Studies in Second Language Acquisition (2024), **46**, 251–276
doi:10.1017/S0272263123000311

METHODS FORUM

Meta-analysis of second language research with complex research designs

Reza Norouzian[1] and Gavin Bui[2]

[1]Oregon Department of Education, Salem, OR, USA; [2]Department of English, The Hang Seng University of Hong Kong, China
Corresponding author: Reza Norouzian; Email: rnorouzian@gmail.com

(Received 19 January 2023; Revised 29 March 2023; Accepted 09 May 2023)

Abstract

Meta-analyses play an instrumental role in informing second language (L2) theory and practice. However, current (i.e., classic) approaches to meta-analysis are limited in their ability to do so because they often fail to capture the complexity inherent in primary studies' research designs. As we argue in this article, when complex L2 studies are represented by simplistic meta-analyses, the latter cannot reach its potential to contribute to the development of cumulative knowledge. To mitigate this issue, we first discuss the fundamental problems of the classic approaches to meta-analysis of complex L2 research. Second, we introduce an alternative meta-analytic framework that will address those problems. Third, we apply the meta-analytic framework discussed to a complex L2 domain. Fourth, we offer free software to facilitate the conduct of the alternative meta-analytic approach described. Finally, we discuss the implications of this alternative framework for making evidence-based recommendations to the relevant stakeholders.

Introduction

In the last two decades or so, research syntheses have gained considerable traction in second language (L2) research (Chong & Plonsky, 2023; Raeisi-Vanani et al., 2022). One form of research synthesis that has become particularly popular in L2 research is the classic meta-analysis. The appellation *classic*[1] is mainly due to the traditional methods of

[1]We use the term *classic* to generally refer to the methods of meta-analysis that predate meta-regression. For studies with a complex research design (e.g., with multiple treatment groups, posttests, and outcomes), classic methods of meta-analysis often require aggregating effect-size estimates in each study before meta-analyzing them using a fixed- or a random-effects model. However, there is no unified framework as to how that aggregation may be done (e.g., arithmetic averaging, weighted averaging, correlated weighted or unweighted averaging). As a result, in practice there are several subjective actions that go into combining, aggregating, or even discarding (i.e., selective outcome reporting) the data depending on one's style, prior experience, available software, convenience, and/or common traditions in a research domain. In addition, due to the lack of a regression framework, classic methods of meta-analysis are not readily able to allow the focal moderator(s) to interact with the main aspects of the research designs. For example, understanding how

synthesizing the quantified evidence (i.e., effect-size estimates) obtained from the primary studies (Lipsey & Wilson, 2001). The main goal behind classic meta-analysis is to "summarize a stream of research with an average effect size" and make it available to a wide audience of researchers, practitioners, and other stakeholders (Raudenbush & Bryk, 1985, p. 94).

Undoubtedly, the common use of classic methods of meta-analysis has informed our understanding of L2 theory, research, and practice (e.g., see Plonsky et al., 2021, for a review of the state of meta-analysis in bilingualism). However, given the recent methodological reform movement taking place in the field (Byrnes, 2013; Gass et al., 2021; Norouzian, 2021; Norouzian et al., 2018, 2019; Norouzian & Plonsky, 2018) and a growth in the complexity of research designs (e.g., use of multiple treatment groups, multiple posttests, and multiple outcomes or measures) employed in L2 studies (Norouzian, 2020, 2021), a few questions about the use of classic methods of meta-analysis need to be addressed:

1- To what extent can classic methods of meta-analysis capture the complex nature designs found in L2 research?
2- If to a limited extent, how can we modernize our methods of meta-analysis so as to aptly capture the complex nature of L2 study designs?
3- What tools are available to L2 researchers for a transition into more flexible methods of meta-analysis?

The remainder of this article seeks to provide practical answers to these questions. First, we discuss the fundamental problems with the classic methods of meta-analysis that prevent them from capturing the complex nature of the study designs used in L2 research. Second, we introduce an alternative framework, *multivariate multilevel meta-analysis* (3M), for meta-analyzing L2 research with some of its additional capabilities discussed in our supplementary document available at: https://rnorouzian.github.io/m/3msup.html. Third, we apply the alternative framework to a well-known, yet complex L2 literature. Fourth, we offer free software programs to enable the practical use of the new framework discussed. Finally, we discuss the implications of this alternative framework for making evidence-based recommendations to the relevant stakeholders.

Fundamental problems with classic meta-analysis of L2 research

Sampling dependence inherent in L2 studies is ignored

As Norouzian (2020) has discussed, the design of empirical L2 research has grown over time in complexity. Specifically, L2 research often tends to involve one or more of the following design aspects: (a) inclusion of multiple treatment groups along with a control/comparison group, (b) measurement over multiple occasions, and (c) measurement over multiple outcomes.

Take, for instance, Shintani and Ellis (2013). The study investigated the effect of direct corrective feedback (D) and the metalinguistic explanation (M) on ESL (English

a focal moderator (e.g., proficiency) associates with the effect of a phenomenon of interest (e.g., corrective feedback's effect on written accuracy) may not be readily explored across measurement occasions (pre- and posttests) even though the primary studies are longitudinal. Readers may refer to the section titled "Finding may be biased" for more details.

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

as a second language) writers' implicit knowledge of the English indefinite article. To do so, the researchers adopted the design depicted in Figure 1.

As shown in Figure 1, Shintani and Ellis (2013) randomly assigned their classes (RC) to three groups (i.e., D, M, and control). Following an initial observation (O_1) of their participants' knowledge of English indefinite article, they provided their intended treatments to each group. The two follow-up observations (O_2 and O_3) were made to measure the possible change in participants' implicit knowledge of the English indefinite article over time in each group.

From this study, one can commonly obtain six effect-size estimates, in this case standardized mean differences (SMD; i.e., Hedges's *g*). Namely, we can obtain two pre-feedback effect estimates at O_1, two immediate, post-feedback effect estimates at O_2, and finally two delayed, post-feedback effect estimates at O_3, all comparing the treatment groups with the same control group.

However, these effect-size estimates are likely dependent on (i.e., correlated with) each other due to the design aspects (A) and (B) adopted in Shintani and Ellis (2013). For instance, at each observation time, the effect-size estimates across the treatment groups are likely dependent due to being obtained by comparing the performance of each treatment group against that of the same control group (design aspect A). Being a pivot point, if the control group put up a superior performance, all of the treatments' effects would become smaller. Conversely, if the control group put up a low performance, all treatments' effects would become larger. Thus, treatments' effects could vary together (i.e., be correlated) to a measurable degree as a function of their common control group's performance.

Additionally, the effect-size estimates across the occasions for each treatment group are also likely dependent due to being obtained over repeated measurements (design aspect B). Specifically, because the same participants were measured over the three occasions (O_1, O_2, and O_3) in each group, their measurements across these occasions contain an estimable degree of correlation, likely more strongly so between the adjacent occasions than those further apart (Trikalinos & Olkin, 2012). The same correlations carry forward to any statistic that is obtained from these measurements' summaries such as effect-size estimates inducing dependence among them.

Indeed, Shintani and Ellis (2013) also measured their participants using an additional grammar correction test as a way to examine the participants' explicit knowledge development on two occasions (O_1 and O_3). This would be considered a second outcome in the study. Thus, a new set of four effect-size estimates may be obtained from the same participants on this additional grammar correction test, if explicit knowledge development is of interest to the meta-analyst. If so, then this new set of four effect-size estimates on explicit knowledge development would be correlated with

RC	O_1	D	O_2	O_3
RC	O_1	M	O_2	O_3
RC	O_1		O_2	O_3

Figure 1. Design of Shintani and Ellis (2013).

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

each other due to the design aspects A and B. But now these effect-size estimates on explicit knowledge development for each treatment group would also be correlated with the previous set of effect-size estimates on implicit knowledge development (design aspect C).

Because correlated effect-size estimates often stem from the use of the summary data from the same sample of participants, this form of inherent dependence in the effect-size estimates is sometimes referred to as *sampling dependence* (Stevens & Taylor, 2009). Note that sampling dependence can occur in correlational L2 studies, too. For instance, if a correlational study examined the relationships of multiple subscales of L2 motivation with L2 achievement, the reported correlation coefficient estimates (i.e., effect-size estimates) for the subscales would likely be correlated with each other. This is because the same sample of participants has been measured on multiple subscales (design aspect C).

Classic methods of meta-analysis are ill-equipped to account for the sampling dependence inherent in effect-size estimates obtained from studies with one or more of the design aspects discussed above. Consequently, meta-analysts often find themselves making ad hoc decisions that oversimplify the nature of L2 studies (for a review of these decisions, see Hedges, 2019). In addition, ignoring the sampling dependence threatens the validity of inferential conclusions regarding the summary effects and accurately estimating the magnitude of the heterogeneity in the underlying effects (Pustejovsky & Tipton, 2022).

