A Systematic Review of Research Designs and Tests Used for Quantification of Treatment Effects in *ARELE* 13–28

Shuichi TAKAKI
Fukushima University

Akira HAMADA
Nihon University

Keisuke KUBOTA
Graduate School, Fukushima University

ARELE
— Annual Review of English Language Education in Japan —
Volume 29
March 2018 所載論文
Published by
THE JAPAN SOCIETY OF ENGLISH LANGUAGE EDUCATION

A Systematic Review of Research Designs and Tests Used for Quantification of Treatment Effects in *ARELE* 13–28

Shuichi TAKAKI
Fukushima University
Akira HAMADA
Nihon University
Keisuke KUBOTA
Graduate School, Fukushima University

Abstract

The present study reviewed the quality of quantitative research in *Annual Review of English Language Education in Japan* (*ARELE*) volumes 13 to 28 from two perspectives: research designs and test traits. In articles on foreign language pedagogy including *ARELE*, quantitative data have been used as scientific evidence to determine if a teaching method is effective. This requires us to revisit the types of research designs and tests that have been employed to quantify the efficacy of instructions used in English education in Japan. In this systematic review, we targeted research that claimed a causal relationship between a pedagogical intervention and its outcome and coded 398 *ARELE* articles according to research designs and test traits for a meta-analysis. Results showed that 59% of primary studies adopted less appropriate designs involving threats to internal validity. Similarly, research that used less appropriate tests was more likely to overestimate and/or underestimate the treatment effects. Although our research community has continuously provided information about statistical reforms, such as statistical power and statistical indices to be reported, findings further suggest that we should revisit methods of designing our quantitative research.

1. Introduction

Annual Review of English Language Education in Japan (*ARELE*) is an aggregation of knowledge regarding English education in Japan. A dominant methodology used in *ARELE* is quantitative research (Mizumoto, Urano, & Maeda, 2014), in which statistical methods are used to investigate the efficacy of a pedagogical intervention, defined as *treatment effects*, on individual and/or classroom problems. Accordingly, English education researchers in Japan have discussed how to improve language teaching practice by considering the quantified effects of a treatment as one of scientific evidence to be referenced.

However, Terasawa (2015) has pointed out that the quality of evidence provided by English education research conducted in Japan was relatively low (particularly in primary school English education). He argued that this low quality resulted from the unfamiliarity with quantitative research designs. Norris, Plonsky, Ross, and Schoonen (2015) recommend specifying the research design adopted to extrapolate a treatment effect based on numeric data. Nevertheless, Mizumoto et al. (2014) showed that only 28 out of 120 articles in *ARELE* used (quasi-)experimental designs for group contrasts and were acceptable to discuss treatment effects.

This study, therefore, set two goals: (a) to review the appropriateness of research designs frequently used in *ARELE* from the viewpoint of experimental validity and (b) to examine the bias of treatment effects reported in *ARELE* across research-design types using meta-analysis. Before that, we begin with summarizing the importance of considering threats to experimental validity in designing quantitative research.

1.1 Experimental Validity in the Estimation of Treatment Effects

Quantitative research uses test scores as evidence of the efficacy of educational treatments. The quality of the results depends on whether test scores really reflect the treatment effects, which is regarded as experimental validity, or comprehensive validity of quantitative research. Reichardt (2009) classified experimental validity into *construct validity*, *internal validity*, *external validity*, and *statistical conclusion validity*. Dörnyei (2007) made a similar classification as follows:

Experimental validity	→	Measurement validity	Research validity
		= Construct validity	= Internal and external validity

In conjunction with this framework, the following sections will overview the factors that affect the appropriateness of quantitative research for a further meta-analysis.

1.1.1 Measurement Validity: Threats to Construct Validity

Construct is learners' latent traits, abilities, or characteristics measured by a test. Because psychological construct cannot be measured directly, the construct must be inferred from performance elicited by a test (Taylor, 2013). In other words, construct validity is not simply a property of a test, but rather represents to what extent inferences or interpretations derived from the test outcomes are appropriate (Messick, 1996).

One major threat to construct validity is construct irrelevant variance, which indicates that factors irrelevant to the construct influence test outcomes (e.g., Messick, 1996). For example, when a study aims to quantify a treatment effect on second language (L2) reading skills, the use of a written recall test in L2 causes construct irrelevant variance. Because the test requires learners to write what they understood in L2, the test score should reflect both writing and reading skills in L2. Thus, invalidated tests prevent the accurate estimation of a treatment effect.

Another threat to the appropriateness of score interpretations refers to test reliability (e.g., Messick, 1996). Even if a test is developed to reflect a targeted construct, the inconsistency of test scores reduces the validity of the estimated treatment effect. Plonsky and Gass (2011) showed that effect sizes differed by whether reliability coefficients were reported ($k = 61$, $d = 0.42$) or not ($k = 45$, $d = 0.96$) across the international journals on applied linguistics. The difference in test difficulty also distorts the results when researchers use alternate forms in pretest-posttest contrasts.

