This is a pre-publication version of a manuscript that has been accepted at *Language Learning and Technology* (accepted: 15 June 2020).
This version is not the final version and is subject to change.

# Dialogue systems for language learning: a meta-analysis

Serge Bibauw
Universidad Central del Ecuador; KU Leuven; UCLouvain

Thomas François
CENTAL, UCLouvain

Wim Van den Noortgate, Piet Desmet
ITEC, an imec research group at KU Leuven

The present study offers a meta-analysis of effectiveness studies on dialogue-based CALL, i.e., systems allowing a learner to practice a second language (L2) by interacting with a conversational agent ("bot"). Through a systematic inclusion and exclusion process, screening 419 publications, we identified 17 relevant meta-analysable studies. We made use of Morris and DeShon's (2002) formulas to compute comparable effect sizes across designs, including $k = 100$ individual effect sizes, which were analysed through a multilevel random-effects model.
Results confirm that dialogue-based CALL practice has a significant medium effect size on L2 proficiency development ($d = 0.59$). We performed extensive moderator analyses to explore the relative effectiveness on several learning outcomes of different types and features of dialogue-based CALL (type of interaction, modality, agent embodiment, etc.). In particular, our study confirms the effectiveness of form-focused and goal-oriented systems, system-guided interactions, corrective feedback provision, and gamification features. Significant effects for lower proficiency learners, and on vocabulary, morphosyntax, holistic proficiency and accuracy are established. Finally, we discuss evolutions to be expected in dialogue-based CALL and the language learning opportunities it offers.

## Introduction

The central aim of this study is to evaluate the effectiveness of dialogue-based computer-assisted language learning (CALL) for second or foreign language (L2) learning. Dialogue-based CALL encompasses all applications allowing one to practice an L2 through written or spoken conversational interactions with an automated agent, be it a voice-only virtual assistant, a computer-controlled character or a physical robot. Recently, with the increased prevalence of chatbots and digital personal assistants, a renewed attention has been brought to the use of similar dialogue systems for language learning purposes, and commercial applications are being developed (for instance, *Duolingo Bots* was released in 2016). Yet, beyond the hype, the question remains: how effective are these systems for learning a foreign language? The purpose of the present research is to establish whether and to what extent the use of such dialogue-based CALL applications has an impact on the development of learners' L2 proficiency, as it is commonly assumed by the proponents of these systems, and which instructional and study design characteristics moderate the size of the effect. We attempt to address these questions through a meta-analysis of existing effectiveness studies about these applications.

## Dialogue-based CALL

Many names have been given to systems implementing dialogic interactions with an automated agent for language learning purposes: *intelligent tutoring systems*, *conversational agents*, *dialogue systems*, *chatbots*, etc. We gather under the term *dialogue-based CALL* all efforts to make a learner of a foreign language have a dialogue—i.e., a sequence of conversational turns—with any sort of automated agent (chatbot, robot, embodied agent, speech interface, non-player character in a virtual world, etc.) as a language learning task, be it written or spoken. This definition sets dialogue-based CALL apart from other types of

language learning technology. First, interactions occur as part of a meaningful conversational context, rather than as isolated items in many tutorial CALL activities. Second, the interlocutor is the *system*, rather than another human as in computer-mediated communication (CMC). And third, the dialogue *is* the L2 task, not a means of providing scaffolding (pedagogical agent) or instruction in the learner's native language (tutorial dialogue).

A general assumption behind many of these systems is that the meaning-oriented practice of an L2 contributes to the development of the learner's proficiency in that language and that, even though a native speaker would be the ideal conversation partner, an automated agent can provide such practice in contexts where expert speakers are scarce (Sydorenko, Daurio, & L. Thorne, 2018). The idea finds a theoretical foundation in the interactionist approach of second language acquisition: through the dialogue, learners receive input, feedback, and opportunities for output, negotiation of meaning and noticing, which are all essential for L2 development (Ellis & Bogart, 2007). While not all dialogue-based CALL systems provide corrective feedback or complete negotiation of meaning, they all provide input, output and various forms of interactional feedback (Basiron, 2008).

Empirical effectiveness studies on CMC—text-based chat in particular—have already demonstrated that similar interactions with humans have significant effects on language learning outcomes (Lin, 2015a). In certain conditions, they might even have a higher impact on L2 speaking proficiency than face-to-face interactions (Ziegler, 2016). We hypothesise that well-designed dialogue-based CALL systems could provide learning opportunities comparable to CMC. Besides, these systems offer a few advantages over their human counterparts: permanent availability, infinite patience when needing to repeat or to correct, and potential for systematic adaptivity to the learner. They also offer a low-anxiety environment for language practice, which could raise learners' willingness-to-communicate (Fryer & Carpenter, 2006).