Hierarchical dependence in underlying effects of L2 studies is ignored

That L2 studies can often produce multiple effect-size estimates due to the design aspects discussed in the previous section introduces sampling dependence among them prior to their meta-analysis. Another form of dependence may arise when we intend to envision a meta-analytic model for such effect-size estimates obtained from the literature.

Part of the job of a meta-analytic model is to statistically link the effect-size estimates we obtain from the literature to their underlying population effects (i.e., true effects). In its simplest form, when a group of L2 studies can each provide multiple effect-size estimates, then, by virtue of belonging to the same study, their underlying true effects are likely more similar in each study than those in other studies. This similarity amounts to a measurable degree of overlap (i.e., correlation) among the true effects in each study forming clusters of true effects across the studies. Because this second form of correlation occurs owing to the clustered nature of the underlying true effects envisioned for the meta-analysis of studies with complex research designs, it is sometimes referred to as *hierarchical*[2] *dependence* (Stevens & Taylor, 2009).

Classic methods of meta-analysis, which to date are the norm in the field, are unable to adopt a meta-analytic model for L2 studies that sufficiently allows for the hierarchical

[2]3M is also capable of accommodating other forms of dependence in the true effects (or equivalently random effects) that are not hierarchical. For instance, when meta-analysts have a theoretical reason and/or empirical support (e.g., better model fit) that the different measures used within and across the individual studies could themselves produce similarity (i.e., overlap) in the true effects for each unique measure, then the dependence in true effects can separately arise at the study level as well as at the measure level, with no hierarchy existing between these levels. It is customary to refer to the true effects separately structured for, in this case, unique studies and measures as being crossed (or for true effects to be cross-classified by them) rather than hierarchical (for details see Fernández-Castilla et al., 2019).

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

dependence among their underlying true effects. This inability has negative effect on the conclusion validity of our meta-analytic findings such as incorrect precision in average effect estimates, inflated Type I error rates, incorrectly narrow confidence intervals, and incomplete delineation among the sources of variation in the true effects (Matt & Cook, 2019).

Comparison across categories of substudy moderators may be invalid: The k situation

Perhaps one of the fundamental needs of L2 meta-analysts is to be able to compare meta-analytic findings across the theoretically distinct categories of their moderators. As Norouzian (2021) discussed, L2 moderators such as the type of treatment (e.g., direct feedback vs. meta-linguistic explanation) are often found at the substudy levels of L2 studies. That is to say, such moderators can vary within the same studies. For instance, L2 studies can each examine multiple types of treatment. Thus, comparing different types of treatment in a literature often entails obtaining multiple effect-size estimates from a combination of the same and different studies. This results in a situation where the number of effect-size estimates (usually denoted by "*k*" in L2 meta-analyses) obtained across the categories of a moderator (e.g., types of treatment) exceeds the number of the studies associated with those categories. For example, in a meta-analysis where 20 studies have employed direct feedback and/or indirect feedback, we may be able to obtain 20 effect-size estimates for direct feedback and 14 effect-size estimates for indirect feedback (i.e., 34 effect-size estimates from 20 studies). We refer to this situation as the "*k* situation."

In the presence of a *k* situation, the effect-size estimates across the theoretically distinct categories of a moderator are not independent of each other.[3] For instance, effect-size estimates that make up the direct feedback category and those that constitute the meta-linguistic feedback category are not independent of each other. This is because some of these effect-size estimates have come from the same studies and carry along their sampling dependence. Placing these dependent effect-size estimates into seemingly independent categories (e.g., direct feedback vs. meta-linguistic explanation) and then comparing across them via traditional tests (e.g., $Q_{between}$ test) to answer a research question (see the next section) may lead to invalid conclusions depending on the extent of the *k* situation.

Classic methods of meta-analysis are ill-suited to address the *k* situation. As noted above, this inability is, once again, rooted in the lack of a framework to account for the dependence among the effect-size estimates classified into seemingly independent categories (see Plonsky et al., 2023).

Findings may be biased

Meta-analysts are often interested in examining the relation between the effect of a phenomenon of interest (e.g., the effect of feedback on written accuracy) and a range of

[3]Aside from dependence, moderators found in the *k* situation by design vary within and across studies. As a result, it is good practice to measure and distinguish between their within- and cross-study effects on (or association with) the average meta-analytic effect because the two could potentially differ in magnitude and direction. Frequently, their within-study effect (or association) is of immediate interest to researchers (for details see Fisher & Tipton, 2015).

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

other substantively moderating variables (Shadish et al., 2002). Cronbach's (1982) *UTOS* paradigm is what is often invoked as an *organizing framework* for this purpose (Becker, 1996; Cook, 1991; Matt & Cook, 2019). Under this framework, substantively moderating variables in the literature are divided into four dimensions, as applicable—namely, participants or more broadly units (*U*), treatments (*T*), outcomes (*O*), and settings (*S*). The meta-analyst then explores the association between instantiations of each dimension and the effect of a phenomenon of interest (Cook, 1993, 2013).

However, to validly estimate such substantive associations, an additional nuisance dimension needs to be accounted for as well. This additional nuisance dimension merely reflects the nonsubstantive and/or methodological (collectively denoted by *M*) differences across the studies selected for meta-analysis. As a result, the *M* dimension itself is not of theoretical interest or a target for meta-analysis. Rather, this dimension acts as a regulator to ensure obtaining unbiased substantive associations (Aloe et al., 2020; Becker, 2017; Becker & Aloe, 2008; Cheung & Slavin, 2016).

For example, L2 studies in a meta-analysis sample might likely differ in several substantive instantiations of *UTOS* dimensions (denoted by lower-case letters) such as in (1) the characteristics of participants (*u*'s, e.g., their proficiency level), (2) the types of treatment (*t*'s, e.g., types of corrective feedback), (3) the types of outcome (*o*'s, e.g., types of linguistic features targeted), and (4) the context of the study (*s*'s, e.g., language institute vs. university). However, the same L2 studies might also differ in ways that merely reflect their nonsubstantive and/or methodological differences. For instance, whereas some studies might have been able to randomly form their study groups (a true experiment), others might have been unable to do so purely by chance, for example, due to logistical constraints (i.e., a quasi-experiment). Or for example, some studies might have taken longer to complete and others have taken less time to complete (i.e., difference in study length). Such nonsubstantive and/or methodological differences need to be always accounted for in the background so the substantive association between *UTOS* dimensions and the effect of the phenomenon of interest can be validly assessed in the foreground (Berlin & Antman, 1992; Tipton et al., 2019).

Classic methods of meta-analysis are not readily capable of accommodating both the nonsubstantive and/or methodological confounds along with the substantive moderators. This incapability is due to the classic methods' *fragmented* approach toward moderators in a meta-analysis. That is, classic methods can commonly take one moderator at a time. This limitation is evident when, for example, we see tables presenting the results of numerous isolated moderator analyses in an L2 meta-analysis disregarding the connections between the substantive features of interest and the methodological choices that produced them. Consequently, classic methods of meta-analysis may not provide a valid picture regarding the association between the *UTOS* moderators (e.g., proficiency) and the effect of the phenomenon of interest (e.g., teacher feedback on written accuracy). In addition to biasing these associations, such a shortcoming in classic methods of meta-analysis can also bias the amount of heterogeneity in the effects explained by the *UTOS* moderators and, therefore, that which cannot be explained by the *UTOS* moderators.

An alternative approach to meta-analysis in L2 research

It should seem obvious that a starting point for an alternative approach to meta-analysis in L2 research, therefore, should account for the four main shortcomings of the classic

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

methods of meta-analysis discussed earlier. In the next section, the specifics of this alternative approach, *multivariate, multilevel meta-analysis*, will be discussed.

Multivariate, multilevel meta-analysis

The flexibility of regression analyses over the commonly used ANOVA techniques is well established in L2 research (Plonsky & Ghanbar, 2018; Plonsky & Oswald, 2017). Also well documented in L2 methodological literature is the fact that multilevel models (MLMs) add new capabilities to the regression analyses by accounting for the various types of dependence in the data (Cunnings, 2012; Gries, 2021). Multivariate, multilevel meta-analysis (3M) is a flexible approach to meta-analysis that builds on the capabilities of MLMs in the context of meta-analysis.

In its simplest formulation, 3M allows for examining the association between the average size of the effect (dependent variable) of an L2 phenomenon of interest (e.g., effect of teacher feedback on written accuracy) and a set of *(M)UTOS* moderators (independent variables) taking into account two important facts. First, studies could produce multiple dependent effect-size estimates prior to their meta-analysis (*sampling dependence*). Second, their corresponding underlying true effects are likely more similar in each study than those in other studies forming clusters across the studies (*hierarchical dependence*). The application of these two features in 3M is conceptually explained in the next two sections.