Mackey and Gass (2015) recommend the use of tests with validity and reliability proven by pilot and/or previous studies. However, Sakai and Koizumi (2014) demonstrated that in *Kanto-Koshinetsu Association of Teachers of English (KATE) Journal*, 25 out of 63 tests (40%) were newly created, seven tests (11%) were used with a modification, and only 31 tests (49%) were taken from previous studies. Moreover, reliability coefficients were reported in 21 out of 47 tests (45%), and interrater-reliabilities were included in five out of 20 tests (25%). According to Mizumoto (2014), 74 out of 186 *ARELE* articles (40%) specified reliability indices. These results require us to reconsider how the research outcomes that we have accumulated so far would vary according to the test traits.

1.1.2 Research Validity 1: Threats to Internal Validity

Internal validity refers to the appropriateness of causal inference derived from experimental results (Cook & Campbell, 1979), or to what extent the research outcomes reflect the effects of instrumental variables (Dörnyei, 2007; Taylor, 2013). Reichardt (2009) has indicated that most threats to internal validity occur due to a less appropriate research design and/or procedure adopted in an experiment. According to Haebara, Ichikawa, and Shimoyama (2001), the following six factors threaten internal validity in educational settings:

1. *Maturation:* Participants' growth attributed to the passage of time influences the outcome.
2. *History:* Participants' changes attributed to external events influence the outcome.
3. *Testing:* Participants' changes attributed to a pretest influence the outcome.
4. *Instrumentation:* Changes from pretest to posttest influence the outcome.
5. *Statistical regression:* Regression to the mean influences the outcome.
6. *Selection bias:* Inherent differences in participants influence the outcome.

For instance, selection bias was seen in Hagiwara and Ojima (2007), which concluded that English education for early elementary children negatively influenced their first language (i.e., Japanese) proficiency. The evidence they provided was that the Japanese test scores of the children who experienced early English education were lower than those who did not. However, Hoshino (2009) reanalyzed the data and found no difference between the groups after controlling for factors such as the academic background of the parents. In the same way as Terasawa (2015) emphasized the risk of selection bias, a variety of threats to internal validity trick us into reaching incorrect conclusions about the efficacy of particular pedagogical interventions.

The most appropriate way to exclude these threats to internal validity is random assignment, or a "randomization procedure for assigning the experimental units to the treatment levels" (Kirk, 2009, p. 23). However, random assignment would be difficult in educational settings, particularly when researchers use an intact class for their practice. Alternatively, a practical and effective way to alleviate threats to internal validity is to control uninteresting factors (e.g., Dörnyei, 2007) or to elaborate research designs (e.g., Reichardt, 2009; Taylor, 2013). As explained in Section 1.2, for example, the use of a research design that compares two groups at pretreatment and posttreatment resists threats to internal validity.

1.1.3 Research Validity 2: Threats to External Validity

External validity refers to the generalizability of causal inference derived from an outcome of a study (Cook & Campbell, 1979), or to what extent the inference can be applied to situations in which treatment, students, times, outcome variables, and/or settings are different (Reichardt, 2009). Even an experiment with high internal validity can be under threats to external validity. A result obtained from one study cannot always be applied to another because the treatment effect depends on the interaction between the treatment and the participants (Taylor, 2013).

Whereas random sampling is a typical procedure for ensuring external validity, replication and meta-analysis studies can also establish generalizability if a diverse group of participants is collected (Plonsky, 2012). Thus, external validity is not supported by individual research unless it adopts a random sampling procedure in large-scale experiments and/or surveys. Given that a lot of research in *ARELE* is a hypothesis-generating type based on individual researchers' interests (Mizumoto et al., 2014), the present study did not investigate the external validity of each article. However, we will later discuss how to generalize the results of treatment effects.

1.2 Quantitative Research Designs

Quantitative research begins with an experimental design proposing how to collect numerical data. Appropriate research design ensures the validity of the quantification of treatment effects. Many researchers have classified quantitative research into two types: correlational and experimental (e.g., Dörnyei, 2007; Haebara et al., 2001; Mackey & Gass, 2015). The first goal of correlational research is to "determine whether a relationship exists between variables and, if so, the strength of that relationship" (Mackey & Gass, 2015, p. 189). On the other hand, experimental research in English language education attempts to establish a causal connection between a pedagogical intervention and its outcome (Terasawa, 2015). Because the present study was interested in treatment effects, or the causal relationships between independent variables (i.e., treatments) and dependent variables (e.g., language skill improvements), we focused on the types of experimental research designs to create a coding scheme for our meta-analysis.

