

CARNEGIE LEARNING

GUIDE TO INTERPRETING EVALUATIONS

Carnegie Learning's Cognitive Tutor Results Reports present data that we have collected and received about the effect that our Cognitive Tutors have had on student performance. Understanding and interpreting these reports (and, indeed, any results report) can be challenging. This short guide is intended to help.

DOES IT WORK?

The fundamental question researchers try to answer when evaluating an educational initiative is "Does the program work?" At first glance, this may seem to be a simple question, equivalent to "Do students learn from this program?" But more careful consideration reveals the complexity of this question. It is not enough to ask whether students learned something from an educational experience. That is, at best, a minimum expectation. In fact, it would be surprising to find an educational initiative that was remotely worthwhile which couldn't show some student learning, especially over a full school year.

Instead of simply asking "Does the program work?", well-designed research tries to answer the question "Does the program work better than what students would have otherwise done?" In some settings, this amounts to asking whether, in the time allotted to the program, students were able to achieve at a higher level than students spending an equivalent amount of time using another program. In other settings, the question might be whether students were able to reach a given level of performance in a shorter time than students using a different program.

These more sophisticated questions require that we compare students' performance to something. In Carnegie Learning's Results Reports, we talk about three different kinds of comparisons: control groups, equivalent comparisons and expectations.



Control Groups

The gold standard in evaluation is comparison to a control group. In such a study, students who are participating in the program (the “experimental group”) are compared to those who are not (the “control group”). In our studies, students using the Cognitive Tutor program are considered the “experimental group.”

The logic behind a control group study is that, if the only difference between the two groups is the form of instruction that they are given, then any differences in achievement between the groups must be due to the form of instruction. Many of the procedural methods used in control group studies are designed to ensure that the two groups start out as nearly equal as possible and are not influenced by anything other than the curriculum they are given.

Perhaps the most important consideration in establishing the comparability of two groups in a control-group study is their initial knowledge and ability. In some studies, the two groups are given a pretest, to show that students in the two groups start with roughly the same level of knowledge. A similar method is to show that, as a whole, students in the control and experimental group received equivalent grades in previous classes.

If the groups are found to be different in their initial ability (or other identifiable factor), statistical methods can be used to determine how much of the difference in final ability is attributable to differences in initial ability.

Even if students are matched on initial knowledge, we can never be absolutely sure that the comparison is “fair.” Lots of things might be different between the groups. One group might have a more experienced teacher; one group might have more supportive parents; one group might be hit with the flu, causing more absences.

As a practical matter, it is impossible to measure and account for all possible differences between groups. For that reason, the best control groups are those that are most similar to the students in the experimental group. Students in the same class or in a different class taught by the same teacher make an excellent control group. Those options are rarely practical, though, so students in a different class in the same school or a different class in the same district may be used.

Random assignment is another way to try and ensure that the groups are comparable. If students are taken from the same large group and split into two groups (control and experimental) at random, then we may have some confidence that students do not differ on some factor related to success.

If the students being studied are in a school, though, it is rarely practical to randomly assign students to groups. Typically, it is enough that students were assigned to groups based on factors unrelated to the study (like school scheduling considerations).



1200 Penn Avenue; Suite 150; Pittsburgh, PA 15222;
1-888-244-7569; www.carnegielearning.com

Finally, the larger the number of students in the study, the more confident we can be that the results are intentional and not due to some unintended factor influencing one group but not the other. Statistical methods can be employed to tell us how many students are enough and how large a difference between the groups is a reliable difference.

In Carnegie Learning's research reports, studies using formal control groups are indicated as using a *matched control group*.

Equivalent Comparisons

Although control groups provide the most reliable information about the effectiveness of an educational intervention, studies that use formal control groups can be difficult and expensive to conduct. In some cases, when we have data on the performance of students in a program, we can identify a reasonable comparison. For example, we might be able to compare student performance to that of students in a previous school year or to performance on national or state exams. Since the researchers do not have access to the comparison group in this kind of study (unlike in a control group study), there is a greater potential that we may be misled by a false comparison. Factors other than the introduction of the Cognitive Tutor program may account for different student performance from one year to the next, for example. The challenge in this kind of study is to eliminate such sources of error and ensure that the results that are reported are really due to the influence of the Cognitive Tutor program.

In Carnegie Learning's research reports, studies using equivalent comparisons are indicated as using a *comparison group*.

Comparisons to Expectations

In many cases, schools will collect data without having an appropriate control group or even an equivalent comparison. Typically, such reports are accompanied by indications that the school is pleased (or displeased) with the students' results. While there is no explicitly identified comparison group in this kind of data, these studies make an implicit comparison to people's expectations. This constitutes the least reliable kind of data, since we may have originally set our expectations too low (or too high), but it can still be useful information.

In Carnegie Learning's research reports, studies comparing student data to expectations are indicated as *no comparison group* studies.



STATISTICS

Even in the most carefully designed and implemented study, there is always a possibility that the results are due to chance. There is always the possibility that something unrelated to the educational program being studied might have led one group to do better than the other. One of the purposes of statistics is to tell us how likely it is that the results that were found would be found again, if we repeated the same study.

