By Anthony R. Montebello and Maureen Haga

To Justify Training, Test, Test Again

Despite a $34-million sales bonanza that made a training program at R. R. Donnelley & Sons Company look incredibly good, simple analysis showed that other factors may have deserved the credit. A series of statistical tests provided management with the answers it needed.

Getting budgets approved for training is hard today. Squeezed by unparalleled pressure for cost control and incessant demands for productivity gains, line managers are forced to make tough choices between HR-related programs, such as training, and alternative investments, such as automation. Often, they base their decisions on suppositions about what the different investments will yield.

This is quite a tenuous and uneasy situation for the HRD function and staff. Under increased demand for productivity-related development programs, training budgets are burgeoning faster than the rate of inflation—according to a recent *Training Magazine* survey, they increased in the past year by an average of 7%. When tangible dollar delivery expenses, such as travel, lodging and time off the job are factored in, total costs swelled by 20%. Ironically, these bloated training budgets are difficult to justify because the payoff of training more often than not is elusive.

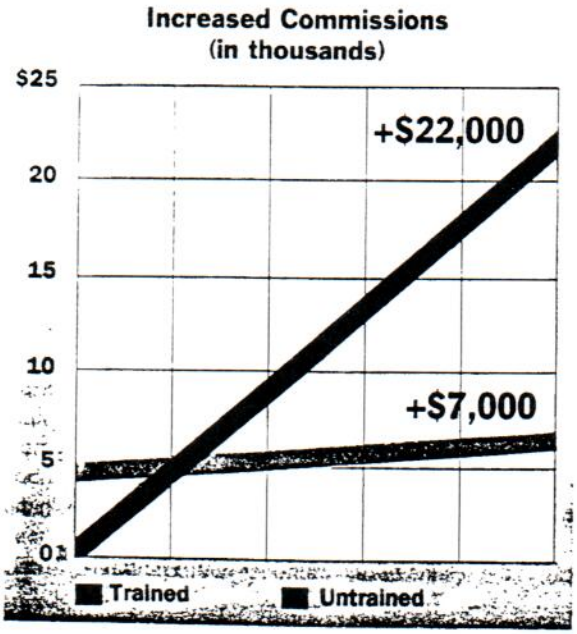
That's why, more than ever before, HRD professionals must evaluate the impact of training. Says William J. Healy, president of Iselin, New Jersey-based VMI Learning Systems, "Companies are frustrated by investment in training programs that don't feature measurement of results."

These measurements, though necessary, aren't easy. The training department at Chicago-based R.R. Donnelley & Sons, a commercial printing firm, found this out when it tracked and evaluated the results of a training program that took place in May 1990 for senior sales representatives. The training bill added up to more than $125,000, including the costs for instructor certification, seminar delivery and seminar participation. That doesn't count the opportunity cost of keeping the sales representatives away from the office for four and a half days. Senior executives and sales managers understandably wanted to be shown that the investment in both money and time was worth it.

In the end, the trainers proved it was. Along the way, however, they weren't so sure. Initial data suggesting the program had produced an eye-popping $34 million windfall proved to be inconclusive. Further analysis repeatedly showed that

Trained Group Outperforms Control Group

Factors such as gained experience and marketing campaigns contributed to an increase in commissions between two years for a group of sales representatives who received training and a group that didn't. However, commissions for the trained group far exceeded those of the untrained group, indicating that training had an impact.



what appeared to be impressive numbers actually were insignificant. Only through perseverance did the trainers succeed in their goal. Here's how they did it.

Donnelley determines measures to evaluate training results. The training department teamed up with St. Louis-based Psychological Associates, an international consulting and training firm, to develop a measurement system. The team had to look at three issues in designing the system, which were how to:

- Obtain relevant measures of training impact to evaluate its effectiveness
- Design the experiment to balance practical realities with scientific control
- Analyze the results to reach sound conclusions that easily can be communicated to sales supervisors and senior managers.

The first step for the team was to figure out what measures to use in evaluating the results. To do this, it had to consider measurement validity, which refers to whether or not the measurements chosen focus on the right thing. In training evaluation, the right thing is the behavior or result the trainers are attempting to change or influence. For example, if a company develops training to change attitudes toward minorities, observations of service representatives' behavior toward minorities would be more valid than transactions per hour.

Reliability was a related concern for the team. Measures are considered reliable if they consistently measure the variable of interest. A yardstick, for example, is a much more reliable measure of height than subjective judgments of people eyeballing an object. Customer observations of whether a clerk smiled or used a customer's name probably are more reliable than supervisory ratings of customer courtesy or bedside manner. Objective measures, such as sales or production figures, usually are the most reliable.

