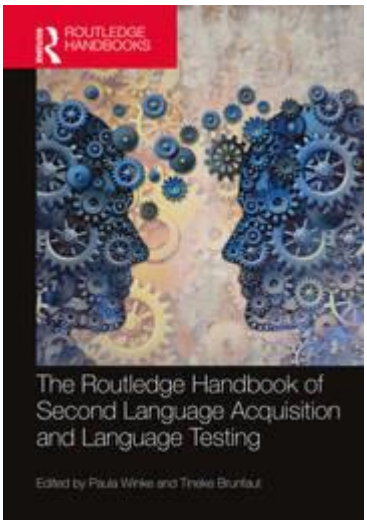
[Preprint version] Please cite as:
Suzuki, Y., & Koizumi, R. (2021). Using equivalent test forms in SLA pretest−posttest research design. In P. Winke & T. Brunfaut (Eds.), The Routledge Handbook of Second Language Acquisition and Language Testing (pp. 456-466). New York: Routledge.



ABSTRACT

Pretest-posttest design requires multiple testing on the target constructs of knowledge and skills over time. It thus necessitates equivalent test forms to reliably assess L2 development while reducing the influence from extraneous factors such as practice effects. In this chapter, the nuts and bolts of equivalent test use are presented for test item construction, pilot data collection, and test administration. In addition, a focused survey was conducted on the practice of using multiple equivalent test forms in a subdomain of instructed L2 research. Recommendations are presented for appropriate use of equivalent test forms in pretest-posttest design.

Background

Pretest-posttest experimental design is essential when the aim is to identify the effects of different treatments on second language (L2) acquisition. This experimental research design typically requires two or more test forms (e.g., pretest, immediate posttest, and delayed posttest). These test forms should be interchangeable in terms of content, difficulty, and other characteristics. Forms that have such equivalency are called equivalent forms, alternate forms, or parallel forms (Henning, 1987).[1] They facilitate consistent score interpretations and thus help in objectively assessing L2 development over time.

Equivalent test forms also minimize a threat to the internal validity of pretest-posttest design used in experimental research (Brown, 1988). When participants are administered the same test before and after the treatment, the experience gained during the first test may influence the performance on the second test. This effect is called a *practice effect*. It refers to "the fact that taking two tests with the same or similar content may result in a higher score on the second test, despite there being no increase in ability in the skill being measured" (Davies et al., 1999, p. 148).[2] The discussions provided in this chapter focus on practice effects observed when the same test is used in pretest-posttest design. There are at least three major sources of practice effects: (a) test-takers remember (some elements of) test item content; (b) test-takers notice some patterns or intentions of researchers in the tests and learn a strategy to perform on the test better; and (c) they get used to the test situation and perform better for the second time. It may not be possible to completely eliminate the second and third possibility by using equivalent test forms. These risks, however, can be reduced by including many filler test items and inserting a longer interval between multiple test administrations. These unintended and undesirable practice effects jeopardize the interpretations of score changes from one test to the next. In other words, practice effects make it impossible to fairly assess learning gains that can be attributed to the experimental treatment. This concern is equally important for observational studies where the development of L2 learners is being tracked. In

[1] Although these three form types may sometimes be distinguished by definitions (e.g., AERA, APA, & NCME, 2014), they are considered the same in this chapter.
[2] Repeated testing can also lead to lower scores due to fatigue and decreased motivation.

order to minimize practice effects, researchers should be equipped with several methodological options, such as using equivalent test forms. The appropriate use of equivalent test forms and related techniques contributes to scientifically rigorous research in both fields of second language acquisition (SLA) and language testing (LT).

In the sections that follow, researchers will be given advice on how to design equivalent test forms for a pretest-posttest design study and to account for the equivalence of test forms and their use from experimental and statistical points of view (e.g., counterbalancing and test equating). This will be followed by a brief overview of the current practice in using equivalent test forms in a subdomain of instructed SLA research to demonstrate that they are insufficiently exploited. Finally, some recommendations for using equivalent test forms in SLA and LT research are given.