1.2.1 Experiments

Randomized controlled trial (RCT). Kirk (2009) characterized experimental research as (a) the manipulation of one or more independent variables; (b) the comparison between treatment and control groups, randomly assigned; and (c) the use of a blind procedure to reduce experimenter-expectancy effects. In any field, researchers emphasize the importance of random assignments because of the need to ensure that participants sampled from a particular population had an equal and independent opportunity in a sampling procedure (e.g., Haebara et al., 2001; Kirk, 2009; Mackey & Gass, 2015). A blind procedure is sometimes excluded from the definition of RCT (see a meta-analysis of Norris & Ortega, 2000). If we strictly adopt the above three requisites for RCT, there would be few experiments in *ARELE* considered "pure."

Repeated measures design. Haebara et al. (2001) suggested the validity of balancing the effects of mediators involved in participant traits with a repeated measures design (hereafter, RM). They stated that when multiple treatments are within-participants variables, in theory, individual differences in participants can be counterbalanced. Although the pretest-posttest contrasts derived from each participant appear to employ RM design (Morris & DeShon, 2002), following Haebara et al., the present study classified *ARELE* articles as an RM design only when the researchers counterbalanced the order of treatments and measurements to avoid order and testing effects.

In educational research, pure experiments are not always workable due to many limitations, such as research ethics and practical constraints. Particularly, most research designs used in foreign language pedagogy were in the framework of quasi-experiments when they employed intact classes (Norris & Ortega, 2000; Plonsky & Gass, 2011). Quasi-experiments include various forms of research designs; however, researchers should recognize the threats to internal validity involved in each design (Haebara et al., 2001; Reichardt, 2009). In the following sections, we will overview four quasi-experimental designs that have been commonly adopted in English language education in Japan in conjunction with possible threats to experimental validity.[1]

1.2.2 Quasi-Experiments

One-group-pretest-posttest design. The research scheme of this design (hereafter, OGPP) is [Pretest | Treatment | Posttest]. Haebara et al. (2001) introduced it as one of problematic designs. On the basis of Reichardt (2009), we should account for five types of threats to internal validity: maturation, history, statistical regression, testing, and instrumentation (see Section 1.1.2). When implementing a relatively long-term intervention (e.g., effects of a particular program on listening skills in Grades 4–6), researchers must reject an alternative interpretation that the change in children's development is simply because they are cognitively matured after the pretest (i.e., maturation). In the same way, the threat of history is critical unless researchers provide evidence that the score changes between a pretest and a posttest do not reflect any other untargeted variables such as extracurricular activities. In a longitudinal study with no comparison groups, a variety of measurements must be used for further statistical analyses to remove as many mediator variables as possible (e.g., Terasawa, 2015). In'nami (2012) has also stated that the OGPP design should not be used because it causes regression to the mean effect.

Testing and instrument effects will occur when using invalidated tests. If a pretest and a posttest were totally the same, taking the pretest would improve the posttest score as a testing effect. Even if researchers used different tests, substantial differences such as difficulty, constructs, and procedures would cause an unreliable estimation of the change occurring between the two tests. While Norris et al. (2015) have recommended reporting assessment reliability, Plonsky and Gass (2011) showed that only 64% of primary studies reported either rater or instrument reliability. Regarding *ARELE*, Mizumoto et al. (2014) did not focus on test reliability; therefore, we classified the test traits described in *ARELE* articles in terms of reliability and comparability of test scores.

Nonequivalent-group design. Nonequivalent-group design, in which "different participants receive different treatments and the relative effectiveness of the treatment is assessed by comparing the performances of the participants across the different groups" (Reichardt, 2009, p. 54), is applied to remove the five threats to internal validity. Whereas a group contrast is possible without a pretest, visualized as [Group A: Treatment | Posttest; Group B: No Treatment | Posttest] (hereafter, NEGP), the threat of selection bias occurs because there are no means of ensuring that, for example, the participants in Group A might have higher language skills than those in Group B at the beginning of research (e.g., Haebara et al., 2001).

Many researchers agree that the use of both pretests and posttests in a nonequivalent-group design (hereafter, NEGPP) is one of the most credible designs in quasi-experiments [Group A: Pretest | Treatment | Posttest; Group B: Pretest | No Treatment | Posttest] (Dörnyei, 2007; Haebara et al., 2001; Mackey & Gass, 2015; Reichardt, 2009). According to Reichardt (2009), the role of a pretest is to adjust the effects of selection bias. For example, the mean difference of change scores (i.e., the mean gain scores from the pretest to the posttest) between two groups is regarded as the estimate of the treatment effect. Another statistical procedure for removing the selection bias is to use the propensity score analysis. In our meta-analysis, *ARELE* articles were also coded in terms of the types of research designs to examine the negative effects of threats to internal validity.

1.3 Research Questions

To examine the issues discussed so far, namely, (a) the appropriateness of research designs adopted in *ARELE* from the viewpoint of experimental validity, and (b) the variance of treatment effects reported in *ARELE* possibly caused by threats to experimental validity, the present study is designed to answer the following three research questions (RQs):

RQ1. What types of research designs and tests are used in quantitative research of *ARELE*?
RQ2. To what extent do treatment effects reported in *ARELE* differ across research designs?
RQ3. To what extent do treatment effects reported in *ARELE* differ across test traits?