A third consideration was the timing of the measurement. Often, measures of results aren't possible until several months after training. The longer the time span, the more likely it is that other events, such as a downturn in the business, turnover or additional training, will have an impact on the measure selected. Other considerations in selecting a measure of change have to do with practical considerations, such as cost, availability, accessibility and so on.

The team determined that relevant measures for evaluating the effectiveness of Donnelley's sales training would be behavioral changes and the business results brought about by these changes. The team asked itself, "What is the new behavior that we want to see?" The answer was that it wanted to see if people could close new business as a result of training. Therefore, closure ratios were the measure the team would be using. (See Table 1, this page, for other possible measures.)

ing session took place in May 1990. Twenty-six sales representatives from nine different product groups attended the four-and-a-half day session. The participants all were senior-level, commission-only workers with a minimum of three years' selling experience at the company.

Before the sessions began, the trainers asked each of the participants to prepare background on a real case. The case needed to involve a prospect whose sale the participant couldn't close. It also had to be a case with which the participant had been working for a while so that a meaningful role-play simulation could be

Measures of Behavioral Change and Business Results

To evaluate whether its sales training course had any impact, R.R. Donnelley & Sons considered the following measures of behavioral change and business results.

Behavioral Change Measures:	Business Results Measures:
• Number of calls	• New business development
• Closure ratio	• Sales volume
• Customer evaluations	• Cost of sale
• Manager evaluations	• Average order size
• Peer evaluations	• Add on sales
• Percent of objections managed	• Commissions
• Product and service mix sold	• Market penetration

Table 1

Next step: Developing a measurement system. To help determine what measure to use, the team investigated the four basic testing designs, which are derived from combinations of pre- and post-testing of either one or two test groups. In the two-group designs, one group is an untrained control. (See Table 2, opposite page, for an explanation of the four designs.)

The first evaluation method in which the training team engaged was measuring the results of one trained group. The train-

constructed during the training program. In addition, it had to represent significant new business potential that realistically could be pursued soon after the seminar: It had to be challenging but attainable.

During the seminar, participants honed their selling skills by participating in sales simulations that were videotaped and critiqued by a team of peers. They learned about four personality types and how to handle clients who have each one. Before completing the seminar, each participant had to develop a strategic plan for closing

the deal in his or her case study. The trainers kept copies of these plans for follow-up purposes.

Using real cases in the training proved beneficial for at least three reasons. First, it increased participant motivation by focusing skills on an important, real-world issue. Second, by encouraging immediate implementation of learned strategies, it minimized the possibility of problems with the transfer of training. Finally, having real sales situations provided the trainers with a quick and dirty means to evaluate training transfer, because follow-up on *hit rate* and total sales generated is straightforward and relatively easy.

Trainers checked on the application of the participants' strategies beginning six months after the training was completed and then every six weeks for a 12-month period. This time frame paralleled the typical cycle for closing the commercial-printing company's large and complex sales. Most participants needed nine to 12 months to execute their strategies. The results of the follow-up evaluations? Of the 26 participants tracked, 14 closed their sales within 12 months, accounting for $34,239,000 in new revenue. The average sale was nearly $2.5 million, with a range spanning from $35,000 to $15 million.

To compute the cost-benefit of the training, the team divided the revenue produced by the cost of the training. The resulting ratio, 273:1, shows that for each dollar spent, the company realized a $273 return.

Although the results sound good, there's a problem with them. By measuring only the revenue produced by the trained group after the training, there's nothing with which to compare the results. Did performance behaviors or results improve, get worse or stay the same? It isn't clear.

There is some merit to this testing method, however, as R. R. Donnelley found. It allowed the training team to gauge the level of transfer of training. This is important because if there's no transfer, there are no results. However, this wasn't enough for the company. It needed more evidence, which required a more sophisticated testing design.

By applying the one group, pre- and post-training design method, the team was able to add a baseline for comparative purposes, adding a measure of integrity. It chose commission data of new business for its objective measure. The team members determined that commission data for the year prior to and the year following the training would provide a reliable, valid and relatively short-term measure of performance.

The team found that the trained group increased its commissions by an average of approximately $22,000 from one year to the next. These results are more compelling than those found in the first test, but there's still a problem. This data doesn't provide enough information to rule out other reasons for the increase. For example, any of the following could have affected the commissions data:

- Other events, such as price decreases, promotional events, decreases in competition or incentive programs
- Natural changes in trainees, such as aging, increased experience, increased knowledge or personality changes
- Special or preferential treatment, such as extra attention from a manager, additional follow-up training, special coaching or contacts with other trainees

Without a comparison group, it's impossible to tell whether the increase in commissions was a result of the training or the other factors.

Comparing the after-training commissions data of two groups—one trained and the other untrained—provided some insight. The training team established a control group by searching the sales organization database and randomly selecting people from each of Donnelley's nine product groups represented in the trained group. The people selected for the control group met the same experience criteria as those in the trained group.