Key Concepts

Practice effect: The influence of one test on the next one when the same test is given multiple times to the same participants. Practice effect is also called recall effect, memory effect, test order(ing) effect, carryover effect, and test familiarity effect.

Equivalent forms: Test forms that have very similar content, difficulty, and other pertinent characteristics. Test development based on test specifications and statistical analysis is required to ensure equivalency of test forms. Equivalent forms are also called alternate or parallel forms.

Test equating: A statistical method of adjusting scores on multiple test forms to ensure that they can be used interchangeably.

Key Issues

Here, the nuts and bolts of equivalent test design and its use are discussed in relation to three stages of research. In the test design phase, researchers need to create test forms to assess the same construct for the same purpose, with the same content and statistical specifications, to the best of their abilities. In the pilot testing phase, they should examine the actual psychometric characteristics of the test forms to demonstrate evidence of equivalency. In the main study, they should implement a design that ensures test equivalency by using counterbalancing techniques.

Test Design

Development of language tests requires test specifications to account for how tests can measure target constructs (e.g., target linguistic ability). Based on these specifications, multiple test forms that contain different items but fulfill very similar item characteristics are created. The test specifications typically include detailed descriptions of content areas, format, difficulty, the number of test items, ratios of different item types, item order, and how to administer the test and score the responses, as well as specifications on any other pertinent elements, such as test directions (Davies et al., 1999).

By covering some of these key features of test design, we explain how to create equivalent test forms in this section. For illustration purposes, we focus on grammar tests because the acquisition of grammatical knowledge remains one of the most researched linguistic domains in SLA research. Outcome measures for the acquisition of linguistic rules can be broadly categorized into: (a) controlled tests, where target linguistic rules are obligatorily tested for receptive or productive use; and (b) free production tests, where test-takers are given control over the linguistic features they use for production (Norris & Ortega, 2000). Examples of controlled tests are a grammaticality judgment test (GJT), an oral elicited imitation (EI) test, a text reconstruction test, a translation test, a picture matching test, a multiple-choice test, etc. Free production tests typically include an oral/written narrative test and an interview test.

In controlled tests, researchers can take a stricter control of learners' responses than in production tests. For instance, one of the widely-used tests, a GJT, typically requires participants to read or listen to a single sentence and to judge whether the sentence is grammatically correct (Spinner & Gass, 2019). Equivalent forms of GJT should therefore have the same number of test items (e.g., 40),[3] as well as equal proportion of item types (e.g., 20 grammatical and 20 ungrammatical items, with a balanced number of items related to targeted constructs, such as 10 past-tense, 10 past-perfect, and 20 filler items), and the order of test items should be (pseudo-)randomized across forms. In addition, other factors should be also controlled across test forms, such as sentence length (e.g., number of words, syllables,

[3] Researchers sometimes use a smaller number of items for a pretest, e.g., when participants are likely to know little about target linguistic rules and a large number of items is likely to lead to undesirable outcomes, such as demotivation and fatigue. In this case, proportions of correct items are compared.

and audio stimulus lengths, when necessary) and lexical difficulty of stimulus sentences (e.g., frequency, familiarity).

While these rules seem straightforward, creating equivalent forms is challenging. For instance, when a GJT is designed to assess learners' knowledge of subjunctive past perfect in English, similar test items for equivalent test forms can be constructed in several ways. Suppose that the item (A) below is the baseline that contains verb form error (i.e., the correct form is "had taken"). A rule of thumb is to create a sentence that is not too different nor too similar to the baseline. Higher similarity between test items increases interchangeability of test scores across forms, but it tends to increase practice effects. Lower similarity can reduce practice effects; however, if test items differ significantly in many aspects, their equivalency becomes questionable. This point is illustrated below.

A) If my father took the car, he would have come here earlier. (Baseline)
B) If my father had taken the car, he came here earlier. (Inappropriate)
C) If my mother took the car, she would have come here earlier.
D) If my mother took the car, she would have left home later.
E) If my mother took the picture, she would have kept it here.
F) If my mother helped my father, he would have been very happy.