2. Method

2.1 Study Retrieval, Coding, and Inclusion

Retrieval. Because Mizumoto et al. (2014) revealed that the number of intervention studies in *ARELE* had been increasing since volume 13, we determined to submit articles in *ARELE* volumes 13–28 (2002 through 2017) into a meta-analysis. In the study search, we used J-STAGE (Japan Science and Technology Information Aggregator, Electronic) to retrieve articles in *ARELE* and found 363 articles. Thirty-five articles were not available in J-STAGE; we obtained their print version from a library. Accordingly, 398 articles were included in further analyses.

Coding. Three independent coders completed coding by use of a scheme shown in Table 1. It was composed of (a) research frameworks, (b) research purposes, (c) research domains, (d) research designs, and (e) test traits. Three pairs of coders separately coded 72 articles in *ARELE* volumes 26–28 to check if the scheme worked well, and then made a final version for the coding (intercoder reliability = 83%). Any disagreements were solved through discussion. Particularly, if studies can be classified into two or more characteristics in the research frameworks, domains, and designs, we discussed which were dominant in the same way as Mizumoto et al. (2014). Some studies adopted two or more research purposes and multiple tests. In this case, we recorded all the purposes and test traits for each article. For the remaining 326 articles, three coders independently coded each of them.

Inclusion. For the present meta-analysis, we defined the following inclusion criteria:

- The article must report quantitative data (389 out of 398 articles).
- The article (i.e., the research categorized as Treatment Effect) must state that its purpose was to examine the efficacy of pedagogical interventions, such as language teaching, materials given to learners, and educational programs for Japanese learners of English (94 out of 389 articles).
- The article must target one of the research domains listed in Table 2; however, any articles were excluded when they used the data obtained from psycholinguistics experiments such as reaction time (85 out of 94 articles).

Finally, 85 articles met our inclusion criteria. The group-contrast designs included 4,937 participants (treatment: n = 2,535, control/comparison: n = 2,402) and the pretest-posttest contrast designs included 5,470 participants. There were no studies that adopted the one-group-posttest-only design to estimate the effects of an intervention in *ARELE* volumes 13–28.

Table 1

Coding Scheme: Categories and Definitions

1. Research frameworks
   - *Correlational:* Quantitative data is used to reveal any relationships between variables from the same group of participants.
   - *Causal:* Quantitative data is used to determine if independent variables cause the changes in dependent variables.
2. Research purposes
   - *Treatment effect:* Quantitative data is used to estimate the effects of a pedagogical intervention.
   - *Cognition:* Quantitative data is used to investigate learner cognition involved in language learning with a psycholinguistic approach.
   - *Quality:* Quantitative data is used to describe the research outcomes qualitatively.
   - *Development:* Quantitative data is used to develop materials, such as tests and questionnaires.
   - *Other:* Quantitative data is not used for the above purposes.
3. Research domains
   - Listening, Reading, Speaking, Writing, Vocabulary, Grammar, Motivation, and Other.
4. Research designs
   - *RCT*: Participants are randomly allocated to either a treatment or a control group. Information about the groups is masked from both researchers and participants (i.e., a blinded experiment).
   - *RM*: Participants are assigned two or more educational treatments. The differences between the independent variables are compared to estimate the effects of a targeted treatment.
   - *OGPP*: There is no comparison group. Treatment effects are estimated by the contrast between a pretest and a posttest.
   - *NEGP*: Participants are allocated to either a treatment or a control group with no randomization. Treatment effects are estimated by the contrast between posttest scores of the two groups.
   - *NEGPP*: Participants are allocated to either a treatment or a control group with no randomization. Treatment effects are estimated by the contrast between posttest scores of the two groups. Pretest scores are used to ensure the homogeneity of the two groups.
5. Test traits
   - *Test used*: The tests used in pretests and posttests are the same or different.
   - *Test reliability*: Scoring reliability (instrument or interrater) are reported or not.
   - *Test equation*: Test difficulty differences are adjusted in some manner, such as a pilot study.

2.2 Meta-Analytic Procedures

All means, standard deviations, and the number of participants were inserted into “metafor” version 1.9-9 for meta-analysis package of R, and transformed into Cohen’s *d* for further analysis. Some of the 85 *ARELE* articles reported multiple datasets for their targeted treatment effects.

Because multiple effect sizes computed from the same sample would over- or underestimate the efficacy of pedagogical interventions, some adjustments based on Lipsey and Wilson (2001) were applied as listed, and we finally obtained 123 datasets for the meta-analysis:

- If two or more tests were used to measure similar constructs for the same sample (e.g., writing complexity evaluated in terms of lexis and syntax), the effect sizes were averaged.
- If two or more tests were used to measure similar constructs for the same sample but their score units were different (e.g., reading comprehension skills evaluated in a multiple-choice test and reading speed), an effect size that showed the largest one was employed.
- If two or more experimental groups were used for comparisons with one control group, an effect size that showed the largest one was employed.

