A Meta-Analysis of Hybrid Language Instruction and Call for Future Research

TÜLAY DIXON,[1] MARYANN CHRISTISON,[2] DANIEL H. DIXON,[3] AND ADRIAN S. PALMER[4]

[1]*Northern Arizona University, Department of English, 705 S Beaver St, Flagstaff, AZ 86011 Email: to283@nau.edu*

[2]*University of Utah, Department of Linguistics/Urban Institute for Teacher Education, 255 S. Central Campus Dr. #2300, Salt Lake City, UT 84112 Email: ma.christison@utah.edu*

[3]*Northern Arizona University, Department of English, 705 S Beaver St, Flagstaff, AZ 86011 Email: dhd23@nau.edu*

[4]*University of Utah, Department of Linguistics, 255 S. Central Campus Dr. #2300, Salt Lake City, UT 84112 Email: adrian.palmer@utah.edu*

In this meta-analysis, we estimate the effectiveness of hybrid language instruction overall and across a number of moderator variables by aggregating effect sizes from 11 studies with 34 samples. Results suggest hybrid language instruction can be just as effective as traditional face-to-face (f2f) instruction, as indicated by the negligible differences between hybrid courses and traditional f2f courses ($d = .14$). Furthermore, studies employing within-group designs indicate that students in hybrid language classes can improve their language skills considerably ($d = 1.47$). This is a positive finding given that many institutions have experienced a surge in hybrid teaching due to the COVID-19 pandemic. We also report on a number of moderator variables that can impact the effectiveness of hybrid language courses, including (a) the amount of reduction in f2f time, (b) the use of online activities provided by textbook publishers, (c) the use of a learning management system, (d) advances in digital technologies, (e) the targeted language skills (e.g., speaking, writing), and (f) whether the data come from initial or subsequent iterations of a hybrid course. Additionally, we offer directions for future research regarding the substantive and methodological issues in the hybrid language instruction domain.

Keywords: face-to-face instruction; hybrid instruction; language instruction; meta-analysis; online technologies

ONLINE EDUCATION HAS GROWN exponentially in the past decade, making it "the fastest growing area of education in the world today, in both the developed and developing worlds" (Simpson, 2012, p. 1). The global movement to online education has been principally fueled by three crucial factors: (a) globalization, (b) an unprecedented movement of people in the latter decades of the 20th century and the first two decades of the 21st century, and (c) remarkable and seemingly endless advances in digital technologies. To these three factors, a fourth must also be added—namely, the unprecedented global health crisis brought about by the COVID-19 pandemic, which has given educators first-hand knowledge of the extent to which curricular changes can be influenced by and are manifestations of social forces. The pandemic has changed educational curricula and how learners are being educated on a global scale (UNESCO, 2020). It has placed the use of digital technologies and online education at the forefront of concerns about how to design and deliver effective online instruction.

Within the context of language teaching and learning, one of the online instruction

The Modern Language Journal, 105, 4, (2021)
DOI: 10.1111/modl.12732
0026-7902/21/792–809 $1.50/0

models that has received much attention is the hybrid model, which is a combination of face-to-face (f2f) and online instruction with a reduction in f2f time. Previous research on hybrid language instruction often reported statistically nonsignificant differences between hybrid language courses and traditional f2f courses in terms of language gains (Rubio & Thoms, 2014; Thoms, 2020). In the current meta-analysis, we go beyond the dichotomous question of whether there is a difference between the two course delivery formats and quantify the magnitude of the difference between the two course delivery formats by aggregating effect sizes across studies. Aggregation of effect sizes across studies helps mitigate some of the methodological limitations of research in this relatively young field of inquiry, such as the fact that sample sizes may be too small or insufficient for accurate statistical measurement or broad generalizations of findings. The current meta-analysis also aims to shed light on some of the factors that can impact the effectiveness of hybrid instruction. These factors include the targeted language skills (e.g., speaking, listening), the amount of reduced f2f time, and the use of online activities provided by textbook publishers, among others.

In the following four sections, we define hybrid language instruction, discuss the need for a meta-analysis of hybrid language instruction, introduce the variables that can impact the effectiveness of hybrid language courses, and present the research questions guiding this meta-analysis.

HYBRID LANGUAGE INSTRUCTION

Online Learning Consortium (OLC), which was previously known as the Sloan Consortium, studies trends in higher education in the United States. OLC has defined and classified online learning based solely on the percentage of content delivered online (Allen & Seamen, 2013). According to this classification, for example, traditional classes are defined as having 0% of the content delivered online. In the 21st century in most parts of the world, traditional f2f classes with no online component are rare in institutions of higher education (IHEs) because even classes that are deemed f2f have access to and often make use of learning management systems (LMSs) such as Blackboard, Canvas, and Edmodo, if only for the purposes of posting a syllabus. Murray and Christison (2017) noted that a classification that is based only on the percentage of content delivered online fails to capture the range of curricular design options available in online learning, which are crucial to understanding how the online environment affects teaching and learning. They proposed the classification in Table 1, in which online technologies are configured in terms of content, activities, and the sequencing and timing of instructional components.

The design option that has garnered the most attention in teaching second languages (L2s) and foreign languages (FLs) in IHEs has been the blended or hybrid course. Although some practitioners in other disciplines make a distinction between the use of the terms *hybrid* and *blended* (see, e.g., Graham, 2005; O'Rourke, n.d.), we use the terms interchangeably (Rubio, 2014) for the purposes of combining f2f instruction with the use of online technologies. For clarity, in this paper we have chosen to use the term *hybrid* and define hybrid courses as described in Table 1, with a reduction in f2f time.

In spite of the popularity of hybrid courses, practitioners remain uncertain about the effectiveness of the design in relationship to solely f2f classes, which have long been considered hallmarks of effective instruction in IHEs. Consequently, it is understandable that maintaining an f2f component while at the same time making use of online technologies to enhance learning is not only attractive for both students and teachers but, in 2020, has also become a necessity. Consider the fact that as of August 2020, 3,278 IHEs in the United States had moved at least a portion of their courses online in response to the COVID-19 pandemic (Education Data, 2020). Consequently, it comes as no surprise that research that focuses on whether online language learning can be as effective as f2f learning and how to blend the two delivery options is of primary concern for many language educators and institutions.

THE EFFECTIVENESS OF HYBRID LANGUAGE INSTRUCTION

The effectiveness of hybrid language instruction has been examined mainly in the L2 and FL teaching contexts in the United States (e.g., Chenoweth & Murday, 2003; Chenoweth, Ushida, & Murday, 2006; Dixon & Christison, 2021; Scida & Jones, 2016; Scida & Saury, 2006; Thoms, 2014). The findings of previous research were synthesized by Thoms (2014, 2020), who reported that hybrid language courses are generally found to be as effective as f2f traditional courses, if not more effective for certain language skills such as writing. These conclusions, however, are heavily reliant on null hypothesis significance testing (NHST) and *p* values used in previous research, both of which are limited in the types of research questions they

TABLE 1
Online Learning Classifications

Classification	Characteristics
Enhanced	Face-to-face classes are supported by course content and activity online.
Blended or hybrid	There is both face-to-face and online activity; the number of face-to-face classes is reduced in favor of online learning.
Flipped	There is both face-to-face and online activity; the face-to-face time is devoted to interactive problem solving based on information available online.
Synchronous online	All instruction is online; students meet virtually online at the same time.
Asynchronous online	All instruction is online; all activities and tasks can be completed asynchronously.

Note. Adapted from *Online Language Teacher Education: Participants' Experiences and Perspectives* by D. E. Murray & M. Christison, 2018, Monterey, CA: The International Research Foundation (TIRF), p. 17. Copyright 2018 TIRF.

can answer. According to Plonsky (2015), *p* values only allow answers for dichotomous *yes–no* research questions (i.e., is there a difference between the X and Y groups?), rather than a rich description of the extent to which the groups differ from one another (i.e., to what extent do the X and Y groups differ?). Consequently, previous research on hybrid language instruction, which generally relied on NHST, could only address whether there is a difference between hybrid language instruction and traditional f2f instruction, rather than the magnitude of the differences between the two course delivery formats. Additionally, it is possible for statistically nonsignificant differences to be large, especially with small sample sizes (Plonsky, 2015). Thus, to answer the question of to what extent hybrid language instruction is effective, the current meta-analysis aggregates effect sizes (i.e., Cohen's *d*) across studies, going beyond the dichotomous thinking of NHST and quantifying the magnitude of the differences between the two course delivery formats.