Sampling dependence in 3M

Despite the available formulas to obtain the exact structure of the sampling dependence among effect-size estimates due to certain design aspects (see Olkin & Gleser, 2009), in practice, determining the exact structure of the sampling dependence in L2 studies is often infeasible for at least two reasons. First, for design aspects B (repeated measurements) and C (measuring multiple outcomes), accurate estimation of sampling dependence requires estimates of the correlation coefficient for the participants' measurements across multiple occasions and outcomes. Yet these required correlation estimates are almost always absent in the primary L2 studies (Plonsky & Oswald, 2017). Second, the three design aspects discussed in the previous sections could occur together in any combination across the studies. This makes the construction of the sampling dependence structure practically difficult and potentially prone to error. Acknowledging these limitations, 3M initially assumes a medium-high correlation (e.g., $r = .60$), as a middle-of-the-way value, among effect-size estimates across all studies and later subjects such an assumption to a sensitivity analysis (Pustejovsky & Tipton, 2022).

Hierarchical dependence in 3M

In theory, 3M is capable of envisioning many potential levels of hierarchical dependence for the underlying true effects of our effect-size estimates. For instance, one assumption to envision might be that the true effects of an L2 phenomenon of interest (e.g., effect of corrective feedback on written accuracy) are more similar in each study than those in other studies (study as a level). But when, for example, treatment groups in each study are tested on multiple occasions, as in Shintani and Ellis (2013) discussed earlier, one might also reason that the true effects for each treatment group in a study are likely more similar than those for other treatment groups in that study (treatment group as a level).

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

In practice, however, (a) limitations in the number of studies and their effect-size estimates and/or (b) having possibly several *(M)UTOS* moderators at various hierarchical levels (e.g., study level and treatment level) of the literature can curtail one's ability to encode the many potential levels of hierarchical dependence into a 3M model (Hedges, 2019). For most purposes in L2 research, a general form of hierarchical dependence might simply acknowledge that the underlying true effects of our effect-size estimates are likely more similar in each study as a whole than those in other studies. The product of such an acknowledgement is that we can obtain the random variation in studies' average true effects (study-level variation) and the random variation in any given study's own true effects (effect-level variation). We say "any given study" because the effect-level variation is somewhat averaged across all studies, thus representing such variation in any given study. Figure 2 symbolically depicts these two sources of random variation with the first study at the top figuratively used as a "given study."

Importantly, these two sources of variation are called *random* because they account for the variation in the underlying true effects of an L2 phenomenon of interest that could not be systematically attributed to any of the *(M)UTOS* moderators used in a 3M model (e.g., variation in magnitude of true effects beyond that attributable to proficiency level differences). This formulation reinforces the belief that despite potentially systematic reasons (i.e., *[M]UTOS* moderators) for variation in the magnitude of true

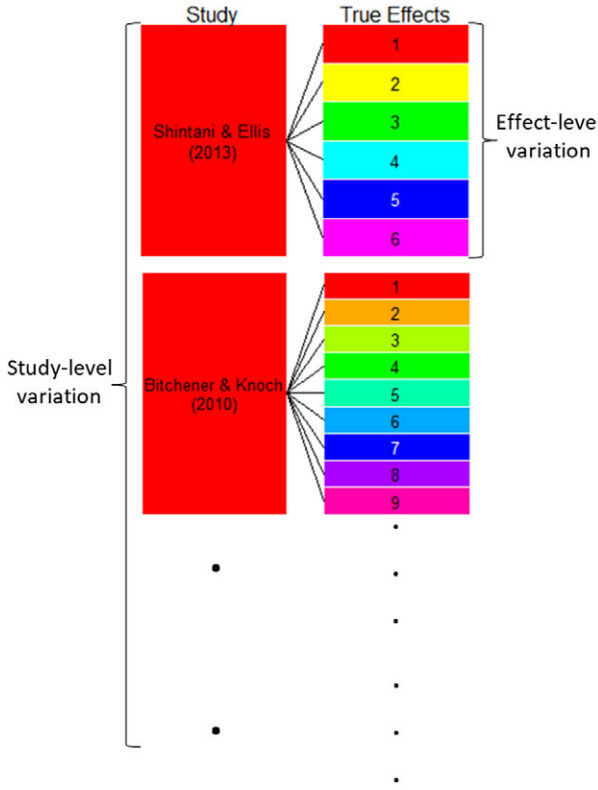

Figure 2. Two sources of random variation captured by 3M.

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

effects for an L2 phenomenon of interest, in the end studies as a whole and their constituent effects could each potentially have their own unique but random differences captured by these two sources of random variation.

For purposes of describing such a model, it is customary to refer to the studies' average true effects as Level 1 (highest level), the underlying true effects of the studies' effect-size estimates as Level 2 (middle level), and the studies' own effect-size estimates as obtained from the literature as Level 3 (lowest level). Thus, this would be an example of a three-level, 3M model.

Applying 3M to a complex literature

To promote the use of 3M in L2 research, we apply it to one of the most polemic areas in instructed second language acquisition (ISLA) research: the effectiveness of written corrective feedback (WCF) in improving L2 writers' accuracy. Because of the nature of the so-called WCF debate, a large and often complex body of studies has contributed to the WCF literature. For substantive purposes, Brown, Liu, & Norouzian (2023) extensively reviewed the details of and meta-analyzed a large set of WCF studies. For the methodological purposes of this article, we randomly selected 35 studies from the full set of the WCF studies in Brown et al.

Eligibility criteria

As in Brown et al. (2023), the WCF studies in the selected sample have all (1) collected data through writing tasks, (2) investigated WCF administered only by teachers and/or researchers, (3) focused on linguistic errors (lexical, morphological, and/or syntactic), (4) measured the effectiveness of WCF through accuracy in participants' new pieces of writing, (5) employed objective measures (e.g., target language use, normed error frequency, ratio of error-free units) to measure development in linguistic accuracy, (6) employed a (quasi-)experimental research design, (7) reported descriptive statistics (e.g., group means and standard deviations or related test statistics such as *t* values) for the calculation of effect sizes, and (8) were written in English.

A brief review of the design complexities of the selected WCF studies will prove crucial in understanding the meta-analytic challenges in this literature. Below, we explore some of the design complexities of the selected sample of WCF studies in more detail.

Design complexities of WCF studies

The WCF literature includes studies with a variety of design features. From those features, four in particular are worth mentioning here. First, WCF studies usually employ multiple treatment groups (*group*) receiving different types of WCF all compared against a control/comparison group. Second, the studies often involve multiple measurements of the same participants (*time*) to assess the durability of WCF's effect on learners' written accuracy. Third, they can also measure multiple outcomes to study WCF's effect on multiple types of linguistic errors (*err_type*). Finally, given the debates over the treatment of the control group's members (e.g., whether they should only engage in writing practice or receive feedback on content) and its subsequent influence on the results of WCF studies (Truscott, 1996, 2004), occasionally studies (Van Beuningen et al., 2012) tend to use more than one type of control group (*cont_type*) against which all WCF-receiving groups are compared. Figure 3 examines,

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

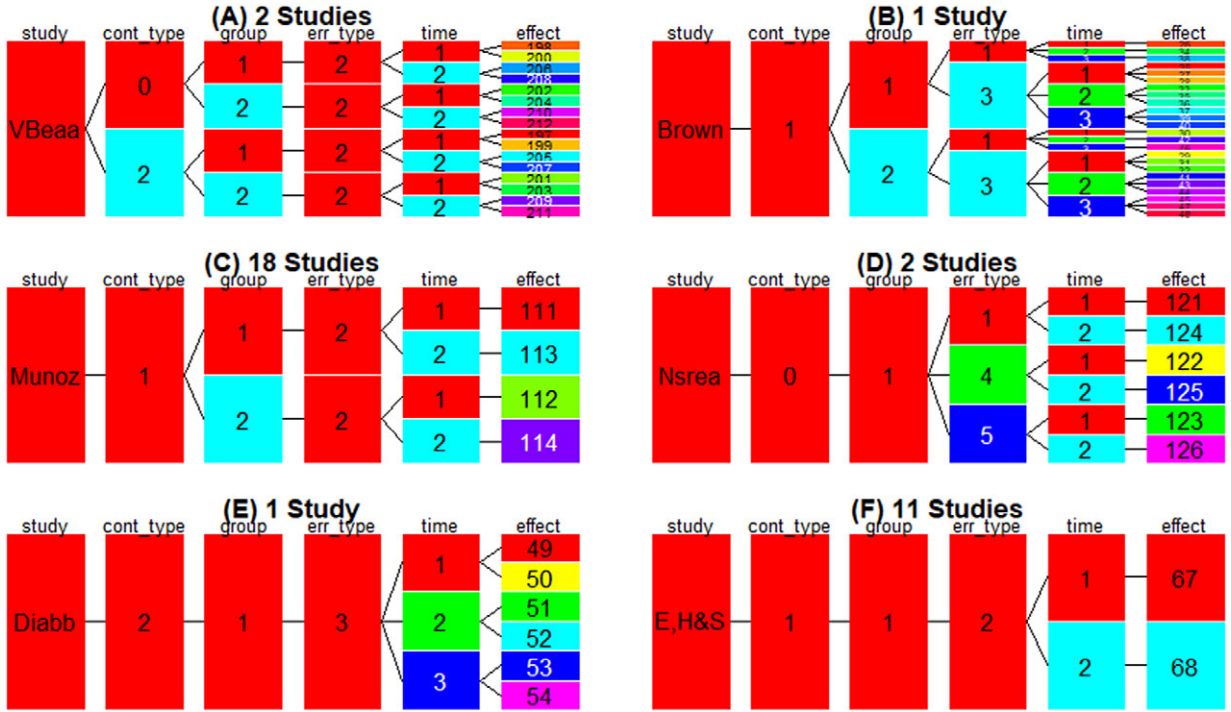

Figure 3. Design complexity in WCF literature.

classifies, and visually displays these design features for our selected WCF studies (to reproduce Figure 3, see: https://github.com/rnorouzian/i/blob/master/1.r).