The team compared the two groups on pre-test commission means using a *t*-test statistical analysis (see "How to Measure Results Statistically," page 87). Although the untrained group was earning slightly higher commission on average (approximately $3,000) than the group earmarked for training, the test indicated that the difference wasn't statistically significant. The team concluded that the groups were equivalent before the training.

This determination was important. Unless the two groups are equivalent

Four Common Experimental Designs

Following is a list of four commonly used experimental designs, arranged in increasing levels of sophistication. R. R. Donnelley compared the designs for how well each could evaluate whether training alone caused the effect.

(Level of Sophistication)	NUMBER OF GROUPS	WHEN MEASURES ARE TAKEN
HIGH	**Two Groups** (Trained and untrained)	**Pre- and Post-test** (Measure before and after training; test for differences on pre-test and post-test measures or compute gain scores.)
	Two Groups (Trained and untrained)	**Post-test** (Measure after training for trained and untrained groups.)
	One Group (Trained)	**Pre- and Post-test** (Measure before and after training for trained group.)
LOW	**One Group** (Trained)	**Post-test only** (Measure after training.)

Table 2

before training, post-training data can be misleading. For example, the untrained group may be outperforming the trained group before the training. Although the training may improve the performance of the trained group, bringing it up to the level of the control group, the post-test data indicates that the training had no effect, because both groups are performing equally.

On the other hand, the trained group may have already been outperforming the control group before going through the sales training. When the post-test data shows the trained group outperforming the control group, it may erroneously be concluded that the training was responsible. In reality, the training had no effect.

What R. R. Donnelley found by comparing post-training mean commissions was that the trained group had surpassed the control group by more than $12,000. However, the *t*-test analysis indicated that even this difference wasn't statistically significant.

The team had three choices. It could abandon the project, assuming that the training had no effect. It could increase the sample size, assuming that the small sample used didn't allow the *t*-test to detect the differences that actually were there. Or, it could select a more powerful statistical tool. It chose the latter.

VITALS

ORGANIZATION
R.R. Donnelley & Sons Company

TYPE OF BUSINESS
Commercial printing firm

HEADQUARTERS
Chicago

EMPLOYEES
33,000

CORPORATE SALES TRAINER
Maureen Haga

YOU SHOULD KNOW
R. R. Donnelley & Sons Company measured the impact of a sales training course and discovered that it yielded a return-on-investment of nearly $13 million.

The most sophisticated of the tools presented earlier is two-group, pre- and post-training design. This method requires the calculation of gain scores, which are the differences between pre-training and post-training figures. To find the gain scores, the team compared the gain in commissions from the year before the training to the year following the training for each group.

The team determined that the average gain in commission that the untrained control group demonstrated—approximately $7,000—was an effect of history (a marketing campaign) and/or maturation (gained experience). But these factors affected both groups equally. In addition, the team monitored treatment effects to ensure that factors such as special coaching by the sales managers didn't differentially affect the trained group. Moreover, all alternative explanations were ruled out, one by one, to show definitive proof that training caused the increase.

Armed with this evidence, Donnelley's HRD team could claim an average

Tracking By Hand

Looking back at the tracking process used to evaluate the impact of sales training, there are some things that Maureen Haga, corporate sales trainer for Chicago-based R. R. Donnelley & Sons, would change. For one, she says she would somehow automate the process. Haga taught three four-and-a-half-day training courses between May and October 1990, in which approximately 65 sales representatives participated. She tracked the progress of each of these individuals for 18 months after they completed their training. (See the main story for the results of tracking the first group of 26 trainees.)

Tracking consisted of both phone conversations and written memos to each of the participants every six weeks, starting six months after their training sessions.

For each individual, Haga had to generate a file consisting of:

- The sales representative's name and the date on which he or she completed the sales training course
- A customer-description questionnaire that the training participants completed on a hard-to-sell customer before the training
- A sales-strategy planning guide for the same hard-to-sell customer, which the participants completed on the last day of their training
- The progress of each participant on his or her sales strategy during an 18-month period. (If a participant left the company or was pulled off the case being tracked, Haga had to remove that person from the study.)
- The date the participants closed their sales and the results of the sale. (For those who didn't meet their goals by the end of 18 months, Haga made a note in their file to this fact.)

In addition to all of this information, Haga had to keep files on new-sales commission data, not only before and after training for the 65 individuals who participated, but also for 65 sales representatives who served as a control group. (Acquiring the commissions data was actually the biggest challenge Haga encountered during the process, due to its confidentiality.) All of this information Haga collected and recorded manually. "It was a nightmare," says Haga.