Apparently, item (B) has some resemblance to the baseline item (A), but it may not constitute a good equivalent item because the target is different (the error is in the main clause rather than the subordinate clause). Thus, the remaining items—(C) to (F)—are better candidates as an equivalent item for item (A). While the number of words and lexical difficulty are controlled, the degree of similarity decreases from (C) to (F). The items (C) and (D) include the same verb phrase ("took the car") in the subordinate clause. However, while the same verb phrase is used in the main clause in item (C), it is changed in (D). In comparison, the item (E) is more different from the baseline. While the same verb "took" is retained, the object is changed ("took the picture") and the main clause is also modified. In the item (F), the target verb phrase is completely changed ("helped my father"). Unless researchers are also interested in the acquisition of verb form (irregular inflection as in "took" vs. regular inflection as in "helped"), item (F) is most likely to alleviate practice effects, while measuring the knowledge of subjunctive past perfect.

Many decisions have to be made when constructing similar GJT items for equivalent forms. Thus, it is essential to explicitly report in research articles how items in equivalent forms were constructed and justify the chosen strategy. For instance, researchers can select item (C) or (D), if their aim is to assess the direct effect of treatment on the verb phrases (e.g., "took the car") practiced in the treatment, and/or if participants' English proficiency is limited. While GJT is a receptive test, a production test such as EI test (i.e., wherein participants listen to a sentence and repeat it under certain time pressure) may require more careful construction. For one reason, learners may find it difficult to repeat the sentence under time pressure for the first time, but would likely perform better on the second test simply due to the practice effect (rather than as a result of gaining grammatical knowledge of a specific construction). If some tests are more susceptible to practice effects, researchers can create a different test form using more dissimilar sentences. Of course, researchers cannot be certain how practice effects pertaining to different tests for a given study develop. That is why it is critical to evaluate test form equivalency by conducting a pilot study, which is discussed in the next section.

When it comes to free production tests of grammatical knowledge, the elicitation of target grammatical structures can be the key aspect of a certain pretest-posttest research design. Test developers should make sure that participants are equally likely to use targeted grammatical structure in different test forms, as this will allow the development of grammatical knowledge to be assessed in the natural context of language production. Prompts (e.g., picture, short passage, and questions) are typically used to elicit spoken and written production. They should be given, prior to the main experiment, to a group of participants (e.g., native speakers, or a sample of L2 learners similar to the population targeted by the main study). Preliminary analysis should determine the usability of each prompt by evaluating some criteria, which can be guided by the previous research (e.g., eight out of ten participants used the target structure for a prompt). Some problematic items will be discarded along the way.

Finally, it is worth noting that there is an emerging interest in developing controlled, real-time sentence processing tests utilizing different measures (e.g., reaction time and eye-movement) as measures of implicit grammatical knowledge (Suzuki, 2017). For instance, Godfroid (2016) conducted one of the first studies in which a reaction-time test called the *word-monitoring test* was used as a part of a pretest-posttest design in instructed L2 research. Readers can easily imagine that developing multiple equivalent forms for this type of test

requires much more effort and careful construction than traditional tests like GJTs. Moreover, as reaction-time and eye-movement measures are highly sensitive and more susceptible to extraneous variables (e.g., lexical access speed, predictability of cues, visual prompts) than traditional measurements, pilot studies should also be designed and conducted more rigorously to verify test equivalence prior to the main experiment.

Pilot Testing

Once similar multiple test forms are created, researchers need to check if these forms are actually equivalent by administering them to a trial sample drawn from the target population. Three main methods of inspecting test equivalency are presently in use.

