CHAPTER NINE
Investigating validity

II. Conceptual exercises

A. Matching

1. b
2. a
3. f
4. h
5. i
6. d
7. e
8. c

B. True or false

1. T
2. T
3. F
4. F
5. F
6. F
7. T
8. T
9. F
10. T
11. T
12. F

C. Short answers

1. Two required activities:
   a. Articulating an interpretive argument (also referred to as a validation argument), which provides the logical framework linking test performance to an intended interpretation and use; and
   b. collecting relevant evidence in support of the intended interpretations and uses.
2. Linn and Gronlund's cautions:
   a. Validity is a quality of interpretations and uses of assessment results; it is not a quality of either tests or test scores.
   b. The validity of a particular test use is a matter of degree.
   c. Validity is always specific to a particular use or interpretation.

d. Validity is a unitary concept.

e. Validity involves an overall evaluative judgment.

3. Three interrelated parts of test use:

a. Score interpretation;

b. score use; and

c. consequences of score use.

4. Two functions:
   a. Providing a guide for test development; and
   b. providing a framework for collecting evidence in support of the intended interpretations and uses.

5. Four considerations:
   a. The intended use of the test and the potential impact of that use;
   b. the way in which we have defined the construct to be measured;
   c. the kinds of tasks included in the test; and
   d. the characteristics of the intended test takers.

6. Two kinds of inferences:
   a. Predictions about future performance beyond the test itself; and
   b. inferences about knowledge or ability that test takers have.

7. Five recommended steps in designing a validation study:

   a. Develop operational procedures (e.g. measures, observations) for collecting data relevant to each claim and counterclaim. These procedures will include both tasks for eliciting performance and procedures for scoring that performance.

   b. State hypotheses about the expected patterns of relationships or differences among the groups or variables observed. For supporting claims, we hypothesize which variables should be related to each other, or which groups should perform differently on tests of the ability to be measured. For rejecting counterclaims, we hypothesize which variables should not be related to each other, or which groups should perform the same on tests of the ability to be measured.

   c. Design a study to collect the relevant data.

   d. Collect data.

   e. Test hypotheses; for correlational designs, test hypotheses using the appropriate correlational procedures; for experimental and comparison group designs, test hypotheses using appropriate statistical tests.

8. Five quantitative approaches to validation:

   a. The analysis of test content;

   b. the analysis of test-taking processes;

   c. the analysis of correlations among scores from a large number of tests;

   d. the analysis of differences among comparison groups in an experimental design; and

   e. the analysis of differences among non-equivalent criterion groups.

9. Usefulness of confirmatory factor analysis (CFA) for analyzing MTMM:

   In a CFA analysis, the pattern of factor loadings is specified and then tested against a set of data. The expected pattern may derive from a theoretical model, or prior research or both. An MTMM correlation matrix provides a basis for specifying a CFA model with several trait factors and different method factors. On the other hand, by looking at the model fit, which indicates the extent to which the correlations estimated on the basis of the CFA model match the observed correlations, one can decide whether the hypothesized pattern of factor loadings is supported by the data or not.

10. Four components of a true experimental design:

    a. Random selection: a sample of participants is selected at random from a population, or large group that comprises the population to which we want our results to generalize.
    b. Random assignment to groups: the sample of participants is assigned to several groups at random.
    c. Different treatments: Each group of participants undergoes a different treatment.
    d. Post-test: Each group is given a test at the end of the treatment.

11. Steps in non-equivalent group design:

    a. Identify and select criterion groups: Several groups of individuals who are clearly at different levels of the ability to be measured are identified on the basis of some independent indicator of the ability we want to measure.
    b. Administer the same test to all groups.

III. Short answers to scenarios

1. Evidence of content representativeness. Evidence of this type provides some support of the claim that the test measures what it is intended to measure.

2. Correlational: Exploratory Factor Analysis. Evidence of these different papers having high factor loadings on a general factor would provide support for the claim that the different papers measure essentially the same ability. However, if the patterns of factor loadings are varied (some papers/sections have high loadings on a factor while others have low loadings on it), this would constitute some evidence in support of the counterclaim, that the sections/papers are not measuring essentially the same ability.

3. Confirmatory factor analysis (CFA) of MTMM correlations. The design of the CFA analysis of MTMM study provides a basis for specifying a CFA model with two or more trait factors (Listening and Reading, in this case) and several different method factors (multiple-choice and open-ended questions, in this case). This model would specify that the two measures have factor loadings on their associated trait and method factors and

zero loadings on all other factors. In applying CFA, the researcher specifies a pattern of expected or hypothesized factor loadings, including the possibility that the factors themselves may be correlated, and then tests this pattern against the correlations that are actually observed. This is done by estimating the factor loadings specified by the model on the basis of the observed correlations, and then using these factor loadings to estimate what the correlations among the observed variables would be if the specified model provides a good 'fit' to the data. The differences between the observed correlations and the correlations estimated on the basis of the CFA model are an indication of how well the model fits the data. The smaller the differences, the better the model fit.

