How to Evaluate a Worker T&D System: Measuring its Effect(s)

Moses Rivera, Ph.D.

August 4, 2025

Downloads

- PDF version of these lecture notes.

Beyond the *Lecture Notes*: Insights and Elaborations

- In this video, I don't summarize these lecture notes. Rather, I provide related insights and elaborations.

[video]

Introduction

- When designing a T&D system, the best first step is to start with the end in mind. What **outcomes** are desired from this T&D system? And how will those outcomes be **measured and evaluated**? That is the focus of this lecture.

Measuring the effect of the T&D system:

Fundamentally, a T&D system is an attempt at **manipulation**—we are trying to **change the workers**. Thus, you can invoke all of the concepts that you learned in your prior research methods courses, especially the topics about **experiments** and quantifying the **effect of some variable**.

If you feel a bit rusty on that, don't worry! This may be a good time to revisit your notes and materials from those courses. In a nutshell, the topics I want to focus on here are:

- psychometrics;
- quantifying the effect(s) of the T&D system.

Below, I will expand on those three topics.

Psychometrics

- **Psychometrics** comes from the words *psychology* + *metric* (measurement). It is:

  > "the branch of psychology concerned with the quantification and measurement of mental attributes, behavior, performance, and the like, as well as with the design, analysis, and improvement of the tests, questionnaires, and other instruments used in such measurement. Also called *psychometric psychology*; *psychometry*." (online APA Dictionary of Psychology, 2025)

  - Thus, in a nutshell, psychometrics is about measuring psychological features. Common examples include: beliefs, cognitive abilities, extroversion, conscientiousness, knowledge, other behaviors, etc.

- In the context of evaluating a T&D system, we would typically measure people's attitudes, beliefs, knowledge, skills, and other outcomes. However, we must use **valid measurements** of those concepts, otherwise our inferences and conclusions may be incorrect.

  - For example, imagine if I want to see whether the T&D system resulted in an increase in the workers' knowledge, but I measure their knowledge via an online multiple-choice test that the workers can do at home, and then most of the workers use artificial intelligence to cheat on the test. In that case, my measure of their knowledge would be invalid, and it would be invalid to use that data to conclude that the T&D system increased the workers' knowledge.

- Part of what makes a measurement valid is when the measurement is reliable. The **reliability** of a measurement procedure is the extent to which that procedure makes consistent measurements. For example, if I measure your height today, and then again tomorrow, my measurement procedure would be *unreliable* if it says you are 6 feet tall today and 5 feet tall tomorrow.

  - **All valid measurements are reliable, but not all reliable measurements are valid**. In other words, measurement validity *implies* measurement reliability; but measurement reliability *doesn't imply* measurement validity. For example, I could use a *Harry Potter House Quiz* to try to measure your intelligence, and the quiz can be *highly reliable* (i.e., producing consistent results every time I measure you), but it is *not* a valid measure of intelligence.

- There are many techniques you could use to quantify the the validity and/or reliability of your measurement procedures. Those are beyond the scope of this course, but they are part of the psychometrics course of this program.

Quantifying the effect(s) of the T&D system, a.k.a. Calculating the difference that it makes

- The **effect** of a cause is the change that the cause creates. Imagine you implement a T&D system on some people, and it causes a change in those people. The **change is a difference** between what things were like before and after the cause. That change (*without* including any other changes that were caused by *other factors* unrelated to the T&D system), is the effect of the T&D system.
  - Although this is straightforward in principle, it's a bit more nuanced in reality. For example, you can't just measure everyone's performance before the training and again after the training, because there are an **infinite number of factors that could have also affected their performance**:
    - Imagine the workers started using artificial intelligence soon after the pre-training measurement. (This is called a **history effect**.) In that case, we wouldn't know whether (nor how much) their performance increase was due to the training versus due to the artificial intelligence.
    - Imagine the workers just needed more time to get accustomed to their jobs, and they were already going to achieve the higher levels of performance without the training. (This is called a **maturation effect**.) In that case, we wouldn't know whether (nor how much) the training helped.
    - Imagine the company's beloved CEO resigned, and consequently all the workers' performance was down that week because they were all in a down mood. (This is also called a **history effect**.) Then, the pre-training measurements were taken that week. Several weeks later, everyone naturally started feeling back to normal, and their performance went up. Then, everyone was trained, and their post-training performance was measured. In that case, we wouldn't know whether (nor how much) their performance increase was due to them feeling back to normal after the bad news, versus being due to the training.
- To explain how we can get around this issue and actually measure the true effect of a T&D system, let me tell you a useful thought-experiment. Imagine you magically could teleport back and forth to a duplicate universe where everything was 100% identical to your home universe. In other words, Universe 1 (your home universe) is identical to Universe 2, down to the smallest atoms and grains of sand. The only way you can tell them apart is because you see an exact duplicate of you in Universe 2. Now, imagine you teleport to Universe 2 and convince your clone to not implement the T&D system at ABC Company in their universe. Then, you go back to your universe, and you implement the T&D system at ABC Company in *your* universe. At this point, the *only difference* between Universe 1 and Universe 2 is the fact that the T&D system was implemented in Universe

1, but not in Universe 2. Then, you request a report from ABC Company about the worker's performance in your universe (Universe 1), and you teleport to Universe 2 and convince your clone to request the same report from ABC Company in their universe. Now, at this point, the resulting difference in performance between Universe 1 versus Universe 2 is truly caused by the T&D system.

