EPSY 5221: Principles of Educational & Psychological Measurement

Classical Test Theory

The classic approach to introducing CTT and the definitions are based on the work of Allen & Yen (1979)[[1]](#footnote-1). Spearman (1910)[[2]](#footnote-2) was an early contributor to the true-score theory.

|  |  |
| --- | --- |
| *X* = *T* + *E* | The observed score *X* is the sum of the true score *T* and the error score *E*. The error score is due to error of measurement where *T* is fixed and *X* is the realization of a random variable. Error of measurement is an unsystematic, random deviation of a person’s observed score from the theoretically observed score. |
| E(*X*) = *T* | The expected value (long-run average) of *X* is *T*. *T* is the mean of the theoretical distribution of *X* scores that we would find in repeated independent testing of the same person with the same test. *T* is defined in terms of the expected test score, not any “real” trait of the person. |
| ρ*ET* = 0 | The error score and the true score obtained by a population of examinees on one test are uncorrelated. The magnitude of the errors of measurement is not related to the magnitude of true scores. |
| ρ*E*1*E*2 = 0 | The error scores for test 1 and test 2 are uncorrelated. Such correlations would indicate that the test scores are systematically affected by factors such as fatigue, practice, mood, or administration environment. |
| ρ*E*1*T*2 = 0 | The error scores on test 1 are uncorrelated with the true scores on test 2. Such correlations would indicate that test 2 measures a personality trait or ability that influences errors on test 1. |
| Parallel tests | Two tests that have observed scores *X* and *X*' that satisfy the above assumptions and, for every population of examinees, *T*=*T*' and . For the variances to be equal, conditions leading to errors of measurement (mood, environment) must vary in the same way for the two tests. The tests must also have equal observed-score means, variances, and correlations with other observed test scores. |

This presentation is based on the work of Haertel (2006)[[3]](#footnote-3), who referenced Feldt and Brennan (1989)[[4]](#footnote-4), the following approach was offered.

CTT begins with the idea that an observed test score is a random sample from a possible set of scores that could have been observed. The observed score is the sum of a true score and an error score. A person, *p*, has a true score that is constant over test forms *f*. Variation in observed scores over forms is the result of form-specific error.

*Xpf*= τ*p* + *Epf*

Form *f* is one of a set of strictly parallel forms. The properties of parallel forms provide for the assumptions necessary to derive the classic CTT statistics. These include identical test specifications which yield identical observed-score distributions (when administered to very large populations), which covary equally and covary equally with other measures (e.g., Z). These result in:

*F*(*Xf*) = *F*(*Xg*) = *F*(*Xh*) = … where σ*XfXg* = σ*XfXh* = σ*XgXh* = … and σ*XfZ* = σ*XgZ* = σ*XhZ* = …

As a test taker is repeatedly tested with different test forms (assuming no memory effect, learning, or fatigue), the expected value of the error score is zero. The expected value over forms *f* of errors for person *p* is zero:

E*f*(*Epf*) = 0

The same is true of a group of persons – as they are tested with any given form *f*, the expected value of the resulting errors would be zero as the group size approaches infinity:

E*p*(*Epf*) = 0

From these assumptions, we find the following results, which technically are not assumptions, but mathematical results of the preceding assumptions:

σ*TEf* = 0 σ*EfEg* = 0

Note that σ*EfXf* > 0, as both the true score and error score are components of the observed score.

We also find that covariance between parallel forms is true score variance:

This presentation is based on the work of Reykov and Marcoulides (2011)[[5]](#footnote-5).

*X* = *T* + *E*

for individual *i*, *Xi* = *Ti* + *Ei*

for measure *k*, *Xki* = *Tki* + *Eki*

This framework depends on the definition of the true score and in some cases, error scores. Both *T* and *E* are unobserved, so some assumptions must be made. Classically:

*Tki* = over a very large number of administrations.

The expected value of practically all possible iid measurements of a given person on a given measure where the construct being measured does not change.

Because *T* is the expected value of *X*, the expected value of *E* must be zero:

E(*X*) = *T*, thus to satisfy *X* = *T* + *E*, E(*E*) = 0.

*E* is random measurement error with a mean of zero. So random measurement error has no effect on true score.

Some suggest an assumption that ρ*TE* = 0. This is actually a property of the decomposition of *X*, true as a necessary deduction from the definition of *T*, *E*. This is equivalent to the property of residuals being independent of predictors. It also follows that the property of ρ*TE’* = 0 is also true.

Another so called assumption, that ρ*EE’* = 0, is not a necessary assumption of CTT, but is a testable hypothesis that may be necessary for some purposes, but not others. This is only relevant in cases where multiple tests are in use.

**Measurement Models Assuming that ρ*EE’* = 0**

Based on the work of Reykov and Marcoulides (2011)[[6]](#footnote-6).

*Parallel Measures*

All measures share the same True scores

Measures differ only in error scores, which have the same variances

*Xk* = *T* + *Ek*

=

*True-Score Equivalence, Tau-Equivalent Measures*

Tests measure same underlying constructs

Measures differ in levels of precision

*Congeneric Measures*

Measuring the same underlying latent construct

True scores are linearly related

*Xk* = d*k* + b*k* + *Ek*

Where d and b are test specific constants.

True scores of congeneric tests are perfectly correlated.

Observed scores are not perfectly corrected because of *Ek*

Congeneric models are appropriate in most behavioral, social, and educational research settings because measurement units lack precision and concrete meaning – they are specific to different tests.

The three measurement models assume a single underlying latent dimension. The three models are nested, and thus can be tested in a CFA framework:

Parallel < tau-equivalent < congeneric

1. Allen, M.J., & Yen, W.M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole. [↑](#footnote-ref-1)
2. Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101. [↑](#footnote-ref-2)
3. Haertel*,* E.H. (2006). Reliability. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education/Praeger. [↑](#footnote-ref-3)
4. Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York, NY: Macmillan. [↑](#footnote-ref-4)
5. Reykov, T., & Marcoulides, G.A. (2011). *Introduction to psychometric theory*. New York, NY: Routledge. [↑](#footnote-ref-5)
6. Reykov, T., & Marcoulides, G.A. (2011). *Introduction to psychometric theory*. New York, NY: Routledge. [↑](#footnote-ref-6)