The only meta-analysis to date on the effectiveness of hybrid language instruction was conducted by Grgurović, Chapelle, and Shelley (2013). Although the focus of their meta-analysis was on technology-supported language learning more broadly, they reported on hybrid language instruction in moderator analyses. They found mixed results on the effectiveness of hybrid language instruction and attributed the mixed results to the variability within the designs of the primary studies included in the meta-analysis. For a subset of studies, they reported that the effect of hybrid instruction was actually negative ($d = -.21$). This aggregated effect consisted of the studies that had established equivalence of groups using a pretest as well as studies that did not use a pretest for equivalence, such as in cases where participants were randomly assigned to groups. However, they reported a positive medium effect size ($d = .63$) for studies that were analyzed using mean gain scores (rather than raw score comparisons). Mean gain scores were used either because the researchers did not test for equivalence or they did not find equivalence for the groups at the time of the pretest. As a result of these mixed results, Grgurović et al. (2013) called for more research on hybrid instruction to further examine its effectiveness.

It has been 8 years since Grgurović et al.'s (2013) meta-analysis, which warrants a reevaluation of hybrid language instruction with the benefit of including current research. Obviously, the importance and timeliness of this research is further emphasized by the COVID-19 pandemic, which has required most language teachers and learners around the world to prepare for learning in hybrid or totally online environments. Out of the 11 studies included in the current analysis, only 3 studies overlapped with the meta-analysis by Grgurović et al. (2013): Adair–Hauck, Willingham–McLain, & Youngs (2000); Echávez–Solano (2003); and Green & Youngs (2001). There are two reasons for this small overlap. First, in the current meta-analysis, we analyze only the studies that were able to establish statistical equivalence of groups at the beginning of the treatment with respect to the construct(s) measured. This decision was necessary in order to increase the level of certainty in attributing the results of the treatment to hybrid instruction rather than to the preexisting differences between groups. We believed that this was an important decision considering the high stakes involved when transitioning from traditional f2f instruction to hybrid instruction. The second obvious reason for the low overlap is that Grgurović et al.'s (2013) analysis could not include any research published in recent years.

MODERATOR VARIABLES

Each primary study on hybrid language instruction has its own unique design, with differences in the setting, characteristics of the participants, and the language skills measured, among other factors such as the choice of online tools or the amount of f2f instruction. Such design and setting differences cannot all be accounted for in an individual primary study, but the extent to which such differences impact the effectiveness of hybrid language instruction can be examined in meta-analyses through targeted moderator analyses. In the sections that follow, the moderators analyzed are discussed and justified in light of previous research.

Language-Learning Outcomes

Concerns have been raised regarding the effectiveness of hybrid instruction on the development of speaking skills. Because of the reduction in f2f time, students may receive less input from the teacher and have fewer opportunities to interact f2f with their peers. However, students in hybrid courses have opportunities for online input. This type of input seems to be important because a number of studies have suggested that the students in hybrid learning environments developed their speaking skills as much as the students in traditional f2f courses (e.g., Rubio, 2014; Thoms, 2014). In terms of the development of writing skills, some studies reported that students in the hybrid group had larger gains in their writing skills than other language skills (e.g., Chenoweth & Murray, 2003; Thoms, 2014). Thoms (2014) hypothesized that the larger gains in writing skills may be due to the fact that the students in the hybrid group spent more time writing and reading when completing online tasks than the students in an f2f context. What seems clear is that hybrid instruction can have varying effects on language skill development, and the current study aims to aggregate these effects through moderator analyses.

Design Features of the Hybrid Course

The ratio of f2f time versus online time can moderate the effectiveness of hybrid language instruction. Dixon and Christison (2021) called for research that examines how the balance between f2f and online activities impact learning outcomes. Similarly, Zhang and Zhu (2018) pointed out that "there is no specific experimental or intervention research that [has] attempted to investigate the different ratio of f2f and online sessions for [blended learning]" (p. 267), calling for future research. The association between the amount of reduction in f2f time and course effectiveness can be examined by correlating the effect sizes obtained from primary studies with the amount of f2f time that was reduced in these studies, which is one of the aims of the current meta-analysis.

Another design element that can impact the effectiveness of hybrid language courses is whether the selected course textbook has accompanying online activities. In some hybrid courses, students are asked to complete online activities that are provided by the publishers of textbooks (e.g., *Nexos* for Spanish). Such activities often claim to give 'intelligent feedback' to students—that is, when students submit their answers to online activities, they receive explanations as to why answers are correct or incorrect and are then directed to certain sections of the textbook for further information. For various reasons, not all language textbooks have accompanying online activities. In the absence of such activities, it becomes the responsibility of the course instructor to design online activities. Creating these types of activities is time-consuming (Young, 2008) and can impact the effectiveness of instruction in other ways because the amount of time an instructor has is a limited commodity.

Course design is a cyclical process (Christison & Murray, 2020) in that each time a course is designed and taught, it can be improved based on the feedback from students and instructors. Thus, it is likely that there is an association between the effectiveness of hybrid language instruction and whether or not it was the first time the course was taught in a hybrid format. Scida and Jones (2016), for example, found that students in a redesigned hybrid Spanish course improved their language skills more than those in the previous iteration of the hybrid course, thereby indicating that evaluation of hybrid courses should consider data beyond their initial offerings that may have resulted in negative or small effects.

Another design feature that might impact the effectiveness of hybrid language courses is whether a LMS is used to deliver instruction during the online days. The use of a LMS can be advantageous as instructors can keep all course information contained in a shared online space. With this in mind, we analyze the extent to which the use of a LMS moderates the effectiveness of hybrid language instruction.

Advances in Digital Technologies

As online technologies continue to evolve, the time frame in which a hybrid language course was taught is likely to impact the learning outcomes. Scida and Jones (2016) stressed "the need for ongoing evaluation of hybrid language programs and reconsideration of blended learning in light of constantly evolving technologies and changing instructional needs and learning paradigms" (pp. 193–194). The technological tools used 20 years ago are now considered out of date. For example, Adair–Hauck et al. (2000) report that the students in the hybrid group used videocassettes during the online days as one of the 'online' asynchronous activities and engaged in other online activities that they could access only on university servers. More advanced online technologies are now available to the designers of hybrid courses, giving both instructors and language learners more flexibility and access to a variety of online tools. We examine the moderating effect of evolving technologies by correlating the effect sizes with time of publication—with time of publication used as a proxy for the time period in which the course was taught.

RESEARCH QUESTIONS

This meta-analysis is guided by the following research questions:

RQ1. To what extent is hybrid language instruction effective based on the aggregated results of primary research?

RQ2. To what extent is the effectiveness of hybrid language instruction influenced by the following moderators?

(a) outcome measures (e.g., four skills, grammar, pronunciation, vocabulary)
(b) design features of the hybrid or blended design (e.g., the amount of reduction in f2f class time, the use of online activities provided by textbook publishers)
(c) advances in digital technologies

METHOD

Literature Search

Using several techniques, we completed a comprehensive literature search. First, we searched the following databases: Academic Search Complete, Education Abstracts, Education Full Text, Education Resources Information Center (ERIC), JSTOR, Linguistics and Language Behavior Abstracts (LLBA), Modern Language Association (MLA) International Bibliography, Proquest Dissertations and Theses, PsychArticles, PsychINFO, and Web of Science as well as Google and Google Scholar. In the searches, we used the following key terms (the asterisk is used for retrieving terms that start with the letters preceding the asterisk but have different endings; e.g., "learn*" would return studies that used the terms *learning* and *learners*):

> (hybrid OR blended) AND ("second lang*" OR "foreign lang*" OR "target lang*" OR FL OR L2 OR ESL OR EFL OR EAL OR ELT) AND (proficiency OR gain* OR achievement OR "lang* learn*" OR "second language acquisition" OR vocabulary OR grammar OR pronunciation OR listening OR speaking OR writing OR reading) AND (intervention OR treatment OR control OR comparison OR experiment OR effect OR impact OR outcome)

In addition to the database searches, we manually searched the following journals using the terms "hybrid OR blended": *Language Learning and Technology*, *CALICO*, *ReCALL*, *Computer Assisted Language Learning*, and *The Modern Language Journal*. Additionally, we checked the literature reviews of relevant studies (e.g., Kraemer, 2008; Rubio, 2014; Thoms, 2014). Using Google Scholar, we also examined the studies that cited some of the influential studies or books on hybrid language instruction (e.g., Adair–Hauck et al., 2000; Chenoweth et al., 2006; Rubio & Thoms, 2014). Last, we examined the studies that were included in previous meta-analyses on hybrid or blended instruction (i.e., Grgurović et al., 2013; Mahmud, 2018). The search was completed during July 2020 and yielded 90 studies, which were further examined using a set of inclusion and exclusion criteria.