Two types of formulas for Cohen's *d* were used according to the research designs.[2] For the group-contrast design, the following formula was used:

$$d = \frac{M_{exp.} - M_{con.}}{\sqrt{\frac{SD^2_{exp.} + SD^2_{con.}}{n_{exp.} + n_{con.}}}}$$

For the pretest-posttest contrast design, the average of two standard deviations of the means of the pretests and posttests were applied for the denominator (Lipsey & Wilson, 2001). To estimate the aggregated effect size of *d*, this study applied a random-effects model in the same manner as Mizumoto et al. (2014). Moderator analyses were further run by use of the random-effects model in terms of research designs and test traits.

3. Results and Discussion

3.1 Research Designs and Test Traits Used in *ARELE* (RQ1)

In *ARELE* volumes 13–28, 389 out of 398 primary studies reported quantitative data. Most research frameworks were causal (k = 196 [50%]). The number of correlational research studies was relatively small (k = 42 [11%]). The coding scheme did not have frameworks for descriptive studies, in which quantitative data was used to describe characteristics of a sample population; however, 113 (29%) studies were categorized as descriptive. Additionally, 38 (10%) studies could not be classified into any frameworks. As stated in Section 2.2, 85 articles including 123 datasets that examined treatment effects were further analyzed in terms of design variability and test traits among research domains.

Table 2 shows the types of research designs reported in *ARELE* among research domains; OGPP was adopted by the largest number of studies (42%), followed by NEGPP (30%), NEGP (17%), and RM (11%). Only 41% of studies in *ARELE* appropriately included a control group to estimate treatment effects (i.e., NEGPP and RM). Even considering that using a control group is not practical in educational settings, only 30% of studies used NEGPP, which was recommended as a quasi-experimental design (Haebara et al., 2001). Compared to this, the meta-analysis by Norris and Ortega (2000) showed about 60% of L2-interaction studies used NEGPP. In addition, the large-scale meta-analysis by Plonsky and Oswald (2014) indicated that 67 out of 92 (73%) L2 studies for treatment effects adopted NEGPP. The small number of RM is also attributed to its impracticability. Multiple treatments and tests in a single experiment are a burden to students. Kusanagi (2014) has suggested other approaches such as the use of an orthogonal array to improve the practicability of experiments.

Research designs also differed within research domains. First, grammar studies had the highest percentage of NEGPP designs (86%), followed by speaking (50%), vocabulary (35%), listening (33%), writing (29%), and motivation (15%). Reading studies were the least likely to use NEGPP in the other research domains. Second, motivation studies were the mostly to use OGPP (77%), followed by writing (64%), listening (42%), and speaking (42%). These percentages suggest that research domains except grammar have provided evidence threatened by internal validity.

Table 2

Types of Research Designs Among Research Domains

	NEGP		NEGPP		OGPP		RM	
Domain	*k*	%	*k*	%	*k*	%	*k*	%
Listening	3	25	4	33	5	42	0	0
Speaking	1	8	6	50	5	42	0	0
Reading	10	40	1	4	8	32	6	24
Writing	1	7	4	29	9	64	0	0
Vocabulary	6	26	8	35	5	22	4	17
Grammar	0	0	12	86	1	7	1	7
Motivation	0	0	2	15	10	77	1	8
Other	0	0	0	0	9	90	1	10
Total	21	17	37	30	52	42	13	11

Table 3 shows the test traits reported in *ARELE* among research domains. Because research designs except for NEGP use two or more tests which quantify treatment effects (Mackey & Gass, 2015), the comparability between alternative tests has to be ensured. Repeated use of the same test as both tests causes the threats of testing (e.g., practice effects). About a half of studies in *ARELE* used the same tests; they carried a risk of threats to measurement validity. Particularly, motivation (85%) and grammar (71%) studies more frequently used the same tests than the others, especially reading (8%) and listening (33%). Mizumoto et al. (2014) reported that the majority of research type in *ARELE* was hypothesis-generating, in which researchers sometimes developed an original test according to their interests. In fact, the present study further categorized the tests into originally-developed or previously-used, and 31% cases were created for the focal research. For example, some grammar studies tested the effects of an instruction on the acquisition of particular grammatical items as the targeted construct, which led to the originally-developed, same test use.

Among studies using different tests such as pretest and posttest, 73% of them implemented test equation in several ways. Many researchers adopted tests equated in previous studies and/or conducted a pilot study to ensure if the tests would work well for their participants. It shows that the difference between test scores reported in those studies can be interpreted as one of evidence for treatment effects. Although a test-retest method (i.e., the same test use) is appropriate in terms of test equality, the validity threats in terms of instrumentation cannot be rejected (Okada, 2015).