The method of *equivalency* is the strictest of the three and requires the test forms to have (a) the same means, (b) the same standard deviations (*SD*s), and (c) strong correlations with each other (i.e., equivalent-forms reliability) or the same correlations with other tests (Bae & Lee, 2011; Henning, 1987). These assumptions should be always verified by administering multiple test forms to the same group. Whether means, *SD*s, and correlations are the same should be established by significance tests, such as *t*-test, with preferably very small effect sizes indicating small differences between test forms. We argue that the three conditions are equally important, although the third type of evidence for correlations is sometimes considered to be optional (see AERA, APA, & NCME, 2014, p. 35, for an example). Furthermore, the third condition (c) can be investigated more rigorously by examining an internal structure of each test form through structural equation modelling, e.g., by checking for the same factor loadings, factor variances, and factor covariances (Bae & Lee, 2011). See Weir and Wu (2006) for an example of this method.

Second, in the method based on *random parallel test forms*, multiple forms are developed using an "item bank" with a large pool of test items that conform to the same test specifications (see Oswald, Friede, Schmitt, Kim, & Ramsay, 2005, for an illustration). These test items are piloted and often analyzed through item response theory (IRT; Henning, 1987), and their item characteristics (e.g., item difficulty, item discrimination) are known. When creating test forms, researchers would randomly select the same number of items for each test form from within each band of item difficulty, with the items balanced on item discrimination as well. This procedure works best when there are more items in the item bank. Since this procedure may produce slight variations across test forms, especially if there are few items

overall in the item bank, it is advisable to check their characteristics using the approach explained in the first method.

The third method is based on *test equating*, which enables researchers to equate test forms by transforming test form scores "onto a common scale which allows for comparison" across test forms (Davies et al., 1999, p. 199). In this case, test forms may have different means and *SD*s, but their scores can be used comparably by employing a conversion table (see Stewart & Gibson, 2010, for an example). According to Kolen and Brennan (2014), equating "adjusts for differences in difficulty among forms that are built to be similar in difficulty and content" (p. 2). Test form equating is performed in three steps (Kolen & Brennan, 2014).

- First, scores from one form are related to a common score scale, which is constructed based on the results (e.g., means) from a norm group (e.g., using linear or nonlinear transformation). Then, as part of this first step, a conversion table is created, delineating the relationship between each raw score of the form and its corresponding scale score.
- Second, raw scores on a new form (e.g., Form Y) are usually equated to those on the old form (Form X). This second step (in which new form scores are equated to the old form scores) involves multiple phases, such as selecting and implementing an appropriate equating design, choosing statistical techniques for equating estimation, and examining the equating quality. Interested readers can consult Kolen and Brennan's (2014) comprehensive guide on the statistical equating methods based on raw scores as well as IRT.
- Third and last, a revised conversion table is generated by adding the information derived from the second step. Table 43.1 illustrates the scale score derived from the raw scores pertaining to three test forms. In this case, Form X is one point easier than Form Y and is one point more difficult than Form Z, to which the scale scores are adjusted. Thus, by using the scale scores, the three test form scores can be represented on a common scale and used interchangeably.

Table 43.1.

Conversion Table of Three Test Forms in a Hypothetical Situation

Scale score	Form X raw score	Form Y raw score	Form Z raw score
⋮	⋮	⋮	⋮
25.0	20	19	21
25.5	21	20	22
26.0	22	21	23
26.5	23	22	24
27.0	24	23	25
⋮	⋮	⋮	⋮

When administering two test forms for equating, there are three options for participant group selection (Kolen & Brennan, 2014). In one approach, two whole test forms are administered to the same participants, with the order of test forms counterbalanced. In the second approach, only a representative sample of items from two forms (called anchor or common items) is administered to two different groups, and the remaining items of two test forms are presented to one of the two groups. In the third approach, participants are randomly assigned to one of the forms—by distributing either test form sequentially, e.g., in the order of seating. These groups of participants are considered equivalent, allowing test scores to be compared across the test forms.

While the first option requires a smaller number of participants, having to take two tests can be overwhelming. In contrast, the second approach requires only one testing event; nonetheless, this strategy requires careful advanced planning including careful selection of anchor items that represent overall test content and its statistical characteristics (e.g., item difficulty, which is estimated before the test equating) and inserting common items in the same locations across forms. The third option can be applied to more than two forms without increasing the test-taking time; however, it necessitates a larger number of participants to obtain stable results. In all the three approaches, equating can be conducted not only in the pilot testing phase, but also after the main test administration and before the main analysis

(Wendler & Walker, 2016). Although checking the equivalency of test forms may appear tedious, it is indispensable for ensuring test equivalency statistically.