Study Eligibility Criteria

Using the inclusion and exclusion criteria listed in Table 2, we screened the 90 studies downloaded from the search and went through two cycles of screening to determine whether the studies fit the criteria. In the first cycle, we first examined whether the studies measured language-learning outcomes rather than perceptual factors such as anxiety, motivation, and attitudes. We then examined whether there was a reduction in f2f time, as per the definition we adopted for hybrid instruction. Identifying the reduction in f2f time was challenging because the terms *hybrid* and *blended* are used inconsistently across the literature. Some studies used the terms *hybrid* or *blended* to refer

TABLE 2
Inclusion and Exclusion Criteria

Criteria for Inclusion	Criteria for Exclusion
1. The study measured language learning outcomes. 2. The study employed an experimental or quasi-experimental between-groups designs with a pretest and a posttest or a within-group design with a pretest and a posttest.	1. The study adopted a hybrid or blended instruction model that did not have a reduction in f2f time. 2. The study measured factors other than language-learning outcomes (e.g., attitudes, motivation, and anxiety). 3. The study did not report descriptive statistics or the reported statistics were insufficient to calculate an effect size.

Note. f2f = face-to-face.

to instruction that used online technologies without a reduction in f2f time. Some studies did not provide clear definitions of the terms *hybrid* and *blended* or did not report how much time the hybrid or blended group spent in class and online versus the comparison or control groups. Because of the challenge in determining whether there was a reduction in f2f time, each article was screened twice by two of the researchers. After the first cycle of screening, we excluded 62 studies, leaving 28 studies for further screening.

The remaining 28 studies defined the terms *hybrid* (or *blended*) in the same way they are defined in this meta-analysis. However, we had to exclude 17 of them for two reasons: (a) no pretests for the dependent variables of interest, and (b) no reporting, or insufficient reporting, of the descriptive statistics required to calculate effect sizes (i.e., means and standard deviations). Our final sample included 11 studies with 34 samples (see Appendix A for a list of the studies used in the meta-analysis).

The final sample of 11 studies examined the effectiveness of hybrid instruction for a number of languages: Spanish ($n = 5$), English ($n = 3$), French ($n = 2$), Chinese ($n = 1$), and German ($n = 1$). Except for Dixon & Christison (2021), all studies were conducted in an FL setting. Besides the three studies that examined the effectiveness of hybrid instruction for teaching English, the remaining studies were conducted in hybrid FL courses in the United States, all of which could be labeled as first-year FL courses as they were either a first- or second-semester FL course. In total, these 11 studies had 1,005 participants.

Coding

The initial coding scheme was tested with five studies and updated based on insight gained from this first round of coding. Using the updated coding scheme, we coded all eligible studies ($n = 11$; $k = 34$) for a number of substantive and methodological features:

1. Study identification (e.g., citation, authors, publication type)
2. Design of the hybrid course (e.g., the LMS used, percentage of reduction in f2f time, whether the course was taught in a hybrid format for the first time)
3. Learner and contextual differences (e.g., L2 vs. FL setting; higher education vs. primary and secondary education [K–12]; target language)
4. Study design and reporting practices (e.g., within-group vs. between-groups design; dependent variables [DVs]; the reliability of the instrument used to measure the DVs; reporting of effect sizes and statistical assumptions)
5. Results (e.g., *n*, *M*, and *SD* for pretests and posttests; *d* values for the within-group and between-groups comparisons)

There are two items in the coding scheme that we anticipated needing further explanation. In determining whether the data for a given DV was within-group data or between-groups data, we considered the equivalence of groups at the pretest. We considered equivalence of groups at the pretest to avoid including any data where the differences might be due to preexisting differences between the treatment and comparison or control groups, rather than the actual treatment of hybrid instruction itself. As a result, we excluded data from some comparison or control groups even though the researcher(s) intended to have a between-groups design, because this equivalence was either not established or not

TABLE 3
Effect Size Interpretation (Plonsky & Oswald, 2014)

Design	Small	Medium	Large
Within-group	Values around .60	Values around 1.00	Values around 1.40
Between-groups	Values around .40	Values around .70	Values around 1.00

reported. To determine equivalence of groups, we considered the following:

1. When reporting the pretest differences between groups, most researchers only referred to significant versus nonsignificant differences. In such cases, using the reported means and standard deviations, we calculated effect sizes (i.e., Cohen's *d*) to check whether the nonsignificant differences were also small differences.
2. In interpreting the effect sizes, we used the benchmarks suggested by Plonsky & Oswald (2014) for L2[1] research with between-groups designs: *d* values around .40 as a small effect size, values around .70 as a medium effect size, and values around 1.00 as a large effect size. We considered groups to be equivalent if the *d* value was .5 or below .5, as *d* values above this cutoff point can be considered in the small to medium range. If the *d* value for the pretest differences was above .5, we treated the data as within-group data, excluding the data from the comparison or control group.

For studies employing a within-group design, for each of the dependent variables, an effect size was calculated by contrasting the mean scores between the pretest and the posttest. For between-groups studies, an effect size was calculated by contrasting (a) the mean scores between the groups at the posttest, and (b) the treatment group's mean scores at the pretest with their scores at the posttest.

Each item on the coding scheme was double coded independently. Once all coding was complete, the interrater reliability was measured using Langtest (Mizumoto, 2015; Mizumoto & Plonsky, 2016). A breakdown of the reliability scores for each item on the coding sheet can be seen in Appendix B. The average percentage agreement was 96.79%, and the average of Cohen's kappa values was .95, indicating a high level of consistency among coders. Through interrater reliability measurements, we identified areas of disagreement in the coding scheme, and these were discussed to reach an agreement before proceeding with data analysis. The coding scheme is available in the https://www.iris-database.org/.

Aggregating Effect Sizes

The effect sizes retrieved from each sample were weighted by their sample sizes. This decision was made because sample size is inversely related to sampling error (Blair & Blair, 2015)—that is, small samples tend to have larger errors in estimating the population parameters. Thus, to be more accurate in our estimates of how effective hybrid language instruction is, we gave more weight to studies with larger samples. In addition to the weighted effect sizes, we also report the unweighted effect sizes to allow for transparent interpretation of results. All effect sizes were interpreted using the benchmarks suggested for L2 research by Plonsky & Oswald (2014), which are listed in Table 3.

Analysis

We report effect sizes that come from within-group designs separately from effect sizes that come from between-groups designs. The reason for this separation is because within-group designs tend to have higher effect sizes than between-groups designs (see Plonsky & Oswald, 2014; Plonsky & Zhuang, 2019). This separation allows for a more precise and informative aggregation of effect sizes. For the moderator analysis, we created subgroups of studies that contained the variables of interest.