There are two basic factors that the statistical methods consider: the size of the difference between the groups and the number of students in the study. The bigger the difference between the groups, the more likely that the difference is reliable. The greater the number of students in the study, the more likely that the difference is reliable. These factors are summarized in the *p* value, which represents the probability that, if the study were run again, the general result (that the two groups are different) would still be found. The *p* value does not say that the size of the difference would be the same, if the study were run again, only that some difference would be found.

In typical studies, a threshold *p* value of .05 is used. A *p* value of .05 says that there is only a 5% probability that the study results were due to chance. If the statistical analysis shows that the *p* value is less than .05, then the results are said to be *statistically significant*.

A related concept is the *confidence interval*. A study might report that, with 95% confidence, the difference between two groups is within a certain range. For example, the study might be able to say, with 95% confidence, that the experimental group is anywhere between 10% and 80% better than the control group. If the confidence interval does not include 0, then the results are statistically significant.

It is important to understand that statistical significance is not the same as practical significance. Statistical significance is dependent on the number of students used in the study. It is possible for a small difference to be statistically significant, if a large number of students are involved in the study. Conversely, a study might find a large difference between groups but still not find statistical significance, if the number of students in the study is small.

For this reason, many studies report *effect size*, in addition to statistical significance. Effect size can tell you the size of the difference between the groups in a study, in a way that allows that difference to be compared across studies (even studies that use different measures of achievement).

A widely-quoted study by Bloom (1984) found that expert human tutors can produce an effect size of 2.0, as compared to traditional classroom instruction. In summarizing 97 studies on the effects of computer-based instruction (not including studies using the Cognitive Tutor) Kulik (1994) found an average effect size of 0.32.



MEASURES

Every research study measures something. In education, this usually takes the form of an exam, but sometimes the measure is something like a teacher report or the percentage of students passing a course. When you evaluate a research report, you should pay attention to what was measured before deciding if the results are meaningful. Different measures are useful for different things. When deciding how much you value a particular study, you should consider the following:

Alignment

A primary consideration is whether an assessment reflects the course that was taught. No test (or other measure) is truly comprehensive and perfectly aligned with curricular goals, so it is important to know which topics, skills and abilities were emphasized in the test and which were neglected.

There are negatives to strict alignment, however. If curriculum designers (or test designers) are overly concerned with aligning the course and the test, this can result in “teaching to the test.” The danger here is that the course objectives narrow to match those that are easily tested. Also, the course may focus on test-taking strategies that help improve performance but do not help students with the content you expect them to be learning.

In a study involving a control group, it may be difficult to identify an assessment that reflects the curricular goals of both the experimental and the control group. To address this issue, some studies ask teachers to rate the relevance of each test question to the course they taught, and performance is evaluated relative to these ratings. Another way to address this problem (and the one of teaching to the test) is to use multiple measures of performance.

Standardized or specially-constructed exams

Standardized tests (like the SAT) have the advantage that they are widely recognized and are designed to vary little from year to year. On the other hand, such tests are often limited to multiple choice (or other formats that are easily scored) and they may not be well suited to measuring some kinds of student performance.

Specially-constructed exams can be well-aligned with curriculum and tailored to test different performance measures, but it can be hard to know what knowledge students need to succeed on the test without seeing it yourself.



Objectivity

Some measures are objective, meaning that different people will come to the same conclusion about the student's level of knowledge. Multiple-choice questions are an example. Other measures are more subjective. Course grades, for example, might be biased by a teacher's feelings about a student or by the value the teacher places on being in a research study. Although essays and performance assessments are somewhat subjective, a well-designed and well-implemented scoring rubric can yield a fairly objective measure.

Fixed or constructed response questions

Different types of exam questions measure different kinds of knowledge, so it is important to pay attention to the types of questions used on an exam. While it is not true that multiple-choice questions invariably test memorization of facts, it is difficult to design such questions so that they tap higher-order reasoning skills. Performance assessments, where students are asked to solve complex, multiple-step problems, can be a better way to assess students' problem solving skills.

CONCLUSION

No research study is perfect, and scientists disagree all the time about whether a flaw in a particular study is a fatal one (invalidating the results) or a minor concession to practicality. This document is intended to point out some of the factors to consider in evaluating research results, but the final decision about whether you believe that the claimed results are "real" rests with your judgment about whether the study was sound and is supported by other studies showing the same result.

We believe that research shows that the Cognitive Tutor programs are the most effective solutions available, but we ask that you not take our word for it. Read the research reports and see if you agree with that conclusion. If you have further questions, feel free to contact Carnegie Learning at researchpartner@carnegielearning.com.

REFERENCES

- Bloom, B.S. (1984). The 2 Sigma Problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13, 3-15.
- Kulik, J. A. (1994). Meta-Analytic studies of findings on computer-based instruction. In E. L. Baker and H. F. O'Neil (Eds.), *Technology Assessment in Education and Training* (pp. 9-33). Hillsdale, NJ: Lawrence Erlbaum Associates.