Counterbalancing Test Forms

Counterbalancing (i.e., systematically allocating equivalent forms to different groups) is another key aspect of pretest-posttest research design, as it helps control for unknown confounding factors over multiple testing occasions (e.g., practice effects). It involves a systematic assignment of multiple test forms to different testing events (e.g., pretest and posttests). Suppose that L2 experimental research involves pretest, posttest I and posttest II with two treatment conditions. Without counterbalancing, for instance, all participants are given Form A in the pretest, with Form B and Form C adopted in posttest I and posttest II, respectively. This fixed matching between forms and test occasions involves a risk that one test form is easier or more difficult than the others due to practice effect.

As shown in Table 43.2, when counterbalancing three test forms, researchers can randomly assign participants into six groups within each condition. A technique called a diagram-balanced Latin square design can reduce the required number of groups for counterbalancing (see Foley, 2004; Keppel & Wickens, 2004, for details). With this technique, when the number of test forms (k) is even, then only k groups are necessary. When the number of test forms is odd, then 2 x k groups are necessary. For example, in using four test forms, researchers should use four groups (not 24), while 10 groups are sufficient for five test forms (not 120).

Table 43.2.

An Example of Counterbalancing of Three Test Forms for Pretest, and Posttests I and II

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Pretest	Form A	Form A	Form B	Form B	Form C	Form C
Posttest I	Form B	Form C	Form A	Form C	Form A	Form B
Posttest II	Form C	Form B	Form C	Form A	Form B	Form A

Note. Complete counterbalance requires 12 groups (two conditions × six groups).

Assuming that multiple test forms are randomly distributed to groups, counterbalancing can help neutralize the practice effect of administering multiple test forms.

In other words, counterbalancing increasing the confidence of interpreting the average score changes from pretest to posttests as an indicator of L2 development for different groups.

Use of Equivalent Test Forms in L2 Form-focused Instruction Research

In this section, we report a small literature survey that examined how equivalent test forms have (not) been used in L2 research. Since a comprehensive review of pertinent literature was unfeasible, we focused on the subdomain of instructed L2 research. Thus, we selected the study conducted by Goo, Granena, Yilmaz, and Novella (2015) as it is the most recent comprehensive meta-analysis on L2 form-focused instruction. In this work, the authors reviewed 34 empirical studies published in 1980 through 2011 that met the following criteria: (1) research involved a (quasi-)experimental pretest-posttests research design; (2) the effects of different treatments (e.g., explicit and implicit grammar instruction) were compared; and (3) research targeted the development of specific L2 features (e.g., grammatical, pragmatic, lexical areas).

For this assessment, we selected 15 experiments (see the references indicated with asterisks) from these 34 empirical studies, which involved the assessment tasks employed on at least three occasions (e.g., pretest, posttest I and posttest II). The authors coded all 15 experiments for (a) the existence of multiple test forms (no different test form vs. more than one test forms vs. all different test forms). If multiple test forms are used, we further coded for (b) content equivalence, (c) item difficulty equivalency, (d) pilot testing (e.g., whether equivalent-forms reliability was computed), and (e) counterbalancing (counterbalancing vs. no counterbalancing).

The analyses of these 15 studies revealed that (a) in more than half of these cases, the researchers employed identical forms for pretest and multiple posttests. Specifically, no use of different forms is identified in nine studies,[4] at least two different forms (e.g., pretest and immediate posttest were different, but the pretest and the delayed posttest were identical) in three studies, and all test forms were different in three studies. In terms of (b), among the six studies in which different forms were used, how different forms were controlled for content similarity was explicitly stated in only four cases.

[4] In three studies, researchers simply used the same items but changed the item order for different forms; we coded these studies as "no different test form."