RESULTS AND DISCUSSION

Publication Bias

Due to authorial and/or editorial bias, not all studies with nonsignificant results reach publication, which is the most common form of publication bias (Norris & Ortega, 2000). To examine the extent to which the sample used in the current study may be affected by publication bias, we created scatterplots with the effect sizes on the

FIGURE 1
Scatterplot of Between-Groups Effects and Sample Sizes [Color figure can be viewed at wileyonlinelibrary.com]

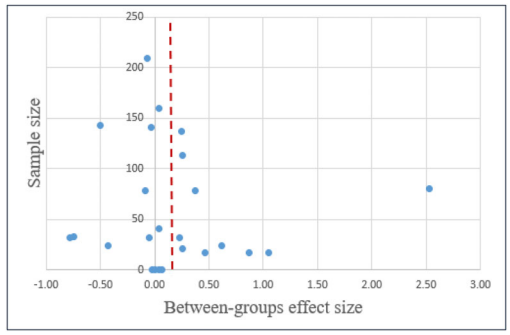
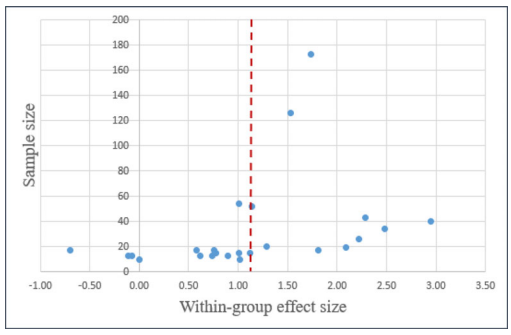


FIGURE 2
Scatterplot of Within-Group Effects and Sample Sizes [Color figure can be viewed at wileyonlinelibrary.com]



x axis and sample sizes on the *y* axis (see Figures 1 and 2). In the absence of publication bias, we would expect to see a triangular funnel with effect sizes spread somewhat equally on both sides of the mean effect size, which is indicated by the dashed line in both figures. In other words, studies with larger sample sizes are expected to converge on the true population value at the upper end of the funnel because larger samples often have less sampling error (Blair & Blair, 2015). The studies with small sample sizes, in contrast, often spread across the bottom of the funnel as they are likely to have more sampling error.

In Figure 1, the effect sizes seem to be equally spread on both sides of the overall effect, suggesting no strong evidence of publication bias for between-groups designs. In Figure 2, there is almost an equal number of effect sizes on both sides of the mean effect size, indicating once again no strong evidence of publication bias. However, it appears that convergence on the true effect size is just beginning to happen for within-group designs, as most effect sizes are spread across the bottom of the funnel. This indicates the need for more primary studies with larger sample sizes in order to reduce sampling error and estimate the true effect size.

Overall Effectiveness of Hybrid Language Instruction

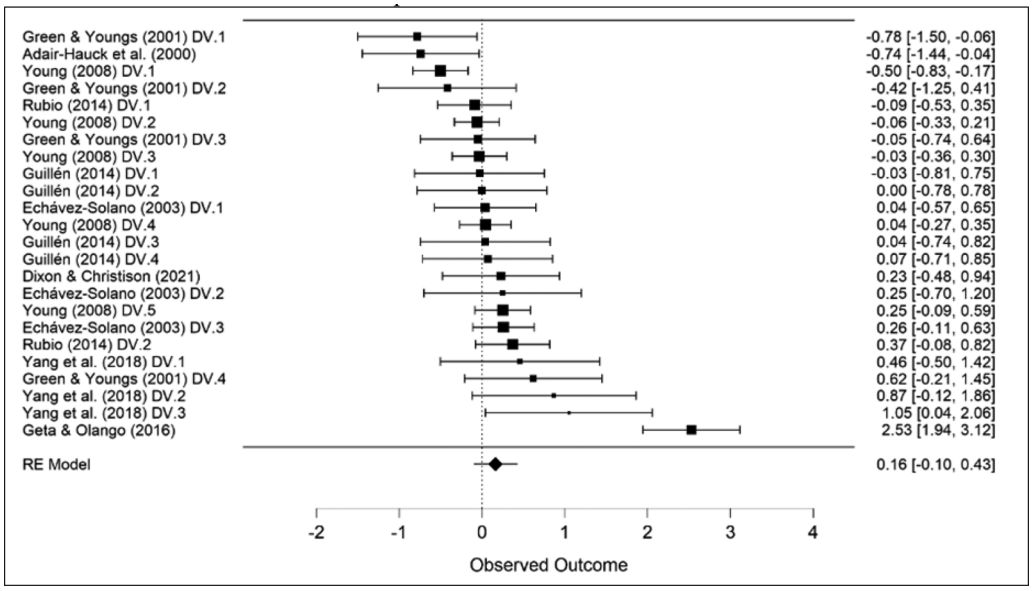
RQ1 focuses on the overall effectiveness of hybrid language instruction. Table 4 presents the overall effect of hybrid language instruction for within-group and between-groups designs separately. It is also important to note that, in the between-groups designs, the control or comparison groups were all considered to be traditional f2f instruction groups, and both the weighted (.14) and the unweighted (.16) effect sizes indicate that there is a negligible difference between hybrid language instruction groups and control or comparison groups (see Table 4). The 95% confidence intervals of [−.10, .43] indicate that the difference between groups, as represented by Cohen's *d*, can be as low as −.10 or as high as .43. In either direction, this difference between groups can be considered negligible, as .10 and .43 are both small effect sizes. For within-group designs, the language gains from pretest to posttest can be considered large ($d_{\text{weighted}} = 1.47$).

TABLE 4
Overall Results for the Effectiveness of Hybrid Language Instruction

					95% CIs	
Contrast	*k*	$M_{d\,(\text{weighted})}$	$M_{d\,(\text{unweighted})}$	*SE*	Lower	Upper
Between-groups	24	.14	.16	.135	−.10	.43
Within-group	24	1.47	1.15	.181	.80	1.51

Note. k = number of samples; $M_{\text{d(weighted)}}$ = mean of effect sizes weighted by sample size; *SE* = standard error; CI = confidence interval. 95% CIs are around the unweighted *d* values.

FIGURE 3
Forest Plot of Overall Between-Groups Effects



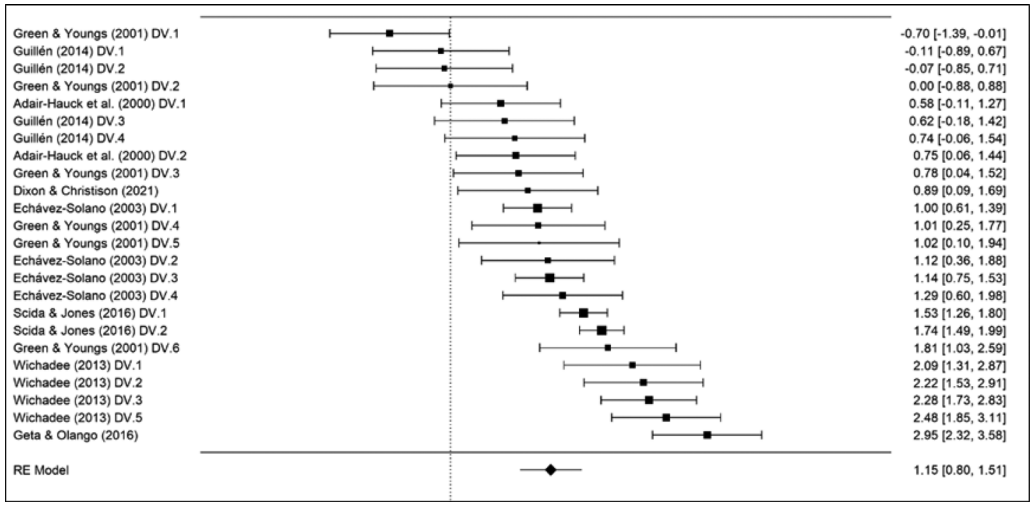
Note. DV = dependent variable; RE = random effects. Numbers following DV indicate different DVs in a single study. Effect sizes are unweighted *d* values with 95% confidence intervals. Squares represent the standard error of each sample, with smaller squares indicating larger standard errors.

