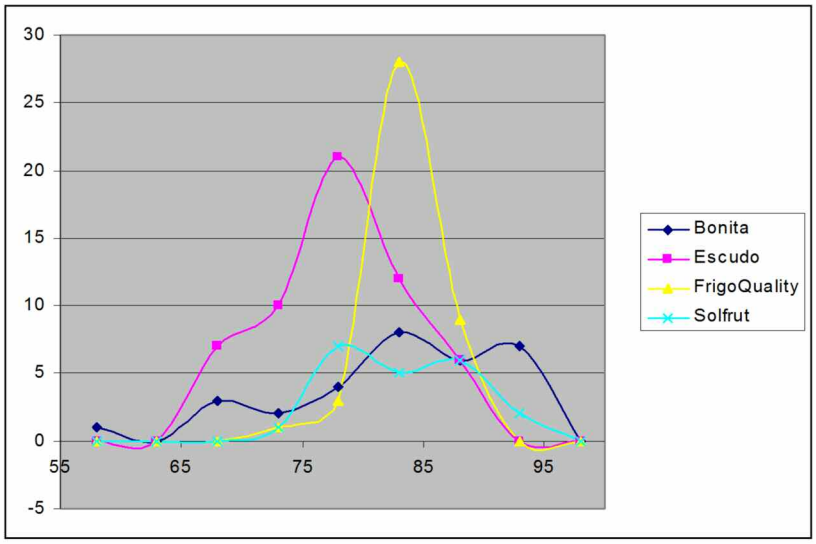
Apple Packingshed Quality Control

RESULTS

We found great variability in terms of people's abilities to make correct decisions. Subjects ranged from 95% to 68% accuracy (which compared to the California study, where individuals ranged from 95% to 59% accuracy). In the California study, participants were also rated in terms of their ability to correctly identify the reasons for making discard decisions. Such skills are particularly important in order to identify staff that can provide effective feedback to employees.

Individuals who are not too sure of themselves are more likely to throw away good fruit. Furthermore, those who are not too sure of themselves may be seen discarding fruit in a packing shed line in order to seem busy. But what fruit are they discarding?

We found that in some instances we had quality control personnel who scored worse than packing shed employees. Some of the benefits from this work include (1) improving communication among management team members; (2) once standards have been developed, more accurately conveying those to packing shed employees; (3) using this tool as both a selection and placement approach to increase accuracy. While we expect some employees to make major gains in terms of quality decisions they make, others will not be able to improve enough given the rate of speed required of them in making these subjective quality evaluations. Accurately identifying borderline fruit is likely to make an important difference in improving the bottom line. In field operations, these same factors need to be considered.



Curves showing highest to lowest accuracy scores at three apple sheds in Chile (Escudo, Frigo Quality, Solfrut) and at a strawberry nursery plant packingshed in California (Bonita) are shown. Given the small numbers of subjects per packingshed, the normal curves show some anomalies.

or a combination of both of these approaches.

One packing shed chose to begin with the cut crown, and follow that with lack of number of roots, mother plants, thick crowns, thin crowns, and lack of root hair. Once again, a pre-printed evaluation sheet can be created and photocopied to save time.

Quality control personnel are trained to always look first for a cut crown. If there is a cut crown they would note this and move to the next plant. Else, they would check for insufficient roots. If

there are not enough roots, they would mark this and move to the next plant, and so on. If, in the quality control inspection the individual gets to the last item, lack of root hair, this would be noted if that problem exists, otherwise the plant would be marked as a good specimen, one that should be packed.

Another way of explaining it would be saying that if a plant is marked as having a thin crown, it means that the plant did *not* have a cut crown, too few roots, was not a mother plant, nor did it have a thick crown. It may or may not lack root hairs, but for our purposes we do not care.

Test Quality Control Personnel

After standards are clear, quality control personnel may now be tested on both pack vs. reject decisions as well as on the reasons for discarding plants.