Reports of test reliability is required regardless of research designs (e.g., Norris et al., 2015). However, only 31% *ARELE* articles reported reliability indices. This result is similar to that of *KATE Journal* (Sakai & Koizumi, 2014), and unfortunately, this discipline is not as familiar as international journals on applied linguistics (Plonsky & Gass, 2011).

Table 3

Test Traits Among Research Domains

	Tests used				Test reliability				Test equation			
	Different		Same		Yes		No		Yes		No	
Domain	*k*	%	*k*	%	*k*	%	*k*	%	*k*	%	*k*	%
Listening	8	67	4	33	3	25	9	75	4	100	0	0
Speaking	5	42	7	58	3	27	8	73	0	0	2	100
Reading	23	92	2	8	15	68	7	32	9	100	0	0
Writing	6	43	8	57	0	0	13	100	2	40	3	60
Vocabulary	11	48	12	52	2	9	21	91	3	100	0	0
Grammar	4	29	10	71	4	40	6	60	1	25	3	75
Motivation	2	15	11	85	7	54	6	46	1	50	1	50
Other	4	40	6	60	1	10	9	90	4	100	0	0
Total	63	51	60	49	35	31	79	69	24	73	9	27

Note. Nine datasets were not applicable for test reliability because they included pause-duration of speech and running words as fluency in writing. In the use of different tests, 30 datasets did not compare test scores between the alternative forms and those omitted from the test-equation category.

3.2 Research Validity: Different Effect Sizes in Different Research Designs (RQ2)

Table 4 presents descriptive statistics of aggregated *d* values for each research design. As the boxplots in Figure 1 show their distributional properties, the comparisons between NEGP and NEGPP showed a different pattern. Whereas both *d* values were distributed equally from each median, the means were higher in NEGPP than NEGP. It cannot be concluded whether the NEGP studies would have obtained the same effect sizes if they had implemented a pretest; however, it is possible that the differences between treatment and control groups before an intervention affected the estimation of treatment effects. The NEGPP studies provided evidence that there were no differences in pretest scores between two groups with an analysis of (co)variance, which was often applied for post-hoc comparisons (Haebara et al., 2001; Kirk, 2009; Reichardt, 2009). Some experiments discourage pretests because they "might alert participants to what the treatment is about" (Mackey & Gass, 2015, p. 203); in those cases, group contrasts must be validated using other measures such as class grades, teachers' ratings, and placement test results (Dörnyei, 2007).

The pretest-posttest contrast showed large positive skews. In other words, some studies that used the OGPP design may have overestimated their treatment effects. The range from the first quartile to median is wider than the range from median to the third quartile. This is consistent with the result that the pretest-posttest contrasts showed higher effect sizes than the group contrasts in L2 pedagogical treatments (Plonsky & Oswald, 2014). Thus, readers of *ARELE* should carefully interpret the efficacy of an intervention that was investigated from both group and pretest-posttest contrasts. Researchers can also provide interpretable effect sizes by use of a reduction formula for between-group and pretest-posttest contrasts developed by Morris and DeShon (2002).

Table 4

Moderator Analysis of Effect Sizes Based on Research Designs

		Estimated effect sizes					Heterogeneity			
Design	*k*	*d*	95% CI	*SE*	*z*	*p*	*Q*	*df*	*p*	I^2
NEGP	21	0.34	[0.10, 0.59]	0.13	2.72	.007	81.07	20	.000	75%
NEGPP	37	0.66	[0.46, 0.85]	0.10	6.70	.000	160.90	36	.000	78%
OGPP	52	0.75	[0.59, 0.91]	0.08	9.35	.000	631.63	51	.000	92%
RM	13	0.78	[0.46, 1.09]	0.05	4.87	.000	86.65	12	.000	86%

We cannot discuss in detail the validity of the RM design adopted in *ARELE* because of the relatively small sample size. First, the *Q*-value (the degree of heterogeneity among the effect sizes taken from each research) did not differ from the NEGPP design but greatly differed from the OGPP one. In this respect, the RM design can be applied in the same manner as the NEGPP design because the counterbalanced order of measurements between the treatment and control sessions reduces the threat of testing effects that appear in the OGPP design.

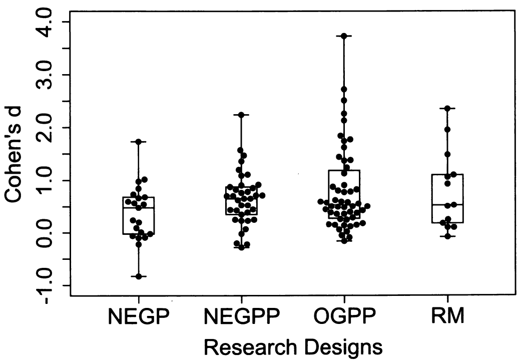

Figure 1. Boxplots of estimated effect sizes based on different research designs.