The aggregated results from between-groups designs suggest that the previously reported statistically nonsignificant differences between hybrid and f2f language courses are also small differences, giving further support to the claim that hybrid language courses can be as effective as traditional f2f courses. This negligible difference between the two course delivery formats is good news for institutions either requiring or considering a reduction in f2f language teaching in favor of online instruction. Within-group results give further positive support for the use of hybrid instruction in L2 learning contexts, as the weighted *d* value was a large effect ($d = 1.47$). These results, however, cannot be generalized to learners outside of a university context because they are based on studies that employed only university students as participants. While we did not intentionally exclude studies that took place outside of a university, our comprehensive search of the literature did not return any studies from non-university populations. We return to this need for research involving non-university language learners in our section on recommendations for future research.

To visually represent the aggregated effects of hybrid instruction, Figures 3 and 4 display the effect sizes obtained from all the studies included in the meta-analysis. The effect sizes in both figures are unweighted *d* values with 95% confidence intervals. Each row represents the results from one dependent variable on a single sample. The squares in the middle of each line represent the standard error of each sample. Smaller squares indicate larger standard errors, which tend to be accompanied with wider confidence intervals. These confidence intervals are visualized by the lines stemming from the squares, and longer lines indicate wider confidence intervals. The rhombus at the bottom represents the aggregated mean value of the unweighted effect sizes, and the lines stemming from it represent the aggregate 95% confidence interval. The figures were produced using the software suite *JASP* (JASP Team, 2020).

In Figure 4, there are three negative effects for within-group designs: one sample from Green & Youngs (2001) and two samples from Guillén (2014). Negative values for within-group designs are often surprising, as one would expect at least some gains as a result of instruction, even if the gains are small. These negative values measured oral skills in the case of Green & Youngs (2001) and measured fluency and pronunciation in the case of Guillén (2014). Green and Youngs (2001) did not address the lack of gains when discussing the results. Guillén (2014), however, stated that a 10-week course was perhaps not enough time to improve fluency and pronunciation. Despite

FIGURE 4
Forest Plot of Overall Within-Group Effects



Note. DV = dependent variable; RE = random effects. Numbers following DV indicate different DVs in a single study. Effect sizes are unweighted *d* values with 95% confidence intervals. Squares represent the standard error of each sample, with smaller squares indicating larger standard errors.

the negative effects found in these three samples, the remaining 21 samples showed positive gains in L2 skill development, giving clear support for the potential effectiveness of hybrid language instruction.

Moderator Analysis

RQ2 focuses on the moderating effects of (a) outcome measures, (b) design features of the hybrid course, and (c) advances in digital technologies. Tables 5 and 6 report the results of several moderator analyses for between-groups and within-group designs, respectively. The first rows in Table 5 list the effects of hybrid language instruction on different outcome measures (e.g., speaking skills and listening skills) for studies using a between-groups design. The category 'other' includes measures that were, out of necessity, grouped together and include pronunciation, vocabulary, fluency, overall proficiency, grammar and editing skills, composite scores for reading and writing, and composite scores for grammar and vocabulary. These measures needed to be grouped because only a single study measured the skill, and at least two samples measuring the same skill would be needed to aggregate effects.

Except for writing, there were negligible differences between groups for all outcome measures (see Table 5). This result indicates that students in the hybrid classes can improve various language skills as much as the students in traditional f2f classes. One finding that needs to be highlighted is that the students in hybrid courses improved their speaking skills as much as the students in f2f courses ($d_{\text{weighted}} = .07$), indicating that the reduced f2f time in hybrid courses does not necessarily jeopardize the development of speaking skills. With respect to writing skills, there is a medium to large effect from hybrid instruction ($d_{\text{weighted}} = 1.20$). Thoms (2014) hypothesized that students in the hybrid instruction group had to read and write more often to complete online exercises, which, in turn, might have had a positive impact on the development of L2 writing skills. Although the weighted and unweighted *d* values indicate a large effect from hybrid instruction on the development of writing skills, the 95% confidence intervals of [−.44, 2.65] around the unweighted mean are quite wide and cross zero. These results suggest that the hybrid groups can be expected to perform on a spectrum ranging from slightly worse to substantially better compared to the traditional groups with respect to their writing skills, a finding highlighting the need for more research in this domain. With such wide confidence intervals and the sample size of two, it is difficult to draw any strong inferences about the impact of hybrid instruction on writing skills. With more primary studies that target writing skills, future meta-analyses can aggregate more precise

TABLE 5
Between-Groups Moderator Analysis

Contrast	*k*	$M_{d(\text{weighted})}$	$M_{d(\text{unweighted})}$	*SE*	95% CIs Lower	95% CIs Upper
Outcome measures						
Reading	2	.02	.02	.159	−.29	.33
Writing	2	1.20	1.10	.788	−.44	2.65
Listening	2	.13	.13	.121	−.11	.37
Speaking	7	.07	.04	.192	−.33	.42
Other[a]	10	−.08	−.05	.123	−.29	.19
First-time hybrid						
Yes	5	.07	.07	.175	−.27	.41
No	11	.14	.15	.252	−.35	.64
LMS used						
Yes	15	.20	.34	.183	−.02	.69
No	9	−.18	−.17	.147	−.46	.12
Publisher activities						
Yes	6	.09	.06	.143	−.22	.34
No	11	.14	.16	.260	−.34	.67

Note. SE = standard error; CIs = confidence intervals; LMS = learning management system.
[a]This category grouped measures including pronunciation, vocabulary, fluency, overall proficiency, grammar and editing skills, composite scores for reading and writing, and composite scores for grammar and vocabulary.

TABLE 6
Within-Group Moderator Analysis

Contrast	*k*	$M_{d(\text{weighted})}$	$M_{d(\text{unweighted})}$	*SE*	95% CIs Lower	95% CIs Upper
Outcome measures						
Listening	3	1.32	1.25	.169	.92	1.58
Speaking	6	.62	.58	.320	−.05	1.21
Cultural knowledge	4	1.06	1.03	.279	.48	1.58
Other[a]	11	1.82	1.46	.318	.84	2.09
First-time hybrid						
Yes	5	.42	.41	.212	−.01	.82
No	10	1.39	1.23	.208	.82	1.64
LMS used						
Yes	12	1.72	1.72	.188	1.35	2.09
No	12	.55	.53	.194	.15	.91
Publisher activities						
Yes	8	1.42	1.22	.144	.93	1.50
No	8	1.35	.98	.398	.20	1.76

Note. SE = standard error; CIs = confidence intervals; LMS = learning management system.
[a]This category grouped measures including pronunciation, vocabulary, fluency, overall proficiency, grammar and editing skills, composite scores for reading and writing, and composite scores for grammar and vocabulary.

estimates, which would likely result in tighter confidence intervals.

For within-group designs (Table 6), the effect of hybrid language instruction is medium for teaching cultural knowledge and large for listening skills and other language skills that were grouped together due to lack of studies measuring these skills (e.g., vocabulary, composite scores for reading and writing, fluency, overall proficiency). The effect of hybrid language instruction on speaking skills is small ($d_{\text{weighted}} = .62$). This small progress in speaking skills cannot be attributed to the fact that students received instruction in a hybrid format because the between-groups effects in

Table 5 show that the students in hybrid classes made as much progress in their speaking skills as those in traditional f2f classes. That speaking skills showed small gains contradicts the recent finding of Winke et al. (2020), who reported that students in lower division FL classes in the United States tend to make greater gains in speaking skills compared to other language skills. The contradiction comes from the fact that the speaking and listening samples in Table 6 all come from first- or second-semester FL courses (i.e., lower division courses) in the United States. However, we would like to note that the present meta-analysis has only three samples for listening and six for speaking. More conclusive generalizations can be made with respect to which language skills develop the most in relation to course levels when there is more research on hybrid language instruction, another direction for future research.

Courses that had been taught in a hybrid format at least once before resulted in much larger effects. This effect was found to be especially common for within-group designs, as seen in Table 6. Courses that were taught in hybrid format for the first time had a small effect ($d_{\text{weighted}} = .42$) whereas the courses that were taught in a hybrid format previously one or more times showed a large effect ($d_{\text{weighted}} = 1.39$). The improved effect may be due to the fact that instructors had more experience with hybrid instruction or had a chance to refine instructional tasks with the added benefit of improved technologies. This finding empirically substantiates the intuition that redesigned courses can result in better learning outcomes and has important implications for stakeholders making curricular decisions based on research results. Language programs aiming to transition from f2f to hybrid delivery formats should keep in mind the cyclical nature of course design and consider collecting data over more than one semester rather than making curricular decisions after an initial offering of a hybrid course.