For strawberry plants, I like to set them out in a numbered sequence from 1 to 100 (or 1 to 150) on a set of tables. In addition, I like to group the samples into groups of five plants per station. So, the first station will have plants 1 through 5, the next will have plants 6 through 10, and so on. The reason for these clusters is to facilitate the testing process (as we shall soon explain).

If we have two sets of tables with 100 plants each, this permits us to test 20 individuals per set, or up to 40 persons at a time.

Before beginning, each individual will fill out the name section. The form also shows the letters A and B as well as the numbers 1 and 2. The examiner may pre-fill out the forms marking either the letter A or B, depending on which 100 plant set the person will evaluate first. The examiner may also mark the number on all of the forms. So, for instance, a form marked with the letter A and number 1 means that the rater will begin with the 100 plant set labeled A and the number 1 indicates the first pass.

Eventually, when A1 participants complete the test, they will be given a new test form appropriately marked and asked to move on to B1 (those who just completed B1 will move to A1). Then, after completing this second set, those who began with A1 will return to the A set with a form labeled A2 (meaning the second time they go through set A); while those who began with B1 will now do B2. In total, then, participants evaluate 200 plants twice, or make a total of 400 individual evaluations.

I prefer *not* to make it clear to participants that they will be evaluating the same plants twice as I do not want them to attempt to memorize their decisions. The large number of plants they will evaluate also helps to deter a memorization approach.

The reason why individuals are given five plants to evaluate at a time is in order to reduce the number of times they will have to change stations throughout the testing process.

Individuals will begin the process starting from a different station where they will evaluate all 5 plants in the cluster. In other words, some will begin with plants labeled 1-5 while others may start at plants labeled 6-10 or 11–15, and so on.

Individuals are to evaluate each of the 5 plants, one at a time, and carefully find the correct spot in the evaluation form to record the answers. Furthermore, they are told to turn around (and thus give their backs to the table) when they are finished recording their answers.

The test administrator then knows how many people have finished evaluating and recording their answers. I prefer to give participants, at first, plenty of time before having everyone move to the next station—and a new cluster of 5 plants. This is specially so because it is may be somewhat cahotic and confusing at first.

For the first few sets of five plants I might wait until everyone has finished and turned their backs to the tables before having them move on to the next set of plants. But after a few rounds I reduce the permitted time for making and recording the responses, until I get to the pre-determined expected time per round (such as 40 seconds per station). The forms should facilitate *marking* rather than *writing out* responses, in order not to delay participants and make

responses easy to read. Other than the participant's name, everything else can be checked off or circled.

Each time participants move from one set of five plants to the next, there will be some who will need to move from one table to another (several tables are needed to accommodate 100 samples). In order to avoid confusion, it is important to provide plenty of support personnel who are ready to guide people to the next set of plants at a different table, when this is not clear.

Also, in order to reduce further chaos and 'cooperative' work among participants, I make sure everyone moves from one station to the next at the same time. I give a signal with some noisemaker, such as a whistle.

Evaluating Quality Control Personnel

The process of evaluating quality control personnel begins early on. In the process of calibrating eyes so everyone sees the plants in the same ways, some individuals will influence others more than they will be influenced.

At one apple packing shed in Chile, for instance, one of the top managers tended to have an opinion that was different than those of the rest of the group, yet he would invariably recognize that others were right. This top manager was never able to argue his case.

Of course, as we already said, this is not an issue of majority rule. In another enterprise participants came to have a different—and very helpful—understanding of what the manager really wanted of them.

By keeping careful records of rater accuracy we can begin to separate the top raters from others. Hopefully, at least two individual will high consistency and accuracy will surface.

When conducting the 30 sample tests, I like to compare the results and identify any disagreements between these top performers. For instance, they may have agreed regarding all plants except for numbers 1, 5, 6, 12, 25 and 27. I give these individuals one blank rating form with these plants circled and ask them to go out and make a joint decision. They do so without the benefit of knowing how each had just evaluated these plant samples.

I am thus able to develop a key of correct answers against which to evaluate rater consistency (test reliability measure) and accuracy (test validity measure). I use this same approach with the 100-plant sample testing (sending the two or three best raters to re-evaluate where there are disagreements).