3.3 Measurement Validity: Different Effect Sizes in Different Test Traits (RQ3)

Table 5 presents descriptive statistics of aggregated *d* values for each test trait. As shown in the left boxplot in Figure 2, both distributions were positively skewed regardless of whether the pretests and posttests were identical. The mean effect size of identical tests was larger than that of different tests. Interestingly, only three studies that used identical tests showed negative effect size, whereas 11 studies that used the different tests did. As Reichardt (2009) suggested, this result implies that even if a treatment had no effect or a negative effect in truth, the use of the same test as a pretest and a posttest could lead to overestimation of a treatment effect.

Table 5

Moderator Analysis of Effect Sizes Based on Test Traits

		Estimated effect sizes					Heterogeneity			
Test traits	*k*	*d*	95% CI	*SE*	*z*	*p*	*Q*	*df*	*p*	I^2
Tests used										
Different	63	0.59	[0.44, 0.74]	0.08	7.79	.000	399.51	62	.000	84%
Same	60	0.73	[0.57, 0.88]	0.08	9.35	.000	651.59	59	.000	91%
Test reliability										
Yes	35	0.48	[0.28, 0.67]	0.10	4.75	.000	172.70	34	.000	80%
No	79	0.75	[0.62, 0.88]	0.07	11.06	.000	861.54	78	.000	91%
Test equation										
Yes	24	0.77	[0.53, 1.01]	0.12	6.33	.000	170.16	23	.000	86%
No	9	0.44	[0.02, 0.87]	0.22	2.04	.041	18.28	8	.019	56%

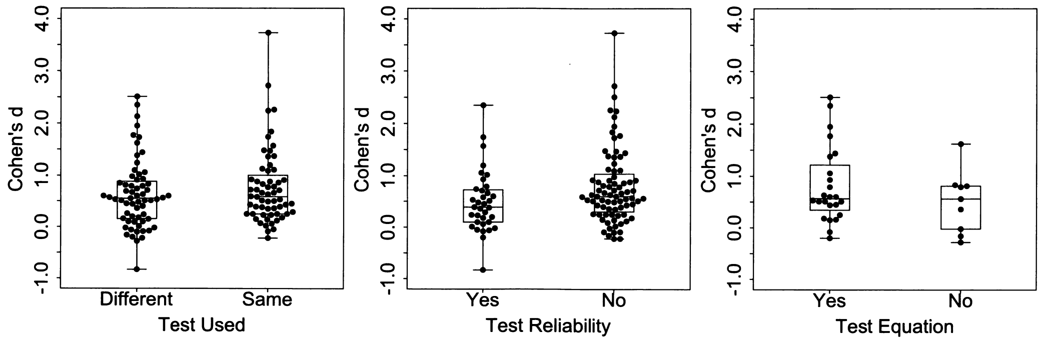

Figure 2. Boxplots of estimated effect sizes based on different test traits.

As for the reliability report, the middle boxplot in Figure 2 illustrates that the distribution of No was positively skewed. Moreover, the mean effect size of No was larger than that of Yes, and the 95% CI of No hardly overlapped that of Yes. This result is fully consistent with Plonsky and Gass (2011) in that studies without reporting reliability indices obtained larger effect sizes.

As the right boxplot in Figure 2 shows, although the sample size was small, the distribution of Yes was positively skewed whereas that of No was dispersed widely, and the mean effect sizes varied between Yes and No. Given that the tests that are adjusted in terms of those difficulty are suitable for measurements, these results indicate that the researchers underestimated the treatment effects of a teaching method when they employed the not-equated tests. In *ARELE* volumes 13–28, only five studies used a test equated through item response theory, a systematic and rigorous method. Whereas an originally created measurement is sometimes required to answer an original research question, this finding shows the need for accurate estimations of treatment effects. When specific research fields mature, the researchers must then ensure both internal and external validity by replicating and synthesizing research (Plonsky, 2012) with validated measurements.

4. Conclusions

The present study systematically reviewed the quality of quantitative research reported in *ARELE* volumes 13–28. The general findings suggest that our research community should pay more attention to designing the quantitative research involved in data collection. While there are different research agendas to be followed in different research domains, the effect sizes derived from individual studies must be interpreted according to the same framework, such as research design (e.g., Plonsky & Oswald, 2014). Particularly, researchers must consider what types of threats to experimental validity will occur when designing their own research. As Plonsky and Gass (2011) as well as the present study have demonstrated, we need to recognize the danger of low-quality experiments because they may distort our interpretation of a pedagogical intervention's effects. Research outcomes are a resource for determining what kinds of treatments, instructions, and programs should be introduced into a classroom (Terasawa, 2015); therefore, our research methods must be improved for our profession to improve.