The use of a LMS appears to have had a moderating effect. For between-groups designs, courses that did not use a LMS had a weighted *d* value of −.18, but courses that used a LMS had a *d* value of .20. Although there is a greater effect for the courses that use a LMS, confidence intervals for both types of courses cross zero, indicating that the use of a LMS may or may not offer an advantage. For within-group designs, however, the advantage of using a LMS is clearer. Courses that did not use a LMS had a small effect ($d_{\text{weighted}} = .55$) whereas the ones that used a LMS had a large effect ($d_{\text{weighted}} = 1.72$). Although the use of a LMS was coded dichotomously in the current study, LMSs were used in various ways in each study, making it difficult to pinpoint which specific aspects of a LMS contribute to language learning. For example, Young (2008) reported using a LMS to transfer the activities in the course textbook to a digital platform, which was also reported in Dixon & Christison (2021), as the selected textbook did not include an online workbook provided by the publisher. Young also reported listening comprehension exercises as well as collaborative asynchronous writing assignments delivered through the LMS. Yang, Yin, and Wang (2018) reported using a LMS for quizzes that tested the comprehension of 10-minute video clips introducing new vocabulary and grammar. These various uses of a LMS indicate that there is more to the dichotomous distinction between whether a LMS was used or not and that what matters is using a LMS in ways that foster gains in the targeted language skills. Despite such varying uses of LMSs, what remains clear is that such activities during the online days would not be possible without a LMS. LMSs allow educators to keep course content centralized, while also allowing students to access course content anywhere and at any time, which gives students autonomy over their own learning and more flexibility in scheduling and opportunities for independent learning (Murray & Christison, 2017). LMSs also make it easy to track and report student progress. Additionally, the new available tools in LMSs (e.g., built-in tools that allow instructors to assign collaborative group projects, give richer feedback either automatically or manually, or even create personalized adaptive learning paths) can allow student–student, student–teacher, and whole-class interactions in a virtual environment.

The inclusion of online activities provided by textbook publishers does not appear to have a moderating effect for between-groups or within-group designs as indicated by the negligible differences in effect sizes reported in Tables 5 and 6. This finding is of importance to language program coordinators and instructors who need to make critical decisions with respect to textbook selection. Given that the activities created or adapted by instructors are just as effective as the online activities provided by textbook publishers, the decision language programs need to make is whether the instructors can spare the considerable time it takes to create and transfer such activities to a LMS. For example, Young (2008) reported that multiple personnel spent an entire semester creating, adapting, and transferring activities to a LMS.

FIGURE 5
Scatterplot of Reduction in Face-to-Face Time and Within-Group Effect Size ($r = -.419$)

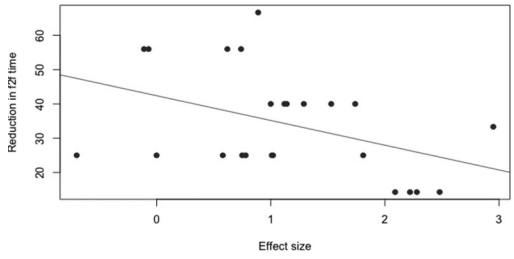


FIGURE 6
Scatterplot of Reduction in Face-to-Face Time and Between-Groups Effect Size ($r_s = .182$; $r = .036$)

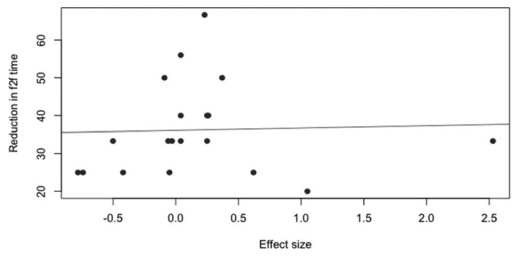


FIGURE 7
Scatterplot of Publication Year and Within-Group Effect Size ($r = .351$)

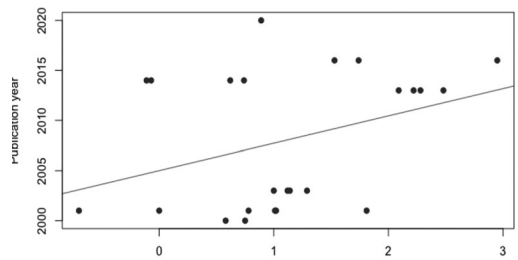
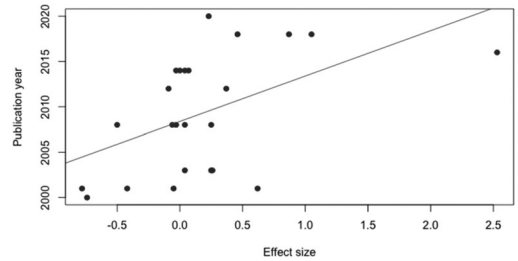


FIGURE 8
Scatterplot of Publication Year and Between-Groups Effect Size ($r_s = .524$; $r = .507$)



To examine whether the amount of f2f reduction has a moderating effect, we ran correlations between the percentage of reduction in f2f time and effect sizes. The assumption of normality was checked before running the correlations. Normality could be assumed for the within-group data (Shapiro–Wilk = .947, $p = .237$), but it could not be assumed for the between-groups data (Shapiro–Wilk = .790, $p < .001$). Thus, we report and interpret Spearman's *rho* for between-groups data.

Figures 5 and 6 present the correlations for within-group and between-groups designs, respectively. In Figure 5, a moderate negative correlation ($r = -.419$) between the amount of reduction in f2f time and effect sizes was found. In other words, as the reduction in f2f time increases, the effect size decreases. However, this trend is not corroborated by the between-groups data, as there is a very weak association between the reduction in f2f time and effect sizes (r_s= .182; $r = .036$). Thus, it is difficult to recommend an optimal percentage for the reduction of f2f time in the design of a hybrid course. Nevertheless, based on within-group data, it appears that program administrators should be cautious when reducing the amount of f2f time, as large reductions in f2f time seem to decrease course effectiveness.

We believe that the focus of the language class (e.g., writing or conversational skills) should be carefully considered when deciding on the reduction of f2f time. While a writing course could be effective with a greater reduction in f2f time, a course whose primary focus is on the development of conversational skills could benefit from a lesser reduction in f2f time, as the development of conversational skills requires co-construction of meaning and turn taking in real time. With more research, it will be possible to examine the effects of differing amounts of reduction in f2f time on various language skills.

To understand the extent to which new technologies affected hybrid language instruction, we ran correlations between time of publication and effect sizes (see Figures 7 and 8). Our initial coding scheme had a category for the years in which the hybrid courses were taught; however, very few studies reported this information. Thus, we decided to use publication year as a proxy for the year the courses were taught. The assumption of normality was met for the within-group data (Shapiro–Wilk = .976, $p = .814$) but not for between-groups data (Shapiro–Wilk = .808, $p <$

.001); thus, we report Spearman's *rho* for between-groups data.

Time of publication appears to have a moderating effect. Although the correlation between time of publication and effect sizes can be considered somewhat weak for within-group data ($r = .351$), there is a moderate association for between-groups data ($r_s = .524$; $r = .507$), indicating that early hybrid courses had smaller effects compared to the hybrid courses taught more recently. This result is not surprising; as technology has improved, so has the effectiveness of hybrid language courses that draw on such technology. It seems that language teachers and learners have found pedagogically sound applications for new technologies, such as the use of LMSs and synchronous tools (such as Zoom, Skype for Business, and GoToMeeting) and have been able to use them effectively in the delivery of hybrid courses. Furthermore, this correlation is likely strengthened by the application of research findings on hybrid instruction, which is a relatively young domain in applied linguistics research.

CALL FOR FUTURE RESEARCH ON HYBRID LANGUAGE INSTRUCTION

Although our extensive search resulted in 90 'hybrid' studies, only 11 studies and 34 samples were found to meet the goals of this study and the inclusion criteria, which was set prior to collecting data. To more accurately aggregate estimates of the effectiveness of hybrid language instruction (i.e., narrower confidence intervals around effect sizes), we need more primary studies that can be included in a meta-analysis of this type. Therefore, we offer suggestions for future primary research on hybrid language instruction.