Using the Gage R&R instrument, quality control personnel may then be evaluated in terms of both their rater reliability (consistency with their own previous decisions) and rater validity (how accurately they performed compared to the answer key). We do this for both pack vs. discard decisions as well as for reason for discarding.

Gage R&R validity scores of 90 or above for pack vs. discard decisions show that quality control personnel have great promise; while validity scores in the mid-80s or above show great promise for determining the reasons why plants ought to be discarded.

Without exception, everyone who will work in quality control needs to be tested using this procedure.

Before leaving this section on testing, it is important to note that there exists great variablity in terms of ability to count total number of plants. As we said, ability to find plant defects is not correlated to ability to quickly and accurately ascertain the total number of plants in a bunch. Quality control personnel need to be tested separately for counting.

Train and Test Applicants

Both applicants and present employees need to be tested using the 100-plant procedure. This needs to be done on a regular basis in order to evaluate the quality control process.

It helps to develop a video of the mechanics of the testing procedure that applicants and other test takers can watch. My experience with both testing and interviewing (Chapters 2 and 3) is that it is easy to provide all of the details and do so with plenty of enthusiasm the

first few times around. But sooner rather than later it is hard to keep up the enthusiasm and thoroughness when explaining the testing process. Answering questions for applicants is also vital.

Some employers blow up photos with samples of plants that show defects or that are borderline in an effort to better communicate with test takers. One of my favorite approaches developed by one packing shed is to encourage test takers to grab plants from a pile, and when they find plants with each of the particular defects that require discarding that they bring it to the quality control person to check for accuracy and understanding. While these procedures extend the testing time, it greatly increases the chances that individuals will do well in the test.

When conducting the 100-plant tests it is important to reserve at least two spots for the top quality personnel to take the test and establish a correct key, as described above.

Test results may be used as selection and placement tools and may also be used to provide training and feedback. Some test takers may be identified as ideal for a quality control function while others may clearly make for excellent supervisors or plant sorters.

For instance, a sorter needs to be fast, reliable and accurate in making pack vs. discard decisions. It is not required for this individual to do well in the portion of the test where accuracy regarding the reasons for discarding is determined.

A top quality control individual needs to be able to quickly, reliably and accurately assess both pack vs. discard decisions as well as the reason for doing so.

A supervisor or runner (we will describe this position in a moment) needs to clearly articulate the reasons why plants need to be discarded in a way that sorters will understand. In addition, they need other communication, supervisory and coaching skills.

In general it is a good idea to encourage sorters to count in terms of bundles of no more than ten plants as they put together the 100 plants per bunch.

Establish Sample Procedure

I heard of a biochemistry professor who enlisted the help of his students to correct a test they had recently taken. Each student got to correct his or her own test, while the professor explained the correct answers.

The next week this same professor announced to the class "that there had been some irregularities" and some of the students had cheated by changing the answers to their test questions. He went on to say that he knew this because he had made photocopies of the tests before handing them out to the students.

The professor went on to say that those who turned themselves in would get an automatic fail grade for the class, but those who did not would be kicked out of the university. My favorite line in this story was when the professor was reported as saying: "By the way, I only made photocopies of some of the tests!"

Some farm supervisors spend their days frantically going from one crew to the next telling pickers not to pick certain fruit. The crew members all agree to change while the supervisor is present, but by the time the supervisor

When paying piece rate, quality incentives take more time to set up but have the greatest potential. Begin by identifying a range of acceptable individual performance. Then set up random quality-control inspections or spot checks. Sub-standard scores can result in additional training or discipline, while good marks earn employees an extra bonus per unit.

Gregorio Billikopf and Jim Kissler

makes the rounds through all the other crews and comes back to this crew, he has to start all over making it clear that he is not satisfied with the quality. I call this the babysitting approach to management.