Bibauw, François, and Desmet (2019) have proposed a conceptual framework for dialogue-based CALL. Dialogue systems are generally categorised into *task-oriented*—in which the user has a certain goal he wants to achieve through the dialogue (booking a hotel, setting an appointment, etc.)—, and *open-ended systems*—where the conversation has no explicit purpose and looks more like small talk. Beyond this general distinction, we proposed a typology distinguishing four types of dialogue-based CALL systems, presented in Table 1. Some of these systems have been empirically evaluated, but little is known about their comparative effectiveness for L2 development.

## A meta-analysis of experimental research

In the last two decades, researchers have carried out experimental evaluations of the learning effects of dialogue-based CALL. Some of these effectiveness studies brought favourable results (e.g., Harless, Zier, & Duncan, 1999), but other promising studies did not find significant learning gains. This inconclusiveness could be imputable to insufficient statistical power, stemming from methodological decisions such as small sample sizes and short treatment duration (e.g., Hassani, Nahvi, & Ahmadi, 2016), but also to an absence of an effect. Looking at the simple juxtaposition of these studies, in some cases presenting imprecise or conflicting evidence, does not allow one to draw clear conclusions. The very small sample sizes among some of these studies (in one case as low as $n = 6$ per condition) make it particularly difficult to obtain significant findings. With a meta-analysis of these results, however, we could aggregate all experimental evidence to obtain a stronger summary effect size, which would offer a more clear-cut view on the general effectiveness of dialogue-based CALL and on the factors that affect its efficacy.

A meta-analysis is a quantitative synthesis of studies, using statistical methods to aggregate and analyse all the compatible effects measured by these studies (Plonsky & Oswald, 2015). It allows one to establish a more accurate estimate of the effects of a certain intervention, going beyond the statistical significance of results in individual studies.

More importantly, considering the diversity of system features, treatment characteristics and methodological choices in dialogue-based CALL studies, a meta-analysis allows us to perform moderator analyses, i.e., comparisons of effects between groups of studies, defined according to certain variables (e.g., task-oriented versus open-ended systems), comparisons that are not made in individual studies. Meta-

Table 1
*Typology of dialogue-based CALL systems with examples.*

| | Narrative system | Form-focused system | Goal-oriented system | Reactive system |
|---|---|---|---|---|
| Constraints | User has to choose from a list of pre-set utterances with different meanings. | Meaning is pre-set (e.g., gap-filling) or constrained (e.g., questions with given answers). | Meaning influenced by set context and tasks. | Open-ended, free dialogue (chatbots). |
| Interaction | System-guided (branching paths) | System-guided | Interactive, less predictable | User-directed |
| Example | CandleTalk (T.-L. Chiu, Liou, & Yeh, 2007) | CALL-SLT (Bouillon, Rayner, Tsourakis, & Qinglu, 2011) | Wilske, 2015 | CSIEC (Jia, 2009) |
| Dialogue excerpt | User (U) is playing a loud student. Their room-mate (S) is complaining.<br>**S:** Excuse me; have you noticed how loud it is in here?<br>**U:** [*choose from list of sentences and pronounce it*]<br>– What? What sound? I didn't hear anything.<br>– Pardon me; what did you say?<br>– Oh, I'm sorry. I was concentrating on the game so I didn't notice. Did I bother you? (...) | At a restaurant. (...)<br>[*Instruction in L1*] Ask-check-politely<br>**U:** [*free oral input*] I would like the check please.<br>[*Feedback on pronunciation and grammar*] | Someone (S) stops you and asks you for directions. [Map with route provided] (...)<br>**U:** [*free written input*] Turn left, in front of the coffee-shop.<br>[*Corrective feedback if erroneous*]<br>**S:** Okay, left in front of the coffee-shop, and then? | User is free to ask or say anything. System reacts to each last message.<br>**U:** [*free written input*] Hello, I am Peter.<br>**S:** Hi Peter. How are you? (...)<br>**U:** I feel very happy to be a student.<br>**S:** I'm a college student and my major is math. What is your major? |

analyses have therefore the potential to inform practice on how to set up effective dialogue-based CALL systems, and to inform research on promising tracks and understudied questions.