The methodological rigor of future primary research could be improved by including pretests for all dependent variables that are measured on the posttests. Without pretests, it is difficult to attribute language gains to the treatment of hybrid instruction itself, as the two groups (i.e., hybrid vs. comparison group) might differ substantially at the beginning of the treatment with regard to the construct of interest. By establishing equivalence of groups before the treatment, we can be more confident in attributing the differences between groups to the treatment itself.

To determine whether the two groups are equal at the beginning of the treatment with respect to the construct(s) being measured, we recommend that researchers not only run inferential statistics, such as *t* tests, but also examine the magnitude of the differences between groups by calculating effect sizes. We base our suggestion on the likelihood of calculating nonsignificant differences with large effect sizes. This is especially the case with small sample sizes because *p* values are largely dependent on sample sizes, meaning that the smaller the sample is, the more likely it is to get nonsignificant results (Plonsky, 2015). Effect sizes, however, are calculated without taking the sample size into consideration. Thus, we recommend that researchers refer to both sources of information to determine whether the groups are at about the same level in terms of their knowledge of the construct of interest.

Another suggestion relates to the amount of detail given in reporting results. In the current meta-analysis, instrument reliability was reported for 15 of the 34 dependent variables. Statistical assumptions were discussed in only 4 of the 11 studies. Effect sizes were reported in only 3 studies, and surprisingly only 1 study provided any interpretations for the reported effect sizes. Additionally, our search identified several studies that could not be used in the current meta-analysis because the descriptive statistics that were needed to calculate effect sizes were not reported, especially the standard deviations associated with the means. Attempts to gather missing data by contacting authors were not successful, as those who replied to inquiries stated no longer having access to their data.

Reporting practices have been a concern not only for the domain of research on hybrid language instruction but also for the field of applied linguistics more generally, a concern raised by many scholars in the field (Larson–Hall & Plonsky, 2015; Norris & Ortega, 2000; Plonsky & Zhuang, 2019; Plonsky & Ziegler, 2016). To advance the research on hybrid language instruction—and the field of applied linguistics—we encourage future researchers to report descriptive statistics (*n*, *M*, and *SD*) for all measures, the reliability of the instruments used to measure the dependent variables, and effect sizes for the comparisons made.

There is still much that is unknown about hybrid language instruction. The need for studies in K–12 settings is critical. In March 2020, the Organisation for Economic Co-operation and Development estimated that there were 421 million children in 39 countries affected by school closures. As a result, children began moving to home schooling, online learning, and combinations of f2f and online learning (World Economic Forum, 2020). By July 2020, the number had grown to 1,184,126,508 children in 143 countries (UNESCO, 2020). Yet, all studies that met our

inclusion criteria were from higher education settings, and all but one was in an FL setting. Thus, we make a plea for future research that focuses on K–12 contexts so that relevant stakeholders can make informed curricular decisions about course delivery formats. In addition, no study to date had empirically tested how different ratios of f2f and online instruction impact learning outcomes, especially with respect to different language skills. We encourage future research on these underexplored areas.

Our penultimate suggestion relates to the use of terminology to describe the configurations of online learning. As a result of this meta-analysis, we became aware of the fact that the terms *hybrid instruction, blended instruction, flipped instruction*, and *online instruction* are used inconsistently across the literature. In the case of this particular meta-analysis, the inconsistent use of these terms made it challenging to identify the studies that employed hybrid instruction. We recommend that researchers take better care in defining terms as they relate to types of configurations for online learning, such as those described in Table 1, so that each configuration is distinct. As digital technologies advance, new opportunities will arise. Further, the COVID-19 pandemic has had a powerful impact on shaping the way instruction has been delivered globally. For example, remote learning emerged in K–12 contexts in response to parents' concerns for safety and wellness. In remote learning, children join f2f classes from remote locations and can interact with their teachers and peers in real time. The pandemic motivated the use of remote learning, but technology allowed for the possibility. It will be interesting to see what other potential configurations arise in the future.

The studies included in the current meta-analysis had reduced f2f instruction time in favor of asynchronous online activities, a feature of hybrid classes that is often praised as it gives students flexibility and autonomy in scheduling. However, this trend may change as a result of the increase in synchronous online teaching during the COVID-19 pandemic. Thus, we expect that hybrid instruction may start featuring more synchronous online activities, replacing or supplementing asynchronous online activities that are currently used in hybrid language courses. Therefore, we recommend future research to consider and report whether the online activities of hybrid courses are synchronous or asynchronous.

Finally, we would like to encourage all researchers to share their data (e.g., raw data, measurement tools, instructional materials) and consider uploading their materials to a central database, such as the IRIS database (iris-database.org). It is by taking this additional step that researchers can increase transparency and, thereby, contribute to replication and meta-analyses efforts in applied linguistics.

NOTE

[1] Used in this case to refer to researching any language acquired beyond one's first without differentiating among first, second, third, and so on, or between L2s and FLs.

Open Research Badges

This article has earned Open Materials badge. Materials are available at https://www.iris-database.org.

REFERENCES

Adair–Hauck, B., Willingham–McLain, L., & Youngs, B. E. (2000). Evaluating the integration of technology and second language learning. *CALICO, 17*, 269–306.

Allen, I. E., & Seaman, J. (2013). Changing course: Ten years of tracking online education in the United States. Accessed 25 June 2021 at http://www.onlinelearningsurvey.com/reports/changingcourse.pdf

Blair, E., & Blair, J. (2015). *Applied survey sampling*. Thousand Oaks, CA: Sage

Chenoweth, A., & Murday, K. (2003). Measuring student learning in an online French course. *CALICO, 20*, 285–314.

Chenoweth, A., & Ushida, E., & Murday, K. (2006). Student learning in hybrid French and Spanish courses: An overview of language online. *CALICO, 24*, 115–145.

Christison, M. A., & Murray, D. E. (2020). *What English language teachers need to know volume III: Designing curriculum* (2nd ed.). London: Routledge.

Dixon, T., & Christison, M. A. (2021). Teaching English grammar in a hybrid academic ESL course: A mixed methods study. In J. Perren, K. Kelch, J.–S. Byun, S. Cervantes, & S. Safavi (Eds.), *Applications of CALL theory in ESL and EFL environments* (2nd ed., pp. 149–169). Hershey, PA: IGI Global.

Echávez–Solano, N. (2003). *A comparison of student outcomes and attitudes in technology-enhanced vs. traditional second-semester Spanish language courses*. (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.

Education Data. (2020). *Online education statistics*. Accessed 25 June 2021 at https://educationdata.org/online-education-statistics/

Geta, M., & Olango, M. (2016). The impact of blended learning in developing students' writing skills: Hawassa University in focus. *African Educational Research Journal*, *4*, 49–68.

Graham, C. R. (2005). Blended learning systems: Definition, current trends, and future directions. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 3–18). San Francisco: Pfeiffer Publishing.

Green, A., & Youngs, B. E. (2001). Using the web in elementary French and German courses: Quantitative and qualitative study results. *CALICO*, *19*, 89–123.

Grgurović, M., Chapelle, C., & Shelley, M. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL*, *25*, 165–198.

Guillén, G. (2014). *Expanding the language classroom: Linguistic gains and learning opportunities through e-tandems and social networks*. (Unpublished doctoral dissertation). University of California Davis, Davis, CA.

JASP Team. (2020). *JASP* (Version 0.13.1) [Computer software].

Kraemer, A. N. (2008). *Engaging the foreign language learner: Using hybrid instruction to bridge the language-literature gap*. (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI.

Larson–Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*, 127–159.

Mahmud, M. M. (2018). Technology and language—what works and what does not: A meta-analysis of blended learning research. *The Journal of Asia TEFL*, *15*, 365–382.

Mizumoto, A. (2015). *Langtest (Version 1.0)*. http://langtest.jp/shiny/kappa/

Mizumoto, A., & Plonsky, L. (2016). R as a lingua franca: Advantages of using R for quantitative research in applied linguistics. *Applied Linguistics*, *37*, 284–291.