Contrast this to the supervisor who makes sure that each fruit bin or box is clearly marked with the employees' numbers so at any time any bin or box may be inspected and the picker evaluated in terms of the quality of the work. And now add the fact that not every box is inspected, but only *some*, just as with our professor above. Effective management, in part, means there are clear rewards for excellent work and consequences for poor performance. And just as importantly, that supervisors do not have to inspect every single item.

The idea, then, is to establish a sampling procedure such that sorters never know when, or how many times, their work will be inspected. I like to make use of a random number generator (this can be done through a calculator or a spreadsheet.)

In Microsoft Excel, for instance, you can use F9 to generate random numbers where the number will be 10 or smaller, using the formula =RAND()*10. You can make that 100 or less or 12 or less (e.g., =RAND()*12, or some other number) by replacing that number in the formula. Make sure to format the cell number so you get 0 decimal places. You may also create a random number generator that has minimum as well as maximum numbers. This would be ideal, for instance, for generating inspection times that only fall in the workday.

With the help of the random number generator you can determine which employee will be inspected, at what time, or which of her boxes. It is possible, this way, that an employee gets inspected at 3:10 PM but this employee will never know if she will be inspected again in a few minutes or not at all for the rest of the day. Sometimes a particular employee will be inspected a few times in a day and at other multiple times.

Create a Feedback Mechanism

Now that we have discussed how samples are taken, let us discuss how these samples are evaluated and feedback is given.

As we discussed in the chapters on employee selection and performance appraisal (Chapters 2, 3, 6, 7) there is a huge danger that raters fall into either the halo or horns rating error.

Whenever possible, quality control personnel should not know who they are rating. There are numerous studies that have shown how raters are subject to influence in numerous ways including: (1) friendship with employees; (2) employee attractiveness; and (3) halo and horns effect, where raters want to be consistent with what they have observed in employees in the past.

Furthermore, it is so difficult to find accurate and reliable raters, that these individuals' time should not be spent moving around to find work to evaluate.

Instead, boxes labeled with bar codes, or some equivalent way of keeping the identity of the sorter completely anonymous, are brought to a rating station where the raters perform their duties, hopefully behind a curtain or in ways they cannot observe where the samples are collected from.

It is ideal, also, if at last two quality control personnel can be hired to do this rating and if bar codes are changed from time to time. These measures will help raters not try and double guess their own work.

A separate person needs to perform counting tests on samples rather than have the main quality control personnel spend their time doing this. Even if a person could do both, the time of the highly keen eyes for defective plants is too precious to have them do counting.

Pre-printed forms are used so that raters can clearly communicate not only a score, but also what steps are needed for employees to improve the quality of their work.

Train Supervisors and Runners

The role of *supervisors* is to deliver the completed evaluation forms to the sorters. It is just as important that

supervisors *celebrate what individuals are doing well* as it is to point out weaknessess and areas of needed improvement. Sorters should not ever feel, "Oh no, here comes my supervisor again!"

In fact, the supervisor acts as a mentor and coach to the sorters, encouraging continued positive behavior and taking the time to coach and instruct workers in those areas where there is a pattern of weakness.

Defective plants—which may not be packed anyway—can be set aside by the quality control raters for supervisors to share with sorters. Instead of the supervisor explaining to sorters what is wrong with these plants, the sorter may profitably be asked to first comment on them. People prefer to point out their own faults.

The ideal quality control program, as we shall see in the next subsection, includes rewards for consistent performance above certain quality thresholds and discipline for falling below minimum quality thresholds. An effective supervisor will be a coach and a motivator of people who is also not afraid to respectfully and positively discuss needed performance improvements (Chapter 21).

Runners are hired to collect random samples to deliver to quality control personnel for evaluation. It is also important for these individuals to be very positive and friendly as they collect samples from sorters. There is no need for runners to understand quality control themselves, although it does not hurt to have knowledgeable individuals who can work in other positions in the shed, when needed.

Based on the results of the evaluations, group training may also be conducted in order to help employees with common challenges.