Just keeping track of all of the information was difficult enough. Having a lot of irrelevant information made tracking results even harder. For example, "In the customer-description questionnaire, a lot of the material that [the participants] gave me was more nice-to-know than need-to-know information," says Haga. "I had to sift through a lot of data to extract what I needed." If she were to go through the process again, Haga says she would ask for more specific information that's directly aligned to the goal.

— Dawn Anfuso

$15,000 gain in commissions attributable to training (the trained group's $22,000 increase minus $7,000 attributed to history and maturation). Now, the *t*-test ratio indicated that the team could be 95% confident that the differences detected were true differences and not simply a result of how the sample was drawn from the population.

Just to be sure of its results, the team performed an Analysis of Covariance (ANCOVA) statistical test, an even more powerful and sophisticated measure. The team used ANCOVA procedures for partialing out or minimizing any pre-test differences. The resulting ratio demonstrated that post-test commissions were significantly different when the pre-test differences were controlled.

Unfortunately, it's nearly impossible to explain ANCOVA in any detail to a layman. The team was unable to use this information to make a simple and straightforward presentation to line management.

Demonstrating the impact of training. Although the statistical results were complex, the data that the team acquired provided a means for the team to perform a return-on-investment (ROI) analysis. This provided an effective means of communicating the impact of training.

Here's what the team did: By multiplying the number of participants in the training session (26) by their average gain in commissions ($22,422), the team calculated the total gain in commissions for the trained group as $582,972. The total gain in commissions for the untrained group from the year before the training to the following year was $191,490, which is the number of sales representatives in the control group (26) multiplied by their average gain in commissions ($7,365). Subtracting the gain of the control group ($191,490) from the gain of the trained group ($582,972) gave the team the gain in commissions produced by the training, which was $391,482.

How to Measure Results Statistically

There are two considerations to keep in mind when selecting statistical procedures: simplicity and power. The most important advice one can offer on the use of statistical procedures is this: Keep it simple. Explaining the results of a complicated statistical test to a layman can be a trying experience.

The other consideration in the choice of statistical procedures is *power*, defined as the sensitivity to detect group differences when in fact they exist. Among other things affecting power is the size of the sample and the choice of the statistical method. Often in evaluation research, the tester must deal with small samples. The *t*-test is most useful for comparing groups when the samples are small—30 or fewer people. However, if you have fewer than 10 or 15 people per group, the power diminishes to the point that it may be better to abandon the entire evaluation.

To measure the impact of its sales-training course, Chicago-based R.R. Donnelley & Sons used the *t*-test. The *t*-test is used to test whether two means or averages, such as the average productivity measures of the individuals within a trained group and an untrained group, are statistically significantly different.

Computational procedures result in a *t* value and associated probability level (p-value). A probability level of .05 or less indicates that the two means are indifferent. More specifically, it means that the testers can be 95% confident that the differences are true differences and not simply a result of how the samples were drawn from the population.

If you hypothesize a direction of the difference between two means (such as, mean A is larger than mean B), then a one-tailed *t*-test is used. This test is more sensitive and powerful for detecting differences than a two-tailed *t*-test. This is because a p-value of .05 (95% confidence level) is divided among the two tails of the distribution of scores around the mean. As a result, the mean of the comparison group must fall within either .025 region (above or below the mean) of the distribution rather than a larger .05 area when a one-tailed test is used.

Computational procedures for the *t*-test are available in any introductory statistics text and on commercially available software programs, such as *Microsoft Excel* and *Statistical Package for the Social Sciences*. —*ARM & MH*

This wasn't enough, however. The team needed sales revenue figures to satisfy top management. Commission for the sales representatives were based on an average of 3% of total sales. Because the team had determined that the total commission revenue gained as a result of training was $391,482, they were able to figure out that the total revenue produced as a result of training was $13,049,433. Subtracting the costs of training ($125,138) from the total revenue produced gave the team a return-on-investment figure of $12,924,295.

The team communicated these results to sales managers and upper executives through presentations at management meetings, using both statistical results and return-on-investment figures. Response has been favorable.

The success of the process has opened the door to future tracking of other high-investment training processes. In addition, it has convinced the sales managers of the benefit of training. They now have empirical evidence that allowing their sales representatives to spend some time for analysis and preparation yields higher results in the long-term than just relying on reactive methods.

The demonstrated results of training have had an unexpected benefit as well. Seeing the positive outcome of the training has persuaded other sales representatives to participate in a similar training course. There are now waiting lists for programs that once required a sales job by the trainers to enlist participants.

Through its efforts, R. R. Donnelley & Sons Company discovered that training does indeed pay. ■

Anthony R. Montebello is vice president of St. Louis-based Psychological Associates.

Maureen Haga is the corporate sales trainer for Chicago-based R. R. Donnelley & Sons Company.

For information on ordering reprints of this article, please see page 12.