In Figure 3, each panel represents a unique combination of the design features found in the WCF literature. Changes in color denote multiplicity in the design feature titled above each column (e.g., *group*, *time*). For instance, we see that all the studies in the WCF literature have measured their participants on more than one occasion (pre- and posttests). This is because across all the panels, *time* always changes in color. In contrast, in panel F, there are 11 studies that have used only one control group, one treatment group, and one error type. The numbers in each box denote the identifiers (e.g., 0 for *cont_type* = the control group only engaging in writing practice) or indices (1 for *group* = the first treatment group in the study) used by the meta-analyst to code these design features in each study in the coding sheet.

Inspecting Figure 3 a bit further, we also realize that the use of multiple control groups (*cont_type*) is very infrequent in our WCF study pool, occurring in only two studies (Panel A). Conversely, having multiple treatment groups is a quite common design feature in the WCF studies appearing in twenty studies (Panels A, B, and C). Additionally, studying WCF's effect on multiple error types (outcomes) occurs in a handful of studies (Panels B and D). Other design features can also be examined in a similar manner. However, notice that all studies have invariably produced multiple effect-size estimates, as indicated by the multiple row indices (*effect*; row-wise location of each effect-size estimate in the data) in each panel.

In short, these design complexities may give rise to (a) having several effect-size estimates from each WCF study, (b) the fact that these estimates are likely dependent on one another in a variety of ways in each study, and potentially, and (c) the existence of both systematic and random variation (i.e., heterogeneity) in the true effects of WCF on written accuracy both at the study and the effect levels. From these three consequences, the last one differs in nature from the other ones in that it involves using a range of *(M)UTOS* moderators backed by theory and/or prior research. Specifically, to explore the degree to which the heterogeneity in the magnitude of the true effects of WCF might be due to the features systematically distinguishing the studies as a whole and/or finer elements within the studies (e.g., participants) from each other, a number of candidate *(M)UTOS* moderators may be considered. A selection of these moderators is provided in the next section.

The (M)UTOS moderators

In theory, there may be a large number of *(M)UTOS* moderators that can potentially account for the systematically explainable heterogeneity in the magnitude of the true effects of WCF. In practice, however, a smaller set may be found and consistently coded in the WCF literature along each dimension of *(M)UTOS*. For the purposes of this study, Table 1 provides 10 *(M)UTOS* moderators, two instantiations of each *(M)UTOS* dimension. Table 1 also includes the general as well as the coding definitions of these moderators (for a fuller list of potential moderators see Brown et al., 2023).

Although space limitation prevents us from applying 3M to all the moderators listed in Table 1, we try to offer a road map by applying the 3M to a substantive moderator (i.e., proficiency) and examine the critical decision points encountered along the way. A description of the data set used for this purpose is presented in the next section.

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

Table 1. Coding scheme for the selected moderators

Abbreviation	(M)UTOS category	General definition	Coding definition
prof	*u*	Participants' proficiency level	Beginner; Intermediate; Advanced
age	*u*	Participants' age group	Teen (12–17); Adult (18–35)
wcf_type	*t*	Type of feedback provided	Direct (correction given); Indirect (error's location indicated); Coded (coding symbols used next to errors); Meta-linguistic (indirect feedback + grammar notes); Meta-linguistic+direct
wcf_scope	*t*	Range of linguistic structures feedback targeted	Unfocused (> 5 structures); Mid-focused (2-5 structures); Highly focused (1 structure)
err_type	*o*	Type of linguistic structures receiving feedback	Mixed (multiple types measured together); Articles; Prepositions; Verb tense; Other types
time	*o*	Outcome measurement timing	Baseline; Posttest 1; Posttest 2
lang_setting	*s*	Language learning context of study	Foreign language setting; Second language setting
res_setting	*s*	Research setting of study	High school; College; Language institute
study_length	*m*	Length of study in weeks from first to last measure	Continuous
des_type	*m*	Type of study design	True experiment; Quasi-experiment

Table 2. Distribution summary of effect-size estimates across studies

Effects	Min effects in a study	Max effects in a study	Average effects in a study
245	2	24	7

The wcf data set

The data set used in this article is called `wcf`. In addition to the *(M)UTOS* moderators described in Table 1, the `wcf` data set includes 245 effect-size estimates in the standardized mean difference (SMD) unit (Hedges's *g*), which is a common metric used in applied linguistics research. As shown in Table 2, all studies contributed more than one effect-size estimate to the database such that, on average, each WCF study provided seven effect-size estimates.

These effect-size estimates are all corrected for a systematic bias inherent in them, especially when obtained from small-sized group comparisons. As is customary, we refer to these bias-corrected effect-size estimates as Hedges's *g* and denote them by `g` in the `wcf` data set.

Accompanying these effect-size estimates are their sampling variances, denoted by `var` in the `wcf` data set. Sampling variances are simply the squared version of the estimates' standard errors (Viechtbauer, 2007). They denote how dispersed, in the variance metric, the distribution of effect-size estimates might be, if the exact same group comparisons used to obtain the effect-size estimates were to be replicated

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

infinitely many times. Although these sampling variances indicate how much each effect-size estimate in a study might *vary* in its replication sampling, they convey no information regarding how different estimates in that study might *covary* with one another likely due to the design aspects discussed earlier (see *Sampling dependence inherent in L2 studies is ignored*).

The `wcf` data set also uses the name of the study authors as the identifier for each study. Thus, variable `study` provides a unique name for each study. Needing to be unique to each study, if the same authors contributed more than one study to the study pool, their names for each study were distinguished from one another by appropriate suffixes (e.g., Author A vs. Author B).

Similar to studies, each effect-size estimate is associated with a numeric identifier. This variable, which is called `effect` in the `wcf` data set, simply denotes the row-wise location of the effect-size estimates. For example, the first effect-size estimate in the first row gets a value of 1, the second one gets a value of 2, and so on. Therefore, such numbers have no real numeric value. We will need this variable to be able to distinguish the effect-level variation from the study-level variation in the true effects (see preceding text).

Together with all the *(M)UTOS* moderators shown in Table 1, the `wcf` data set is built into the R programs introduced in the next section. Thus, readers will automatically have access to the data set, once they install the programs in the next section.

Fitting 3M models with software

One common goal[4] in meta-analysis is to describe the relation of each *UTOS* moderator to the magnitude of the average effect of a phenomenon of interest (e.g., WCF on written accuracy), one relation at a time controlling for the *M* moderators. Consistent with this descriptive goal, research questions (RQs) may then take a general form, leaving the door open for each moderator to be tested in a separate 3M model. For instance, a general RQ might be to what extent do the characteristics of learners or, more broadly, units (*U*), treatments (*T*), outcomes (*O*), and settings (*S*) influence the effect of *X* (e.g., WCF) on *Y* (e.g., written accuracy)?

To fit 3M models for research questions similar to what was mentioned above, we suggest using the following suite of R programs accessible by running the following in R or RStudio:

```
(1) source("https://raw.githubusercontent.com/rnorouzian/i/master/3m.r")
```

These functions use the metafor package (Viechtbauer, 2010) and the clubSandwich package (Pustejovsky, 2022) in R (R Core Team, 2022). As noted above, to achieve our descriptive goal, moderators listed in Table 1 shall be separately fit using a 3M model. In doing so, it is highly advisable (see *Findings may be biased*) to always include the *M* moderators in each model to ensure a higher degree of conclusion validity in the results.