IV. Research designs

A. Comparing approaches to ESL reading instruction

1. An equivalent-groups design would be appropriate. This could entail multiple groups of students from the program, with the students randomly assigned to different treatment groups for the study. Students would be given different kinds of instruction and then take reading tests at the end of the course.

2. In order to collect evidence in support of the claims that these two reading methods (phonics-based and whole language-based) will have different effects on reading ability, this experimental design study has four comparable groups, three in different kinds of reading instruction, and one group receiving no instruction. Group 1 would be instructed using the phonics method and Group 2 using the whole language method. There are two control groups: Control Group 1 receives general reading instruction, while Control Group 2 receives general language instruction instructed.

Groups	Treatment	Post-test
Experimental Group 1	Phonics-based	X_p, $X_{wl,}$ X_{gr}, X_{glp}
Experimental Group 2	Whole language-based	X_p, $X_{wl,}$ X_{gr}, X_{glp}
Control Group 1	General reading	X_p, $X_{wl,}$ X_{gr}, X_{glp}
Control Group 2	General language	X_p, $X_{wl,}$ X_{gr}, X_{glp}

3. At the end of the instructional period, all groups take four tests:

   (1) One that covers the aspects of Reading that are the sole focus in the phonics-based instruction (X_p);

   (2) one that covers the aspects of reading that are the sole focus of the whole language based instruction (X_{wl});

   (3) one that is a general reading test (X_{gr}); and

   (4) one that is a general language proficiency test (X_{glp}).

4. The claims and counter-claims are operationalized by the hypothesized differences in group performance. Thus, we would expect Experimental Group 1 (E1), which receives the phonics-based method, to perform best of all the groups on the phonics-based test, with Experimental Group 2 (E2) and Control Group 2 (C2) to perform better than Control Group 2 (C2). We would also expect group E2 to perform best on the whole language-based part of the test, with E1 and C1 performing better than C2. In addition, we would expect groups E1, E2 and C1 to perform differently on the general reading test, with all three performing better than C2. Finally, we would expect group C2 to perform best on the general proficiency test, with E1, E2 and C1 performing about the same.

   (1) Phonics-based test: $\overline{X}_{E1} > \overline{X}_{E2}, \overline{X}_{C1} > \overline{X}_{C2}$

   (2) Whole language-based test: $\overline{X}_{E2} > \overline{X}_{E1}, \overline{X}_{C1} > \overline{X}_{C2}$

   (3) General reading test: $\overline{X}_{E2} \neq \overline{X}_{E1} \neq \overline{X}_{C1} > \overline{X}_{C2}$

   (4) General language proficiency test: $\overline{X}_{C2} > \overline{X}_{E2}, \overline{X}_{E1}, \overline{X}_{C1}$

If these patterns of differences hypothesized in 1 and 2 are observed, this would be evidence in support of the claims that the different approaches to Reading enable students to perform best on what they were taught. If any of the three Reading groups (E1, E2, C1) out-performed the others on the general Reading test, then this would support the claim that that method is more effective than the others. If all three Reading groups performed better than group C1 on the three Reading tests, this would also constitute

evidence for rejecting the counter-claim that performance on these tests is influenced by general language proficiency.

5. To test these hypotheses and these claims, four separate one-way analyses of variance could be used to determine if the observed differences among the group means on the four tests are statistically significant. (It would also be appropriate to use a multivariate analysis of variance to analyze all four tests together.)

B. Differences among writing classes

1. A non-equivalent criterion-groups design would be appropriate. This study would involve students in all three intact writing level classes. Students are not randomly assigned to the different programs for the study.

2. The claims and counter-claims are operationalized by the hypothesized differences in group performance. Based on prior experience, it is possible to hypothesize that the Advanced Level group would perform better in Writing (in terms of mean scores) when compared to the Intermediate and that the Intermediate Level group would perform better than the Beginning Level group:

$$\bar{X}g1 > \bar{X}g2 > \bar{X}g3$$

3. In order to collect evidence in support of any of the claims stated above, a three-group design is necessary. Group 1 is the Advanced Level group, Group 2 is the Intermediate Level group and Group 3 is the Beginning Level group. All three groups would receive appropriate writing instruction for the semester, and at the end of the instructional period, all groups would take a test of narrative writing (X_w).

Groups	Treatment	Post-test
Group 1: Advanced Level	Regular instruction; no special treatment	X_w
Group 2: Intermediate Level	Regular instruction; no special treatment	X_w
Group 3: Beginning Level	Regular instruction; no special treatment	X_w

4. The test is a narrative writing test best designed as a general writing test in a paper and pencil mode with no bias in favor of any of the groups in terms of topics.

5. To test these hypotheses and these claims, a one-way analysis of variance should be used to determine if the differences among these means are statistically significant.