In comparison with other meta-analyses in applied linguistics, we propose a few methodological advances. One is to use a common (raw) effect size metric for within-group and between-group effects, allowing easier comparison through experimental designs. Another one is the use of multilevel modelling to include multiple effect sizes from single studies, with their respective covariates and characteristics (Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013). Our methodological procedures are fully detailed in the following section and appendices, including our full data set and processing script in R.

## Research questions

The research questions that have guided this meta-analysis are:

RQ1 How effective is dialogue-based CALL in general for L2 development?

RQ2 How do different implementations of dialogue-based CALL, distinguished by characteristics of instructional and system design, compare to each other in terms of effectiveness on diverse language learning outcomes?

## Methodology

### Data collection and selection

We followed a systematic and reproducible data collection procedure, summarised in Figure 1. The first step was a database search in major scientific databases, with a search query associating keywords for dialogue systems and language learning (search syntax in Appendix A) in January 2018. It was completed by an auxiliary manual collection strategy through ancestry search (references mentioned in the previously found publications) and forward citations (new publications citing the previously found ones). This resulted, after pruning duplicates, in a total pool of 419 records.[1]

After screening publications for availability, the remaining 386 articles underwent a full-text review based on the definition of dialogue-based CALL given above: 250 papers were kept. This excluded studies that, although using bots for second language learning, only explored scaffolding interaction, typically in L1 (e.g., Arispe, 2014), which we considered outside the scope of dialogue-based CALL. Finally, we retained only the publications presenting empirical effectiveness studies, i.e., quantitative studies reporting measurements of the effects of a dialogue-based application on a certain outcome variable. This final effectiveness corpus totals 39 papers.

As many publications report various outcome variables or measurements, possibly on different samples of participants, each reported series of measurement was recorded as a separate effect size. Considering our intention to perform moderator analyses, we opted to maximise granularity by including the smallest possible aggregation levels for effect sizes. We identified $k = 138$ individual effect sizes mentioned in the publications.

Only one article explicitly reported effect sizes (S. Lee et al., 2011); for all the others, it was necessary to compute them based on disclosed summary statistics. For this reason, we could not include studies not reporting means, standard deviations (*SD*), or alternate summary or test statistics. We contacted the authors to obtain the missing data, but had limited success, despite warmly appreciated answers from most. We also had to exclude effects from a between-subjects study whose alternate condition did not match our control condition (no treatment) and lacked other reference data (no pretest) (Wang & Johnson, 2008). Finally, because our meta-analysis focuses on the effects on L2 development, we excluded six publications measuring other outcome variables, such as motivation. In the end, we analysed $k = 100$ effect sizes, corresponding to 17 publications, 17 dialogue-based CALL systems (sometimes variations of the same system), and 11 research teams.

### Coding

Each of the articles and effects sizes were further analysed and coded according to an extensive coding scheme including publication, system, treatment, population and outcome categories of variables. The variables and their possible values or definitions are presented in Table 2, and the coding process is described in Appendix A.