[4]We say one common goal because there could be other goals as well. For example, it is possible for L2 meta-analysts to attempt to find the smallest set of *(M)UTOS* moderators that most strongly associate with the magnitude of the average effect of a phenomenon of interest. This is helpful then in informing teachers and other stakeholders of the main *UTOS* factors that are most important in moderating the effectiveness of the phenomenon of interest (teacher feedback on writing accuracy). The details of this approach fall outside the scope of this introductory article.

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

A 3M model may be fit in two steps. First, recall that sampling dependence was inherent in the effect-size estimates obtained from the WCF literature due to the studies' own designs prior to their meta-analysis. As such, prior to meta-analysis, we need to introduce this dependence among the effect-size estimates using the following in R:

(2)
```
V = with(wcf, impute_covariance_matrix(var, study,
r=.6))
```

To better understand the above R function, recall that we already have the sampling variances (`var`) for our effect-size estimates in our `wcf` data set, as we always should for any meta-analysis. However, we need to build the sampling covariances around them to account for the inherent sampling dependence among the effect-size estimates in each study. The R function `impute_covariance_matrix()` helps adding these sampling dependencies (i.e., covariances) around those sampling variances. To do so, we need to supply the function with the sampling variances (`var`), the level in the data where the dependencies among effect-size estimates occur (in our case `study`, as effect-size estimates in each *study* are likely correlated), as well as a value or a set of values for the correlations among those effect-size estimates in each study (`r`). As indicated earlier, the idea is to use a medium-high correlation (r = .60), as an initial guesstimate to represent a constant sampling dependence among effect-size estimates (Pustejovsky & Tipton, 2022). These elements (i.e., `var`, `study`, and `r`) are what we input to `impute_covariance_matrix()` in R code (2). For more details regarding `impute_covariance_matrix()`, readers can refer to the function's help page by running `?impute_covariance_matrix()` in R.

The second step involves setting up the 3M model in R. The model consists of three main parts: (1) a place for defining the *(M)UTOS* moderators of interest and the effect-size estimates (argument `yi`); (2) a place for defining the structure of sampling variances and, if any, sampling dependencies (argument `V`); and (3) a place for defining the structure of hierarchical dependence (argument `random`).

To better understand this second step, suppose that we want to examine the extent to which proficiency influences the effect of WCF on written accuracy over time (i.e., from pre to posttest) in the WCF literature. Part one of the model using Table 1 *(M)UTOS* moderator abbreviations may be defined in R as follows:

(3)
```
yi = g ~ prof* time + study_length + des_type
```

This formulation includes the effect size (`g`), the focal *UTOS* moderator (`prof`), the outcome measurement timing (`time`) interacted (`*`) with the focal moderator to understand the moderator's influence on effect size over time, and our two methodological (*M*) moderators `study_length` and `des_type`, which will need to be always controlled for as nuisance variables (see *Findings may be biased*). Notice that methodological moderators are always added using + signs to properly function their controlling role.

Part two is to simply supply the sampling dependencies (`V`) defined earlier in R code (2) above. Finally, part three is to provide the structure of the hierarchical dependence. With respect to hierarchical dependence, we recognize that true effects are clustered (i.e., nested) within each study leading to the estimation of two sources of random variation (see *Hierarchical dependence in 3M*). Part three for our 3M model may be defined in R as follows:

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

(4) `random = ~1|study/effect`

In addition to these steps, it is highly recommended that hypothesis tests from the model use the potentially more conservative (i.e., smaller) degrees of freedom (`dfs`) offered by the "contain" method[5] (see Viechtbauer, 2010) in the R package metafor:

(5) `dfs = "contain"`

Tying these steps together, the 3M explores the extent to which proficiency influences the effect of WCF on written accuracy over time, which may be defined in R as follows:

(6)
```
m_prof = rma.mv(yi=yi, V=V, random=random, dfs=dfs,
data=wcf)
```

Detecting outliers in 3M models

Before finalizing our 3M model, a critically important step is to ensure that the model is not negatively affected by outlying estimates of effect size. A useful approach to do so is to examine how each effect-size estimate interactively influences a 3M model (Viechtbauer, 2021). Interactive outliers are those that from multiple perspectives are considered outlying. Three such perspectives include an effect estimate (1) having a standardized deleted residual that exceeds ± 1.96 (such estimates do not fit the 3M model well), (2) exerting a large influence on the results once deleted (an estimate's Cook's distance exceeding the upper limit of a box plot of all estimates' Cook's distances), and (3) leveraging considerable influence on its own predicted true effect value given the meta-analytic weight and the moderator values attached to it (an estimate's Hat value exceeding the upper limit of a box plot of all estimates' Hat values).

The idea behind detecting interactive outliers is to find effect-size estimates that *simultaneously* meet the three perspectives discussed above. For our `m_prof` model, we can achieve this goal by running the following in R with the results shown in Figure 4:

(7) `out = interactive_outlier(m_prof)`

If detected, interactive outliers are indicated by their row-wise location in the Figure generated by the R function `interactive_outlier`. In our case, out of 245 estimates of effect size obtained from the WCF literature, only one effect-size estimate (not an entire study) on row 116 is identified as an interactive outlier. The `interactive_outlier` function automatically removes the interactive outliers and provides the new data set:

(8) `wcf_new = out$new_data`

Because this new data set is smaller than the original one due to the removal of one interactively outlying effect-size estimate, we also need to redefine our sampling dependence structure initially defined in R code (2) with this new data set:

[5]For details, readers are referred to the "Tests and Confidence Intervals" section of the metafor package documentation by running the following in R or Rstudio: `? rma.mv`.

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

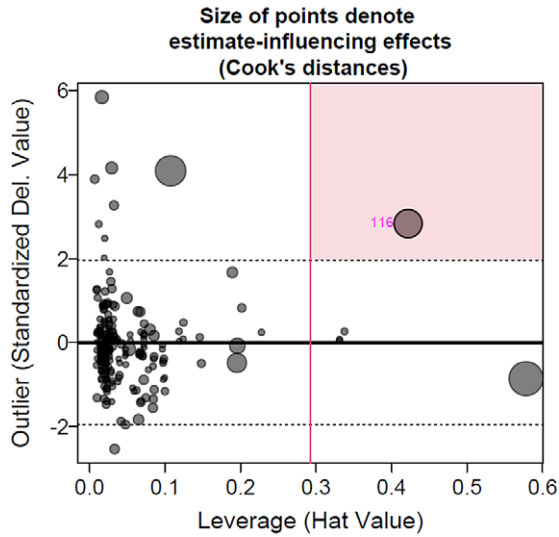

Figure 4. Detection of interactive outliers.

```
(9) V_new = with(wcf_new, impute_covariance_matrix(var,
    study, r=.6))
```

Finally, we can update our original 3M model (`m_prof`) with this new data set (`wcf_new`) and the new sampling dependence structure (`V_new`):

```
(10) m_prof_new = update(m_prof, V = V_new, data = wcf_new)
```

Exploring results from a 3M model

At its core, a 3M model is a multilevel regression model (see *multivariate, multilevel meta-analysis*). This means that the output will largely resemble that from multilevel regression analyses. However, this output is often perceived to be difficult to interpret for researchers in applied linguistics (Plonsky & Oswald, 2017) as well as those in other fields of study (Fox & Weisberg, 2018). The interpretations become particularly subtle as moderators in the model, including the categorical and the continuous ones, grow in number.

Additionally, recall that in addition to the moderators involved in our research question (i.e., how does *proficiency* influence the effect of WCF on written accuracy over *time*?), other variables (i.e., *M* moderators) in the 3M simply act as bias-controlling agents. Such bias-controlling agents, as noted previously, do not necessarily merit an interpretation and may be adjusted for in the background (Berlin & Antman, 1992; Tipton et al., 2019).

The R function `post_rma()` helps us obtain a piece of output that is focused on our research question, keeping the *M* moderators adjusted for in the background. For our research question, the function may be used as follows:

```
(11)    post_prof = post_rma(m_prof_new, ~prof*time)
```

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

Table 3. 3M results for the influence of proficiency on WCF effect over time

	Proficiency	Time	Mean	*SE*	Lower	Upper	*t*	*p*	Sig.
1	Advanced	Baseline	−0.500	0.307	-1.136	0.137	−1.628	.118	
2	Beginner	Baseline	0.575	0.485	-0.432	1.581	1.184	.249	
3	Intermediate	Baseline	0.112	0.138	-0.175	0.399	0.812	.426	
4	Advanced	Post1	0.156	0.307	-0.481	0.793	0.509	.616	
5	Beginner	Post1	0.685	0.492	-0.336	1.705	1.391	.178	
6	Intermediate	Post1	0.660	0.140	0.370	0.949	4.727	.000	***
7	Advanced	Post2	−0.144	0.445	-1.066	0.778	−0.324	.749	
8	Beginner	Post2	1.071	0.671	-0.319	2.462	1.598	.124	
9	Intermediate	Post2	0.654	0.147	0.350	0.959	4.456	.000	***

In (11), `m_prof_new` represents our fitted 3M model in R code (9). The `~prof* – time` part indicates the moderators involved in our research question. These primary moderators are also the ones defined in our 3M model in R code (3). The output from `post_rma` is displayed in Table 3.