Establish Behavioral Consequences

In traditional management schemes the supervisor's role is often to fight against the current. The employer and supervisor hope for good quality but the crew workers or packing shed employees are rewarded only for quantity of work. Generally, a piece-rate paid worker can make more money by ignoring quality. Thus the constant battle between what motivates the farm enterprise and what motivates the employee.

The general idea of a quality control program is to create standards above which employees will receive a quality bonus and another standard below which employees will be subject to discipline. With this systematic approach to quality control, we can reward people for achieving the same objectives that interest the farm enterprise or packing shed.

If we have selected individuals using the testing process we described earlier, then it means that people are capable of recognizing good quality. When people are not selected through a test, on the other hand, we do not know if they are not performing because they are not motivated or because they do not have the required abilities.

In the development of standards at a strawberry nursery, it is a good idea to speak in terms of errors permitted per 1000 plants. This can be done in terms of number of plants with cut crowns, thin crowns, and so on, or in a more general basis.

If employees excel in their performance, then the bonus is based on all the boxes packed for the day or the week, not just those boxes that were inspected. Whether bonuses are based on a daily or weekly performance, frequent feedback is necessary. One advantage of the daily bonus approach (even if it is paid on a weekly basis) is that employees are less likely to get discouraged if they had a bad day that could affect the whole week.

Another approach is to use a stepping ladder approach in which employees are rewarded incrementally as they reach higher and higher levels of quality performance. At least in theory it makes sense that once employees have reached the highest levels, if they are able to stay at these high levels of quality work, that the training wheels are removed, at least partially.

In other words, in order to motivate employees they may be given a small

When properly set up, a quality incentive changes the role of a supervisor to that of a coach who can help workers, and the farm enterprise, achieve their quality goals. When workers are only paid by the piece, with no quality bonus, foremen often feel they are swimming against the current as they try to enforce quality expectations.

bonus if they reach a first step in the quality ladder, and higher pay as they reach higher steps. Lower steps that are seldom needed may be eliminated with time.

The total percentage of the pay that goes toward a quality bonus has to be high enough to make the work towards achieving quality results worth the effort.

When employees are paid by the piece, the idea is that next time you want to increase the piece rate, to instead put that money towards the quality bonus. Or, once employees can see that they can achieve the quality levels you are rewarding, it may be possible to take a greater amount toward paying a quality bonus.

Whatever the quality bonus, it needs to be sufficiently large so that the same employee will make more money when achieving the quality bonus rather than producing more work without the bonus.

One packing shed manager told me that the day after the bonus pay was instituted, that productivity was reduced substantially. Over the next few days, productivity increased and soon employees were producing high quality work and being just as productive. What this tells me, at least partially, is that once employees get accustomed to doing the job right, that it does not take that much more additional effort to do so. Of course, some of these hypothesis need testing through carefully conducted research.

A cherry farmer may pay $3 per box picked with a potential multiplier of 1.084 for good quality or 1.25 for superior quality (about 25 or 75 cents per box, respectively). Three workers picking 24 boxes each in a day would earn $72 (no bonus), $78.05 for good work, and $90 for superior work.

Finally, there needs to be some sort of quality control and quality bonus for quality control employees.

Summary

Most types of agricultural work require both productivity and quality. Focusing on the strawberry nursery and apple industries, we saw how quality determinations should not be taken for granted. Creating a quality control process entails calibrating management and quality control personnel decision making and developing selection tests. We also spoke about providing feedback and paying for good quality, not just for productivity. The concepts can be generalized to pears, cherries and other fruits—both in the field and in packing sheds. We have also implemented them in animal agriculture.

Chapter 11 References

1. Billikopf, Gregorio; Pons, Macarena; Grant, Juan H.; and Muñoz, Pablo. Improving Fruit Sorting Accuracy. *Good Fruit Grower Magazine*, October 2007, pp. 20-21.
1. Billikopf Encina, Gregorio; Pons Porcile, Macarena; Grant Loyer, Juan Horacio; y Muñoz Vitta. “Control de calidad: packing de manzana”. *Revista Frutícola Copefrut S.A.*, Vol. 28, No 3, 2007, pp. 130-134. Chile.