However, it was more difficult to identify evidence of (c) difficulty equivalency.[5] In only two of the six studies, a clear description of the process employed was provided for controlling item difficulty. These studies coincided with the ones that reported conducting (d) pilot testing (Benati, 2004; Tode, 2007). In one study (Benati, 2004), "the three versions were balanced in terms of difficulty and vocabulary during a pilot experiment" (p. 216). In the other case (Tode, 2007), the author provided the explanation of a pilot study, stating that its aim was checking equivalent-forms reliability and indicating that the random parallel test method was employed. In the pilot phase, Tode (2007) created six equivalent test forms (which served as a pretest and five posttests) and administered six tests to a group of students in the same school (i.e., another sample from the target population). These six forms were created by selecting test items from a large pool of items with estimated difficulty. Tode (2007) reported that "all Pearson correlations among the six forms . . . [were] over .90" (p. 19). The creation of six forms in the pilot study is commendable, as it allowed the author to interpret the test score differences straightforwardly. Overall, however, none of the authors of the 15 studies included in the current methodological review examined means, *SD*s, and correlations or conducted test equating, which are two of the three available methods for ensuring test equivalency (see the Pilot Testing section).

In terms of (e), counterbalancing was not employed in any of the six studies in which researchers used different test forms. This is another unfortunate finding of this methodological review.

Overall, the current analysis clearly suggests that methodological and reporting practices in the L2 research domain are deficient (cf., Plonsky, Marsden, Crowther, Gass, & Spinner, in press). Given the importance and impact of meta-analyses in form-focused L2 instruction research (e.g., Goo et al., 2015; Norris & Ortega, 2000), it is surprising to find that, in a significant number of empirical studies in this field, the authors failed to assure the assumption of test equivalence. As the current analysis is far from comprehensive, the practices in other domains are unknown. However, we emphasize that creating, using, and providing detailed descriptions of equivalent test forms is an important aspect of SLA

[5] For example, Muranoi (2000) stated that "Item sequences were shuffled and several content words were replaced with others equally familiar to the students" (p. 642). We found it difficult to conclude that this author controlled "difficulty" (slightly different from familiarity) of items in the study, and thus coded it as "no explanation of difficulty equivalence."

pretest-posttest research design. Moreover, given that this approach is not consistently adopted, there is immediate need for establishing guidelines for constructing and using equivalent forms. To adhere to procedures to adequately use equivalent test forms, expertise in the LT field (test development, interpretation and use, see Carr, 2011, for instance) benefits researchers when they critically analyze and interpret previous studies as well as their studies from methodological and measurement perspectives. Interdisciplinary collaboration across SLA and LT fields is thus an effective means to improve the quality of research.

Recommendations for Practice

Test Design

Researchers should first develop detailed test specifications for multiple test forms and should always report them in publications to demonstrate test equivalence. Asking for external content reviews in advance would help identify problems with test items and test forms, such as inadequate items and extreme content overlaps (AERA, APA, & NCME, 2014). It is also noteworthy that many linguistic journals allow the contributing authors to upload all test items used in their studies as a supplementary online material. Similarly, such materials can be uploaded to a digital depository, such as IRIS (Marsden, Mackey, & Plonsky, 2016). Such materials should also be shared with reviewers as an additional equivalence check.

Pilot Testing

Whenever using new test forms, researchers should conduct a pilot study, as this would allow them to examine the features of each form prior to the main test administration. For the pilot study, it is recommended to include 20−30% more items into the forms than would be employed in the main research, as some problematic items may have to be excluded for numerous reasons (e.g., low item discrimination, attractiveness of options in a multiple-choice format, items of free production test failing to elicit target structures). During this pilot testing, the order of the item presentation should be carefully determined and the order of form presentation should be counterbalanced when multiple forms are administered to the same group, to minimize ordering effects that may skew the analyses of test form equivalency. Note that as long as situations allow, counterbalancing should be conducted prior to the test administration for test form equivalency, because post-hoc counterbalancing may not completely eradicate test difficulty-level differences across test forms.