The column titled *Mean* represents the meta-analytic average effects found for each proficiency level at each measurement time in the WCF literature, *SE* represents the standard errors for those average effects, and *Lower* and *Upper* represent the 95% confidence intervals for the estimates of *Mean.* The column *t* is the *t* value resulting from testing the null hypothesis that each *Mean* is in reality nonexistent (here 0). Thus, the *p* values indicate whether we can reject such null hypotheses using some decision rule for statistical significance (e.g., $p < .05$).

Similar to a primary study, it would be helpful to visualize the patterns represented in Table 3 for our meta-analytic study. The R function `plot_rma()` is designed for this purpose. We can direct the function to plot the proficiency groups' Mean effects over time (`prof~time`):

(12)
```
plot_rma(m_prof_new, prof~time, xlab = "Time")
```

Such visuals have other benefits as well. For instance, looking at Figure 5, we realize that in the WCF literature, low-proficiency learners are likely quite understudied relative to intermediate and advanced learners. This is evident in that the confidence intervals for the beginners are much wider than are those for the other proficiency levels. Additionally, not only have WCF studies largely focused on intermediate learners, this proficiency level has been more longitudinally studied than others have. This is seen in the confidence intervals for the intermediate learners, which are much narrower on the second posttest than those for the other proficiency levels.

Despite these benefits, it is often difficult to compare the groups in any precise manner by solely looking at visualizations such as Figure 5. This is why it is necessary to conduct post-3M comparisons to better explore the findings of the study.

Post-3M comparisons

Although Table 3 and Figure 5 provide useful information about how WCF has worked for each proficiency level at a given measurement time, knowing precisely how WCF's effect compares within and across the proficiency levels requires making post-3M comparisons. Some of these comparisons may be simple (e.g., Prior to receiving WCF, how do the proficiency groups differ at the pretesting occasion?). Others may be more

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

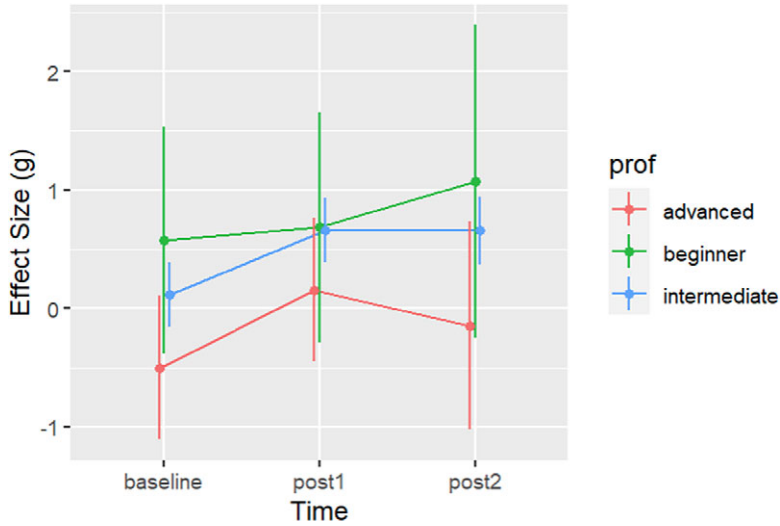

Figure 5. Average meta-analytic effects for proficiency levels over time.

Table 4. Difference between proficiency groups at the pretest

Contrast	Estimate	*SE*	Lower	Upper	*t*	*p*	Sig.
Advanced baseline − beginner baseline	−1.074	0.597	−2.313	0.164	−1.799	.086	.
Advanced baseline − intermediate baseline	−0.612	0.337	−1.311	0.087	−1.817	.083	.
Beginner baseline − intermediate baseline	0.462	0.507	−0.589	1.514	0.912	.372	

complex (e.g., How does WCF's effect for proficiency groups differ from pre- to the final posttest?).

For simple comparisons such as comparing groups at the pretesting time, a pairwise comparison involving the first three *Means* (`include=1:3`) in Table 3 results, which focus on the proficiency groups' average effects at the pretesting occasion, is all that is needed. To do so, we may use the pairwise functionality of `post_rma`:

(13)
```
post_rma(m_prof_new, pairwise~prof*time, include=
1:3)
```

As can be seen in Table 4, no statistically significant differences are found across the proficiency groups at the pretesting occasion (i.e., prior to receiving WCF). This finding may, however, be taken with a grain of salt. This is so because the CIs for the differences are fairly wide, indicating a potential inability to detect such differences, when some could in reality exist (i.e., Type II error in decision making).

With this Type II error possibility, comparing the proficiency groups at any posttesting occasion (e.g., final posttest) without considering their pretesting status may provide an invalid picture regarding how WCF has worked for different proficiency groups at that point. Consistent with this view, we can instead compare the gains

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

each proficiency group has made from the pretest to any posttesting occasion of interest. This is precisely what our second comparison (i.e., How does WCF's effect for proficiency groups differ from pre- to the final posttest?) is targeting. To address our second comparison, we can define the long-term gain for each proficiency group between pre- and second posttest as follows:

(14) Gain = second posttest − pretest

Our guiding plan then is to compare each pair of proficiency groups on such a gain, which will require testing the following contrasts:

(15) Gain (beginner) − Gain (advanced)

(16) Gain (beginner) − Gain (intermediate)

(17) Gain (intermediate) − Gain (advanced)

Such customized contrasts, however, may not be readily tested by statistical software. Instead, we will need to locate the *Means* from Table 3 to form the gains for each proficiency group and then contrast them with one another. For example, for contrast (15) we can easily see that the difference in the gains entails the following mean locations from Table 3:

(18) (8 − 2) − (7 − 1)

The above, in words, reads

(19) (2nd *Mean* subtracted from 8th *Mean*) − (1st *Mean* subtracted from 7th *Mean*)

Contrast (18) can be simplified by applying the − sign between its parentheses:

(20) (8 − 2 − 7 + 1)

Similarly, for contrasts (16) and (17), we will, respectively, obtain the following *Mean* locations from Table 3:

(21) (8 − 2 − 9 + 3)

(22) (9 − 3 −7 +1)

Having determined the location of the *Means* in Table 3 for contrasts (15) through (17), we can test them using `contrast_rma`. To do so, we will supply the function with the Table 3 results (`post_prof`) and the *Mean* locations indicated in (20) through (22) for our customized contrasts:

(23)
```
contrast_rma(post_prof, list(
    "Gain(beginner − advanced)"= c(8,-2,-7,1),
    "Gain(beginner − intermediate)"= c(8,-2,-9,3),
    "Gain(intermediate − advanced)"= c(9,-3,-7,1)))
```

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

Table 5. Difference in long-term gains across proficiency levels

Contrast	Estimate	*SE*	Lower	Upper	*t*	*p*
Gain(beginner − advanced)	0.141	0.656	−1.220	1.502	0.215	.832
Gain(beginner − intermediate)	−0.045	0.549	−1.183	1.092	−0.082	.935
Gain(intermediate − advanced)	0.186	0.384	−0.610	0.983	0.486	.632

The results are shown in Table 5. Based on these results, long-term gains for the proficiency groups in the WCF literature are not statistically different from one another. Once again, however, the fairly wide CIs for these differences prevent us from placing full trust in these results. We, as indicated previously, know that the bulk of WCF research has been focused on intermediate learners, with the other two proficiency groups remaining largely understudied. And this is the underlying reason we tend to see a fair amount of uncertainty in our results.

Note, however, that our customized contrasts focused on the *difference* in gains (not gains themselves) from WCF across the proficiency groups to determine whether WCF might have differentially benefited the proficiency groups. Thus, the lack of a statistically significant finding should not mean that WCF has not produced any benefit for each proficiency group per se. Indeed, we can easily obtain the long-term gains from WCF per equation (14) for each proficiency group using `contrast_rma`. To do so, we can again supply `contrast_rma` with the relevant *Mean* locations for each proficiency group's long-term gain from Table 3 (e.g., 8 − 2 for beginners' gain):

(24)
```
gain_prof = contrast_rma(post_prof, list(
  "Gain(beginner)" = c(8,-2),
  "Gain(intermediate)" = c(9,-3),
  "Gain(advanced)" = c(7,-1)))
```

The results in Table 6 show that the long-term gain in written accuracy from WCF is statistically significant for intermediate learners and so with a reasonable level of certainty, 95% CI [.356, .729]. For other understudied proficiency groups, however, the high level of uncertainty, indicated by wide confidence intervals, prevents a precise estimate of the long-term gains in written accuracy.