Notes

1. A quasi-experiment has two more research designs that are used instead of NEGPP: interrupted time-series design and regression-discontinuity design. Although we did not enter these designs into our systematic review, because keyword search results showed no *ARELE* studies adopted them, they are recognized as credible compared to OGPP (e.g., Reichardt, 2009).
2. We used the formula of Cohen's *d*, in which the effect size is weighed by a sample variance, to examine the differences in the effect sizes among research designs. Although Hedges' *g*, which uses an unbiased sample variance, was computed in the same manner as Mizumoto et al. (2014), the results showed no significant differences between the two effect sizes.

Acknowledgements

This study was supported by JSPS Grants-in-Aid for Young Scientists (B) No. 26770188 and No. 17K13512. We appreciate three reviewers' suggestions to improve our earlier manuscript.

References

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston, MA: Houghton-Mifflin.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford University Press.

Haebara, T., Ichikawa, S., & Shimoyama, H. (Eds). (2001). *Shinrigaku kenkyuhou nyumon: Chousa, jikken kara jissen made* [Introduction to research methods in psychology: Surveys, experiments and practices]. Tokyo, Japan: University of Tokyo Press.

Hagiwara, H., & Ojima, S. (2007, November). *Brain science and foreign language education: Preliminary results of a three-year cohort study using ERP and NIRS*. International Mind, Brain and Education Society Conference 2007, Fort Worth, TX.

Hoshino, T. (2009). *Chousa kansatsu data no toukei kagaku: Ingasuiron, sentaku bias, data yugou* [Statistical science of observational research data]. Tokyo, Japan: Iwanami Shoten.

In'nami, Y. (2012). Test tokuten kaishaku no ryui-ten [Considering test score interpretation]. In Y. Ushiro (Ed.), *Eigo reading test no kangaekata to tsukurikata* [How to design an English reading test: From theory to practice] (pp. 78–87). Tokyo, Japan: Kenkyusha.

Kirk, R. E. (2009). Experimental design. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 23–45). London, England: SAGE.

Kusanagi, K. (2014). Gaikokugo kyouiku to chokkouhyo o mochiita jikken keikaku: Jikken keikaku no kouritsuka o motomete [The use of an orthogonal array in research on foreign language pedagogy for practicability]. *Reports Vol. 4 of 2013 Studies in Japan Association for Language Education and Technology, Kansai Chapter, Methodology Special Interest Group (SIG)*, *4*, 24–33. Retrieved from http://www.mizumot.com/method/04-03_Kusanagi.pdf

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: SAGE.

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

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*, 241–256. doi:10.1177/026553229601300302

Mizumoto, A. (2014). Sokuteisareta sa no bunseki to kaishaku [Analyses and interpretations of mean differences]. In the Japan Society of English Language Education (Ed.), *Trends and current issues in English language education in Japan: Integrating theory and practice* (pp. 376–380). Tokyo, Japan: JASELE.

Mizumoto, A., Urano, K., & Maeda, H. (2014). A systematic review of published articles in *ARELE* 1–24: Focusing on their themes, methods, and outcomes. *ARELE: annual review of English language education in Japan*, *25*, 33–48. doi:10.20581/arele.25.0_33

Morris, S. B., & DeShon, R. P. (2002). Combining effect-size measures in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*, 105–125. doi: 10.1037/1082-989X.7.1.105

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

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning*, *65*, 470–476. doi:10.1111/lang. 12104

Okada, K. (2015). Reliability in psychology and psychological measurement, with focus on Cronbach's alpha. *The Annual Report of Educational Psychology in Japan*, *54*, 71–83. doi: 10.5926/arepj.54.71

Plonsky, L. (2012). Replication, meta-analysis, and generalizability. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 116–132). Cambridge University Press.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*, 325–366. doi:10.1111/j.1467-9922.2011.00640.x

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912. doi:10.1111/lang.12079

Reichardt, C. S. (2009). Quasi-experimental design. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 46–71). London, England: SAGE.

Sakai, H., & Koizumi, R. (2014, August). *Eigokyoiku-kenkyu ni okeru data shushu-houhou: Hakaritaimono o hakaru houhou* [Data collection in English language education research: A methodology to measure targeted constructs]. Paper presented at the 38th Annual Conference on KATE, Chiba, Japan. Retrieved from http://www7b.biglobe.ne.jp/~koizumi/140824KATE2014modified_test_questionnaire_Sakai_Koizumi.pdf

Taylor, C. S. (2013). *Validity and validation: Understanding statistics*. Oxford University Press.

Terasawa, T. (2015). Eigo-kyouiku-gaku ni okeru kagakuteki evidence towa? Shogakkou eigo kyouiku seisaku o jireini [What is scientific evidence in English language education? The case of elementary school English education policy in Japan]. *Reports of 2014 Studies in The Japan Association for Language Education and Technology, Chubu Chapter, Fundamentals of Foreign Language Educational Research Special Interest Group (SIG)*, 15–30. Retrieved from https://www.letchubu.net/modules/bulletin/