Test Administration

In the main study, counterbalancing of test forms is recommended. This can be easily achieved, especially when tests are administered by a computer. Although it is practically more difficult to conduct paper-and-pencil tests in classroom settings, different forms can still be given to different participants. In addition to random assignments of multiple test forms, when a covariate is available (e.g., proficiency test score), researchers are advised to assign participants into groups to ensure that the scores on the covariate will be equivalent. This is an additional measure that helps reduce unexpected sources of threat to internal validity. Furthermore, researchers may conduct more rigorous statistical analysis to take into account a factor of counterbalancing (see Keppel & Wickens, 2004). The main advantage of this strategy stems from stricter statistical control over unexpected confounding factors and greater statistical power; however, computation can be complex and a larger sample size is often required.

When multiple equivalent tests are not available (this is unfortunate but often the case in many SLA studies as shown above), counterbalancing of test forms is mandatory at minimum to compare the scores *at the group level*. It is important to note that interpreting *individuals*' *score changes* may not be meaningful without the tests that are equated properly prior to the main experiment. If test forms are equated, no counterbalancing may not be needed because practice effects can be measured by employing a control group. In any case, explicit reasonable justifications for counterbalancing and non-counterbalancing are necessary for supporting adequate score interpretations and uses of multiple test forms in pretest-posttest research design. In addition to counterbalancing, considering measurement errors and statistical artifacts (regression to the mean and standard error of difference) is also essential, but is often overlooked in interpreting results in pretest–posttest designs (see Koizumi, In'nami, Azuma, Asano, Agawa, & Eberl, 2015; Plonsky, this volume).

Testing Tips

・Create and report test specifications for multiple test forms in publications to account for test equivalence.

・Conduct a pilot study to check the features of each form prior to the main test administration.

・Report some indices (e.g., equivalent-forms reliability) of multiple test forms to support test equivalence.

・Counterbalancing of test forms is recommended.

Recommended Readings

Mackey, A., & Gass, S. M. (2015). *Second language research: Methodology and design* (2nd ed.). New York, NY: Routledge.

This updated book provides a good introduction to the key issues in research design and statistical analysis in a variety of contexts of second language learning.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed. International ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

This book is an intermediate textbook that provides detailed explanations on research design and data analysis. It is a useful resource particularly for the use of analysis of variance (ANOVA).

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer.

This book is a complete summary of what researchers should know to apply theory of test equating into practice.

References

(Asterisks indicate 15 studies included in the methodological synthesis.)

*Andringa, S., de Glopper, K., & Hacquebord, H. (2011). Effect of explicit and implicit instruction on free written response task performance. *Language Learning*, *61*, 868–903. https://doi.org/10.1111/j.1467-9922.2010.00623.x

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Bae, J., & Lee, Y.-S. (2011). The validation of parallel test forms: 'Mountain' and 'beach' picture series for assessment of language skills. *Language Testing, 28,* 155–177.

https://doi.org/10.1177/0265532210382446

*Benati, A. (2004). The effects of structured input activities and explicit information on the acquisition of the Italian future tense. In B. VanPatten (Ed.), *Processing instruction: Teory, research and commentary* (pp. 187–206). Mahwah, NJ Lawrence Erlbaum.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design*. New York, NY: Cambridge University Press.

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford: Oxford University Press.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge, U.K.: Cambridge University Press.

*Ellis, R. (2007). The differential effects of corrective feedback on two grammatical structures. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 339–360). Oxford: Oxford University Press.

*Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition, 28*, 339–368. https://doi.org/10.1017/S0272263106060141

Godfroid, A. (2016). The effects of implicit instruction on implicit and explicit knowledge development. *Studies in Second Language Acquisition, 38*, 177-215. https://doi.org/10.1017/S0272263115000388

Goo, J., Granena, G., Yilmaz, Y., & Novella, M. (2015). Implicit and explicit instruction in L2 learning. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 443-482). Amsterdam, Philadelphia: John Benjamins Publishing Company.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research.* Boston, MA: Heinle & Heinle.