Exploring variation (heterogeneity) in 3M

Besides exploring WCF's mean effects for various proficiency levels, our 3M model allows us to examine how much of the total variation (i.e., sum of σ study- and σ effect-level variations) in the underlying WCF effects is systematically due to proficiency adjusting for the *M* moderators. Stated differently, we can ask the question: to what extent the variation in the magnitude of underlying WCF effects is because they

Table 6. Long-term gains in writing accuracy across proficiency levels

Contrast	Estimate	SE	Lower	Upper	*t*	*p*	Sig.
Gain(beginner)	0.497	0.540	-0.624	1.617	0.919	.368	
Gain(intermediate)	0.542	0.095	0.346	0.738	5.730	.000	***
Gain(advanced)	0.356	0.372	-0.416	1.127	0.956	.350	

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

Table 7. Variation in underlying WCF effects accounted for by the fitted 3M

Model	σ total	σ study	σ effect	*p*	R^2
No (M)UTOS	0.738	0.529	0.514		
m_prof_new	0.699	0.548	0.434	.000	5.200%

come from different proficiency levels controlling for the methodological differences in the WCF studies? To answer this, it is customary to compare our existing model with one that has no *(M)UTOS* moderator in it. This procedure is akin to obtaining the R^2 in regression analyses. To achieve this, we can use `R2_rma()`:

(25)
```
R2_rma(m_prof_new)
```

Table 7 displays the results. As can be seen, our model (`m_prof_new`) has accounted for 5.2% of the total variation (in *SD* unit; σ total) in underlying WCF effects given its moderators. These moderators also tend to collectively be statistically significant ($p < .05$). Clearly, the rest of the variation in the underlying WCF effects remains unexplained by proficiency, perhaps a reason for the meta-analyst to explore other moderators in subsequent analyses.

Using variation (heterogeneity) for evidence-based recommendations

Understanding the magnitude and sources of variation in the underlying WCF effects also provides a window into the distribution of effects in that domain of research (Mathur & VanderWeele, 2019). Supplementing this information with the *Mean* effects (Table 3) or the effects from our main comparisons (e.g., gains in Table 6) provides a highly useful means of offering evidence-based recommendations to the audience of our meta-analysis. For example, we could predict how likely it would be, on average, for L2 learners to positively benefit (i.e., gain) from teachers' written feedback given their proficiency levels. To do so, we can feed Table 6 results (`gain_prof`) for proficiency groups' long-term gains (`gain=TRUE`) into function `prob_rma()` with the results shown in Table 8:

(26)
```
prob_rma(gain_prof, gain=TRUE)
```

Based on Table 8, intermediate learners are, on average, ~81% (column *Probability*) likely to positively benefit from teachers' written feedback in the long run. But a positive gain could include any amount of gain even as small as to be considered practically insignificant. Thus, we can define a minimum threshold for that gain to be practically more meaningful. For instance, we could predict how likely it would be for learners, on average, to gain at least as large as 0.2 on the effect-size scale from teachers' written

Table 8. Prediction for WCF benefitting learners given their proficiency

Term	Target effect	Probability	Min	Max
Gain(beginner)	0 or larger	79.08%	18.73%	99.87%
Gain(intermediate)	0 or larger	81.13%	68.87%	91.52%
Gain(advanced)	0 or larger	71.89%	27.70%	98.20%

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

Table 9. Prediction for WCF benefitting learners by a target effect given their proficiency

Term	Target effect	Probability	Min	Max
Gain(beginner)	0.2 or larger	68.57%	12.05%	99.58%
Gain(intermediate)	0.2 or larger	71.12%	58.23%	84.17%
Gain(advanced)	0.2 or larger	60.03%	19.04%	95.78%

feedback given their proficiency levels. To do so, we can provide `prob_rma` with our target effect of .20 with the results shown in Table 9:

(27) `prob_rma(gain_prof, target_effect=0.2, gain=TRUE)`

Now, we can see that the likelihood of benefitting from teachers' written feedback for intermediate learners, on average, decreases from ~81% to ~71% in the long run (i.e., from pre- to the second posttest controlling for the differences in the studies' lengths). This decrease in likelihood is, of course, expected given that we have set a higher expectation for learning gains from teachers' written feedback. Notice also that in both Tables 8 and 9, our ability to make such easy-to-understand evidence-based recommendations depends, among other things, on the amount of the evidence supporting these benefits. For instance, because beginners and advanced learners are understudied in the WCF literature, the minimum (column *Min*) and maximum (column *Max*) bounds on how much they, on average, can benefit from teachers' written feedback in the long-run are quite far apart. As a result, evidence-based recommendations in this respect may be made with caution (e.g., perhaps using or emphasizing the *Min* probabilities).

Sensitivity of 3M to sampling dependence

Recall from our discussion on sampling dependence (see *Sampling dependence in 3M*) that the medium-high correlation (r = .60) assumed for effect-size estimates in each study was more of a guesstimate. As such, we may want to subject this value to a sensitivity analysis. This means that we would want to assume a range of different correlations for effect-size estimates in each study to see how robust the result of the 3M model (`m_prof_new`) will be to the change in such a guesstimate.

To achieve this goal, we can use the `sense_rma()` function in R, which uses our initial Table 3 results (`post_prof`) as well as the name of the variable in our data denoting the sampling variances (`var`):

(28) `sense_rma(post_prof, var_name = "var")`

Table 10 shows the results of our sensitivity analysis under five different degrees of correlation assumed for effect-size estimates in each study. The last column assesses the variability (in *SD* unit) in the results under all these degrees of correlation. Although there is no preset cut point, almost all mean effects (except for `beginner post2`) are close to one another, resulting in their variability (column *SD*) being fairly small.

For the total variation (in *SD* unit) in true WCF effects (last row), we do see a bit of sensitivity. Specifically, as r becomes larger, the total variation decreases in magnitude at a faster rate than that in other terms in Table 10. Ultimately, the point is to acknowledge the fact that our addressing the dependencies in individual studies, at

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

Table 10. Sensitivity of 3M results to different amounts of sampling dependence

Term	$r = 0.3$	$r = 0.4$	$r = 0.5$	$r = 0.6^{\star}$	$r = 0.7$	*SD*
Advanced baseline	−0.496	−0.497	−0.498	−0.500	−0.501	0.002
Beginner baseline	0.594	0.587	0.581	0.575	0.568	0.010
Intermediate baseline	0.145	0.134	0.123	0.112	0.101	0.017
Advanced Post1	0.142	0.147	0.152	0.156	0.161	0.007
Beginner Post1	0.720	0.708	0.696	0.685	0.673	0.019
Intermediate Post1	0.689	0.679	0.669	0.660	0.650	0.015
Advanced Post2	−0.149	−0.147	−0.146	−0.144	−0.142	0.003
Beginner Post2	1.170	1.137	1.104	1.071	1.039	0.052
Intermediate Post2	0.684	0.675	0.665	0.654	0.644	0.016
Total variation in *SD*	0.743	0.728	0.714	0.699	0.686	0.023

Note. The asterisk denotes the *r* value used in the current model.

least in some ways, is not and perhaps will never be perfect. Rather, it to some extent is sensitive to the degree of correlation assumed for the evidence collected from our complex studies. However, these small sensitivities, in the end, often outweigh a complete ignorance of the dependent nature of the evidence collected from such studies.

The steps illustrated above are important in describing the role of proficiency as a critical variable in WCF effectiveness. But they also provide a road map for exploring all other moderators listed in Table 1. Also, due to space limitation other moderators are not demonstrated here, so the interested readers are encouraged to implement the steps described in the present article for other moderators given the open-data-and-materials nature of the present article.

Conclusion

The alternative view of meta-analysis introduced in this article feels and looks a lot like that of primary study analysis. That is, we collect data according to some plan (e.g., first selecting relevant studies, then selecting relevant evidence within those studies); identify substantive and confounding variables; fit appropriate regression models involving those variables; examine and, if needed, modify the model for outliers[6]; plot the findings; and test various hypotheses to answer our specific research questions.

Although a meta-analysis can be as good as the individual studies that form its basis, in this article, we argued that even with high-quality L2 studies forming its basis, a meta-analysis can misrepresent those studies. This misrepresentation stems from ignoring the evidence dependencies that are due to the studies' design complexities. For instance, we saw that evidence for implicit knowledge development and that for explicit knowledge development could be inherently correlated in L2 studies (see *Sampling dependence*). Yet this methodologically induced relationship has nothing to do with a substantive relationship that could potentially exist between the two concepts, as indicated by theory (e.g., interface hypothesis, see Ellis, 2008). Rather, it merely arises from the design of the studies investigating those concepts. Such design complexities should be addressed in L2 meta-analyses so that the substantive phenomena can be

[6]Besides outliers, another issue not covered here, due to space limitation, is publication bias. Readers are referred to Rodgers and Pustejovsky (2021) for more details.

https://doi.org/10.1017/S0272263123000311 Published online by Cambridge University Press

empirically examined. This is especially crucial in the era of evidence-based social sciences to which meta-analysis often seeks to contribute.