  - In that scenario, we don't have to worry about whether the workers used artificial intelligence to increase their performance, nor whether their performance increased because they naturally got accustomed to their jobs, nor whether they just started feeling better after their beloved CEO resigned. Although all of those factors do impact everyone's performance, the impact is identical between Universe 1 and Universe 2. Thus, Universe 1 and Universe 2 would both have experienced the same enhancements from A.I., the same enhancements from naturally getting accustomed to the job, and the same enhancements from feeling better after the CEO resigned. Therefore, after we implement the T&D system in only Universe 1, if we then see any difference in performance between Universe 1 and Universe 2, it is because of the only difference between the two universes—which was the T&D system. The lesson here is that **the identical universes allowed us to measure what the workers' performance *would have been*** if we didn't do the T&D system. We did that by **comparing two groups that are identical except for one difference: the presence of the T&D system**. In that way, we were able to guarantee that any subsequent changes in performance were actually due to the T&D system.

Creating nearly identical groups

- Because we don't knowingly have access to duplicate universes to do the comparisons to calculate an effect, we must rely on **the next best thing: comparing two groups that are *nearly* identical except for the one thing we are manipulating** (in our case, the one thing we are manipulating is the T&D system). Many scientists call this type of study a **randomized experiment**.
- A simple way to help create nearly identical groups is by **randomly assigning** each person to a group. The beneficial effect of random sampling will also be increased if you use larger groups (this is because of a mathematical fact known as the *Law of Large Numbers*, but that is beyond the scope of this lecture).
  - Notice: this is *not* the same as randomly sampling workers (from a population, e.g., from your company) to be included in your experiment. **Random sampling** helps your sample be representative of your target population, whereas **random assignment to a group** will help your groups be equivalent to each other.
  - Another way to help create nearly identical groups is to assign each

person to a group by using techniques such as **randomized blocks** (aka **randomized strata**) and/or **randomized matches**. These can be excellent techniques if you have good reason to believe your groups aren't (or won't be) nearly equivalent from *completely random assignment*. They are beyond the scope of this lecture, but you can easily find information on these all over the internet.

- Regardless of how you decide to make your nearly identical groups, you will end up with two or more groups that you will compare against each other. In the simplest scenario, you could have one group receive the training (this is often called the **treatment-group**, the **intervention-group**, or the **experimental-group**), and another group (often called a **control-group**) won't get anything (they will just keep doing whatever they would normally be doing).
  - Note: If you have multiple versions of the T&D system, each one could correspond to a different *treatment-group*. So you could have several treatment-groups and one control-group.

What to do if you can't or aren't allowed to make separate groups

- Sometimes you might be asked to implement a **new T&D system without using a control-group** (because the CEO wants everyone to receive the anticipated benefits of the training). In that situation, you can try to suggest using a control-group temporarily, and then eventually letting everyone do the training after you've collected the data you need for calculating the T&D system's effects. However, the company might reject your suggestion and ask you to follow their request. Also, sometimes, you might be hired to evaluate a T&D system that was **already implemented on an entire company without a control-group**. Whenever you can't have groups for comparison, the next best thing you can do is try to get as much data as possible, especially from **before** the T&D system, and **after** the T&D system—and even multiple time-points before and after the T&D system.
  - In typical research methods jargon, such designs are known as one-group designs. Depending on when and how many times you collect data, you could end up with:
    - **one-group post-training data**;
    - **one-group pre-training & post-training data**;
      - the pre-training and/or post-training data can be collected at multiple time-points, in which case it may be called **interrupted time-series data** (*time-series* means *a series of things across time*, and it's *interrupted* because the training occurs at some point *during* the time-series data, therefore the training is said to *interrupt* the time series data).
- Technically, a truly randomized experiment (i.e., with two or more groups that are nearly identical) only requires data from *after* the intervention (e.g., after the training) but you can increase the power and precision

of your statistical analyses by including pre-intervention data (e.g., pre-training data), which could be included as covariates in your statistical model, or they can be used to create randomized blocks and/or randomized matches to help create nearly equivalent groups (see the section above on Creating nearly identical groups).

Suggested Readings

Items indicated with an asterisk (*) are available in the Zotero Group Library for this course.

- *Little, T. D. (Ed.). (2019). Randomized Experiments. In C. S. Reichardt, *Quasi-Experimentation: A Guide to Design and Analysis* (pp. 45–93). The Guilford Press.
  - Charles "Chip" Reichardt is an excellent research methodologist and his 2019 book is a treasure trove. Our Zotero Group Library only includes chapter 4 (Randomized Experiments) of the book, in compliance with U.S. copyright regulations on fair use for educational purposes.