*Kang, H. S. (2010). Negative evidence and its explicitness and positioning in the learning of Korean as a heritage language. *The Modern Language Journal, 94*, 582–599. https://doi.org/10.1111/j.1540-4781.2010.01093.x

*Koike, D. A., & Pearson, L. (2005). The effect of instruction and feedback in the development of pragmatic competence. *System, 33*, 481–501. https://doi.org/10.1016/j.system.2005.06.008

Koizumi, R., In'nami, Y., Azuma, J. Asano, K., Agawa, T., & Eberl, D. (2015). Assessing L2 proficiency growth: Considering regression to the mean and the standard error of difference. *Shiken*, *19*(1), 3–15. http://teval.jalt.org/node/16

*Kubota, M. (1994). The role of negative feedback on the acquisition of the English dative alternation by Japanese college students of EFL. *Institute for Research in Language Teaching Bulletin, 8*, 1-36. https://files.eric.ed.gov/fulltext/ED386023.pdf

*Kubota, M. (1996). The effects of instruction plus feedback on Japanese university students of EFL: A pilot study. *Bulletin of Chofu Gakuen Women's Junior College, 18*, 59–95.

*Leow, R. P. (1998). The effects of amount and type of exposure on adult learners' L2 development in SLA. *The Modern Language Journal, 82*, 49–68. https://doi.org/10.1111/j.1540-4781.1998.tb02593.x

*Loewen, S., & Erlam, R. (2006). Corrective feedback in the chatroom: An experimental study. *Computer Assisted Language Learning, 19*, 1–14. https://doi.org/10.1080/09588220600803311

Marsden, E., Mackey, A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1-21). New York, NY: Routledge.

*Muranoi, H. (2000). Focus on form through Interaction enhancement: Integrating formal instruction into a communicative task in EFL classrooms. *Language Learning, 50*, 617–673. https://doi.org/10.1111/0023-8333.00142

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*, 417-528. https://doi.org/10.1111/0023-8333.00136

Oswald, F. L., Friede, A. J., Schmitt, N., Kim, B. H., & Ramsay, L. J. (2005). Extending a practical method for developing alternate test forms using independent sets of items. *Organizational Research Methods*, *8*, 149–164. https://doi.org/10.1177/1094428105275365

Plonsky, L., Marsden, E., Crowther, D., Gass, S., & Spinner, P. (in press). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research*. Retrieved from https://doi.org/10.1177%2F0267658319828413.

*Rosa, E. M., & Leow, R. P. (2004). Awareness, different learning conditions, and second language development. *Applied Psycholinguistics, 25*, 269–292. https://doi.org/10.1017/S0142716404001134

*Sauro, S. (2009). Computer-mediated corrective feedback and the development of L2 grammar. *Language Learning & Technology, 13*, 96–120. https://doi.org/10125/44170

*Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *TESOL Quarterly, 41*, 255–283. https://doi.org/10.1002/j.1545-7249.2007.tb00059.x

Spinner, P., & Gass, S. (2019). *Using judgments in second language acquisition research*. New York, NY: Routledge.

Stewart, J., & Gibson, A. (2010). Equating classroom pre and post tests under item response theory. *SHIKEN: JALT Testing & Evaluation SIG Newsletter*, *14*(2), 11–18. http://hosted.jalt.org/test/ste_gib1.htm

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics, 38*, 1229–1261. https://doi.org/10.1017/S014271641700011X

*Tode, T. (2007). Durability problems with explicit instruction in an EFL context: The learning of the English copula be before and after the introduction of the auxiliary be. *Language Teaching Research, 11*, 11–30. https://doi.org/10.1177/1362168806072398

Weir, C. J., & Wu, J. R. W. (2006). Establishing test form and individual task comparability: A case study of a semi-direct speaking test. *Language Testing, 23,* 167–197. https://doi.org/10.1191/0265532206lt326oa

Wendler, C. L. W., & Walker, M. E. (2016). Practical issues in designing and maintaining multiple test forms. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 433–449). New York, NY: Routledge.