It should also be noted that there is far more flexibility to the 3M meta-analysis that falls beyond the scope of this introductory article (but see our supplementary document at: https://rnorouzian.github.io/m/3msup.html where we expand on the basic 3M model introduced in this article). However, as a general framework, we would recommend the approach illustrated in this article for use in L2 research. It is certainly our hope that with this alternative approach to meta-analysis, applied linguists would be better able to offer easy-to-understand, evidence-based recommendations to researchers, teachers, teacher trainers, policy makers, and other stakeholders in the language learning and teaching world. The present article, we believe, is a step in that direction.

Data availability statement. The experiment in this article earned Open Data and Open Materials badges for transparent practices. The materials and data are available at https://github.com/rnorouzian/i/blob/master/3m.r.

References

Aloe, A. M., Thompson, C. G., & Reed, D. K. (2020). CUTOS: A framework for contextualizing evidence. In B. Albers, A. Shlonsky, & R. Mildon (Eds.), *Implementation Science 3.0* (pp. 39–52). Springer.

Becker, B. J. (1996). The generalizability of empirical research results. In C. P. Benbow & D. Lubinski (Eds.), *Intellectual talent: Psychometric and social issues* (pp. 362–383). Johns Hopkins University Press.

Becker, B. J. (2017). Improving the design and use of meta-analyses of career interventions. In J. P. Sampson, E. Bullock-Yowell, V. C. Dozier, D. S. Osborn, & J. G. Lenz, *Integrating theory, research, and practice in vocational psychology: Current status and future directions* (pp. 95–107). Florida State University.

Becker, B. J., & Aloe, A. M. (2008, March 24–28). *Modeling heterogeneity in meta-analysis: Generalizing using Cronbach's (M)UTOS framework and meta-analytic data* [Paper presentation]. Annual meeting of the American Educational Research Association, New York, New York, USA.

Berlin, J. A., & Antman, E. M. (1992, May 10–13). *Advantages and limitations of meta-analytic regressions of clinical trials data* [Paper presentation]. Thirteenth annual meeting of the Society for Clinical Trials. Adam's Mark, Philadelphia, PA, USA. https://doi.org/10.1016/0197-2456(92)90151-O

Brown, D., Liu, Q., & Norouzian, R. (2023). Effectiveness of written corrective feedback in developing L2 accuracy: A Bayesian meta-analysis. *Language Teaching Research* [Online]. https://doi.org/10.1177/13621688221147374

Byrnes, H. (2013). Notes from the editor. *The Modern Language Journal*, *97*, 825–827. https://doi.org/10.1111/j.1540-4781.2013.12051.x

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*, 283–292.

Chong, S. W., & Plonsky, L. (2023). A typology of secondary research in applied linguistics. *Applied Linguistic Review*. https://doi.org/10.1515/applirev-2022-0189

Cook, T. D. (1991). Meta-analysis: Its potential for causal description and causal explanation within program evaluation. In G. Albrecht & H. U. Otto (Eds.), *Social prevention and the social sciences* (pp. 245–285). de Gruyter.

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationships. *New Directions for Program Evaluation*, *1993*, 39–82. https://doi.org/10.1002/ev.1638

Cook, T. D. (2013). Causal generalization: How Campbell and Cronbach influenced my theoretical thinking on this topic, including in Shadish, Cook, and Campbell. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (2nd ed., pp. 85–113). Sage.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*, 369–382. https://doi.org/10.1177/0267658312443651

Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford University Press.

Fernández-Castilla, B., Maes, M., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2019). A demonstration and evaluation of the use of cross-classified random-effects models for meta-analysis. *Behavior Research Methods*, *51*, 1286–1304. https://doi.org/10.3758/s13428-018-1063-2

Fisher, Z., & Tipton, E. (2015). *Robumeta: An R-package for robust variance estimation in meta-analysis*. ArXiv. https://arxiv.org/abs/1503.02220

Fox, J., & Weisberg, S. (2018). Visualizing fit and lack of fit in complex regression models with predictor effect plots and partial residuals. *Journal of Statistical Software*, *87*, 1–27.

Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, *54*, 245–258.

Gries, S. T. (2021). (Generalized linear) Mixed-effects modeling: A learner corpus example. *Language Learning*, *71*(3), 757–798.

Hedges, L. V. (2019). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 281–297). Russell Sage Foundation.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage Publications.

Mathur, M. B., & VanderWeele, T. J. (2019). New metrics for meta-analyses of heterogeneous effects. *Statistics in Medicine*, *38*, 1336–1342.

Matt, G. E., & Cook, T. D. (2019). Threats to the validity of generalized inferences from research syntheses. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 489–516). Russell Sage Foundation.

Norouzian, R. (2020). Sample size planning in quantitative L2 research: A pragmatic approach. *Studies in Second Language Acquisition*, *42*, 849–870.

Norouzian, R. (2021). Interrater reliability in second language meta-analyses: The case of categorical moderators. *Studies in Second Language Acquisition*, *43*, 896–915.

Norouzian, R., & Plonsky, L. (2018). Eta- and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research*, *34*, 257–271.

Norouzian, R., de Miranda, M., & Plonsky, L. (2018). The Bayesian revolution in second language research: An applied approach. *Language Learning*, *68*, 1032–1075.

Norouzian, R., De Miranda, M., & Plonsky, L. (2019). A Bayesian approach to measuring evidence in L2 research: An empirical investigation. *Modern Language Journal*, *103*, 248–261.

Olkin, I., & Gleser, L. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357–376). Russell Sage Foundation.

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, *102*, 713–731. https://doi.org/10.1111/modl.12509

Plonsky, L., Hu, Y., Sudina, E., & Oswald, F. L. (2023). Advancing meta-analytic methods in L2 research. In A. Mackey & S. Gass (Eds.), *Current approaches in second language acquisition research* (pp. 304–333). Wiley Blackwell.

Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition*, *39*, 579–592. https://doi.org/10.1017/S0272263116000231

Plonsky, L., Sudina, E., & Hu, Y. (2021). Applying meta-analysis to research on bilingualism: An introduction. *Bilingualism: Language and Cognition*, *24*, 819–824. https://doi.org/10.1017/S1366728920000760

Pustejovsky, J. E. (2022). *ClubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* (R package version 0.5.5) [Computer software]. https://CRAN.R-project.org/package=clubSandwich

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, *23*(3), 425–438. https://doi.org/10.1007/s11121-021-01246-3

R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Raeisi-Vanani, A., Plonsky, L., Wang, W., Lee, K., & Peng, P. (2022). Applying meta-analytic structural equation modeling to second language research: An introduction. *Research Methods in Applied Linguistics*, *1*(3), 100018.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical bayes meta-analysis. *Journal of Educational Statistics*, *10*, 75–98. https://doi.org/10.3102/10769986010002075

Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, *26*, 141–160. https://doi.org/10.1037/met0000300

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

Shintani, N., & Ellis, R. (2013). The comparative effect of direct written corrective feedback and metalinguistic explanation on learners' explicit and implicit knowledge of the English indefinite article. *Journal of Second Language Writing*, *22*, 286–306.

Stevens, J. R., & Taylor, A. M. (2009). Hierarchical dependence in meta-analysis. *Journal of Educational and Behavioral Statistics*, *34*, 46–73. https://doi.org/10.3102/1076998607309080

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods*, *10*, 161–179. https://doi.org/10.1002/jrsm.1338

Trikalinos, T. A., & Olkin, I. (2012). Meta-analysis of effect sizes reported at multiple time points: A multivariate approach. *Clinical Trials*, *9*, 610–620. https://doi.org/10.1177/1740774512453218

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning*, *46*, 327–369.

Truscott, J. (2004). Evidence and conjecture on the effects of correction: A response to Chandler. *Journal of Second Language Writing*, *13*, 337–343.

Van Beuningen, C. G., De Jong, N. H., & Kuiken, F. (2012). Evidence on the effectiveness of comprehensive error correction in second language writing. *Language Learning*, *62*, 1–41.

Viechtbauer W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, *26*, 37–52. DOI:10.1002/sim.2514

Viechtbauer, W. (2010). Conducting meta-analyses in R with the *metafor* package. *Journal of Statistical Software*, *36*, 1–48.

Viechtbauer, W. (2021). Model checking in meta-analysis. In C. H. Schmid, T. Stijnen, & I. R. White (Eds.), *Handbook of meta-analysis* (pp. 219–254). CRC Press.

Cite this article: Norouzian, R. and Bui, G. (2024). Meta-analysis of second language research with complex research designs. *Studies in Second Language Acquisition,* 46: 251–276. https://doi.org/10.1017/S0272263123000311