Murray, D. E., & Christison, M. A. (2017). Going online: Affordances and limitations for teachers and teacher educators. In K. Hyland & L. Wong (Eds.), *Faces of English education: Students, teachers and pedagogy*. New York: Routledge.

Murray, D. E., & Christison, M. A. (2018). *Online language teacher education: Participants' experiences and perspectives*. Monterey, CA: The International Research Foundation (TIRF). Accessed 25 June 2021 at https://www.tirfonline.org/wp-content/uploads/2017/03/TIRF_OLTE_2017_Report_Final.pdf

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

O'Rourke, S. (n.d.). Hybrid learning vs. blended learning: What's the difference? Ring Central blog. Accessed 25 June 2021 at https://www.ringcentral.com/us/en/blog/hybrid-learning-vs-blended-learning-whats-the-difference/

Plonsky, L. (2015). Statistical power, *p* values, descriptive statistics, and effect sizes: A "back-to-basics" approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 23–45). New York: Routledge.

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

Plonsky, L., & Zhuang, J. (2019). A meta-analysis of second language pragmatics instruction. In N. Taguchi (Ed.), *Routledge handbook of SLA and pragmatics* (pp. 287–307). New York: Routledge.

Plonsky, L., & Ziegler, N. (2016). The CALL–SLA interface: Insights from a second-order synthesis. *Language Learning & Technology*, *20*, 17–37.

Rubio, F. (2014). The effects of blended learning on second language fluency and proficiency. In F. Rubio & J. J. Thoms (Eds.), *Hybrid language learning and teaching: Exploring theoretical, pedagogical and curricular issues* (pp. 137–159). Boston, MA: Cengage/Heinle.

Rubio, F., & Thoms, J. J. (Eds.). (2014). *Hybrid language teaching and learning: Exploring theoretical, pedagogical and curricular issues*. Boston, MA: Cengage/Heinle.

Scida, J., & Jones, J. N. (2016). New tools, new designs: A study of a redesigned hybrid Spanish program. *CALICO*, *33*, 174–200.

Scida, E., & Saury, R. (2006). Hybrid courses and their impact on student and classroom performance: A case study at the University of Virginia. *CALICO*, *23*, 517–531.

Simpson, O. (2012). *Supporting students for success in online and distance education* (3rd ed.). New York: Routledge.

Thoms, J. J. (2014). Analyzing linguistics outcomes of second language learners: Hybrid versus traditional course contexts. In F. Rubio & J. J. Thoms (Eds.), *Hybrid language learning and teaching: Exploring theoretical, pedagogical and curricular issues* (pp. 177–195). Boston, MA: Cengage/Heinle.

Thoms, J. J. (2020). Re-envisioning L2 hybrid and online courses as digital open learning and teaching environments: Responding to a changing world. *Second Language Research & Practice*, *1*, 86–98.

UNESCO. (2020). *COVID-19 impact on education*. Accessed 25 June 2021 at https://en.unesco.org/covid19/educationresponse

Wichadee, S. (2013). Facilitating students' learning with hybrid instruction: A comparison among four learning styles. *Electronic Journal of Research in Educational Psychology*, *11*, 99–116.

Winke, P., Zhang, X., Rubio, F., Gass, S., Soneson, D., & Hacking, J. (2020). The proficiency profile of language students: Implications for programs. *Second Language Research & Practice*, *1*, 25–64.

World Economic Forum. (2020). *The world economic forum COVID action platform*. Accessed 25 June 2021 at https://www.weforum.org/agenda/2020/03/3-ways-coronavirus-is-reshaping-education-and-what-changes-might-be-here-to-stay/

Yang, J., Yin, C. X., & Wang, W. (2018). Flipping the classroom in teaching Chinese as a foreign language. *Language Learning & Technology, 22*, 16–26.

Young, D. J. (2008). An empirical investigation of the effects of blended learning on student outcomes in a redesigned intensive Spanish course. *CALICO, 26*, 160–181.

Zhang, W., & Zhu, C. (2018). Impact of blended learning on university students' achievement of English as a second language. *International Journal on E-Learning, 17*, 251–273.

APPENDIX A

List of Studies Included in the Meta-Analysis

1. Adair–Hauck, B., Willingham–McLain, L., & Youngs, B. E. (2000). Evaluating the integration of technology and second language learning. *CALICO, 17*, 269–306.
2. Dixon, T., & Christison, M. A. (2021). Teaching English grammar in a hybrid academic ESL course: A mixed methods study. In J. Perren, K. Kelch, J.–S. Byun, S. Cervantes, & S. Safavi (Eds.), *Applications of CALL theory in ESL and EFL environments* (2nd ed., pp. 149–169). Hershey, PA: IGI Global.
3. Echávez–Solano, N. (2003). *A comparison of student outcomes and attitudes in technology-enhanced vs. traditional second-semester Spanish language courses.* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.
4. Geta, M., & Olango, M. (2016). The impact of blended learning in developing students' writing skills: Hawassa University in focus. *African Educational Research Journal, 4*, 49–68.
5. Green, A., & Youngs, B. E. (2001). Using the web in elementary French and German courses: Quantitative and qualitative study results. *CALICO, 19*, 89–123.
6. Guillén, G. (2014). *Expanding the language classroom: Linguistic gains and learning opportunities through e-tandems and social networks.* (Unpublished doctoral dissertation). University of California Davis, Davis, CA.
7. Rubio, F. (2014). The effects of blended learning on second language fluency and proficiency. In F. Rubio & J. J. Thoms (Eds.), *Hybrid language learning and teaching: Exploring theoretical, pedagogical and curricular issues* (pp. 137–159). Boston, MA: Cengage/Heinle.
8. Scida, J., & Jones, J. N. (2016). New tools, new designs: A study of a redesigned hybrid Spanish program. *CALICO, 33*, 174–200.
9. Wichadee, S. (2013). Facilitating students' learning with hybrid instruction: A comparison among four learning styles. *Electronic Journal of Research in Educational Psychology, 11*, 99–116.
10. Yang, J., Yin, C. X., & Wang, W. (2018). Flipping the classroom in teaching Chinese as a foreign language. *Language Learning & Technology, 22*, 16–26.
11. Young, D. J. (2008). An empirical investigation of the effects of blended learning on student outcomes in a redesigned intensive Spanish course. *CALICO, 26*, 160–181.

APPENDIX B

A Breakdown of Interrater Reliability

Feature Coded	Percentage Agreement	Cohen's kappa
LMS used	100.0	1.00
Textbook provided online activities	81.8	.75
Is it the first time the course is taught in a hybrid format?	90.0	.84
Percentage of reduction in f2f time	100.0	1.00
Target language	100.0	1.00
Setting (SL vs. FL)	100.0	1.00
Setting (higher education vs. K–12)	100.0	1.00
An element of randomness incorporated into sampling	81.8	.75
Dependent variable (e.g., vocabulary, writing, listening)	88.6	.87
Instrument reliability reporting	82.9	.65
Assumptions checked for inferential statistics	100.0	1.00
Equivalence of groups established at the beginning of treatment	94.3	.87
Study design (within-group vs. between-groups)	97.1	.93
Control group pretest *N*	100.0	1.00
Control group pretest *M*	100.0	1.00
Control group pretest *SD*	100.0	1.00
Control group posttest *N*	100.0	1.00
Control group posttest *M*	100.0	1.00
Control group posttest *SD*	100.0	1.00
Treatment or within-group pretest *N*	100.0	1.00
Treatment or within-group pretest *M*	100.0	1.00
Treatment or within-group pretest *SD*	100.0	1.00
Treatment or within-group posttest *N*	100.0	1.00
Treatment or within-group posttest *M*	100.0	1.00
Treatment or within-group posttest *SD*	100.0	1.00
Effect size reporting	100.0	1.00
Average	**96.8**	**.95**

Note. LMS = learning management system; f2f = face-to-face; FL = foreign language; SL = second language; K–12 = primary and secondary education; *M* = mean; *SD* = standard deviation.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.