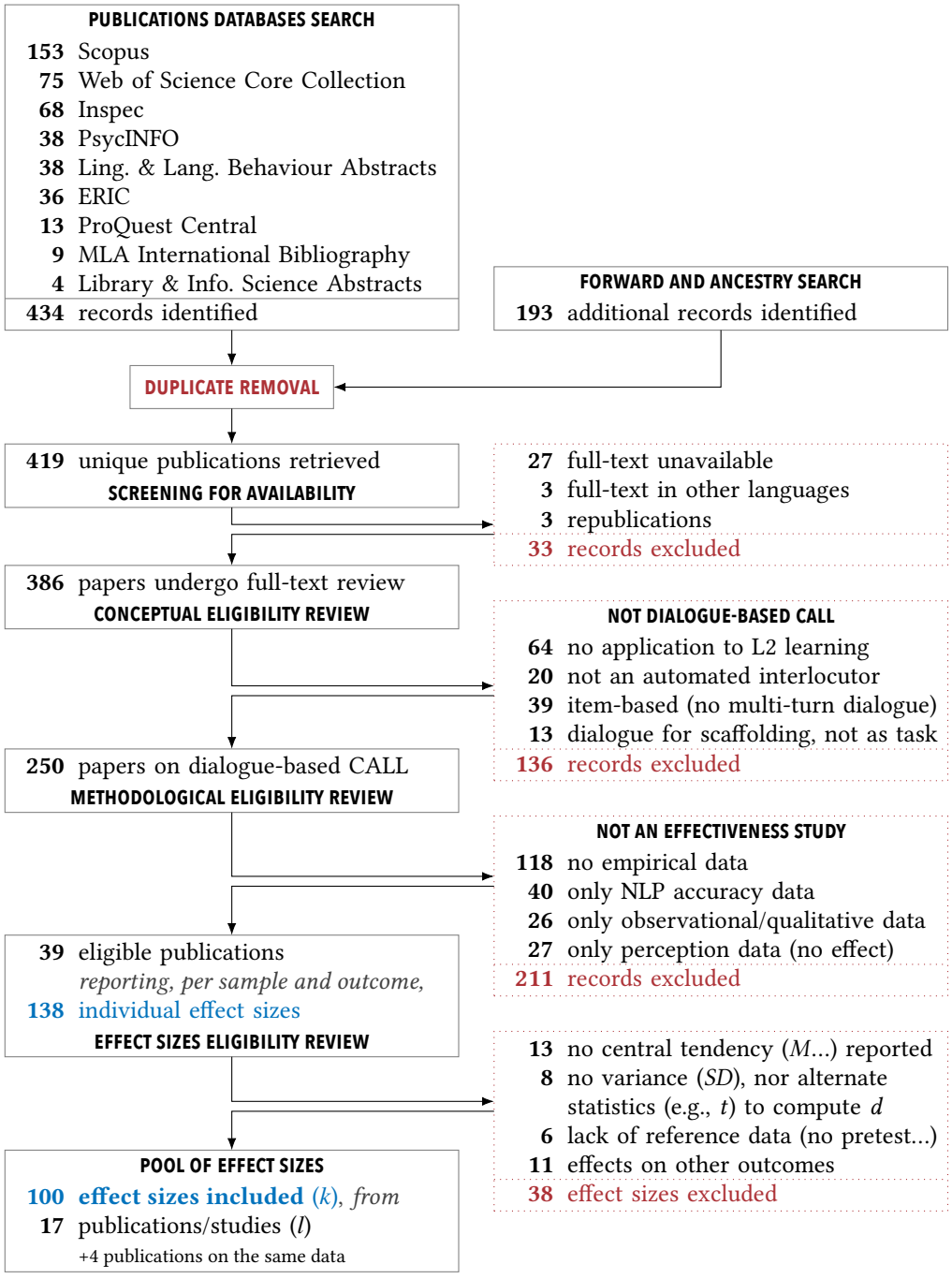
The coding was performed independently by two coders, including the first author, on all studies and effects. The intercoder agreement was computed for all variables as Cohen's kappa, or Krippendorff's alpha for continuous variables and polytomous categorical variables. There was full agreement ($\kappa = 1$) for variables such as age and context. However, the agreement was initially approaching chance level for variables that required a lot of inferencing work, such as time on task and treatment span, because few publications disclose them in an explicit or standardised manner. In such cases, disagreements were subsequently resolved among the two coders, by returning to the original study to reach an agreement, and sometimes by iteratively refining the coding scheme.

### Effect sizes calculation

At the core of a meta-analysis is an aggregation and comparison of individual effects, measured quantitatively. In second language acquisition and CALL research, considering the prevalence of experimental designs, many meta-analyses use Cohen's $d$ or Hedges' $g$, which standardise a difference of means by dividing it by the pooled standard deviation (Plonsky & Oswald, 2015). However, these measures are meant for

[1] The list of considered publications is provided in the Supporting Information.

*Figure 1.* Flowchart of the inclusion and exclusion process of studies and effects.




independent-groups (IG) design, i.e., in studies comparing posttest results from an experimental and a control group. They are not suitable for expressing the *within-group* effect in single-group pretest-posttest design (repeated measures, RM), which requires a formula of standardised mean *change*. Still, other measures may be required for independent-groups pretest-posttest (IGRM) designs that combine features of IG and RM designs.

Morris and DeShon (2002) offer formulas for calculating effect sizes for these designs and for converting them to make them immediately comparable. Therefore, there is no reason to present two different summary effect sizes, one for between-group and an-

Table 2
*Coding scheme for studies*

| Type | Variable | Possible values |
|---|---|---|
| Publication | Publication type | Journal article / Conference paper / Book chapter / Doctoral dissertation |
| Experiment | Experimental design | Independent groups (IG) / Repeated measures (RM) / Repeated measures in independent groups (IGRM) |
| | Group assignment | Random / Intact groups (only for IG/IGRM designs) |
| | Treatment sessions | (number of spaced sessions on the system) |
| | Treatment span | (number of weeks between first and last sessions) |
| | Time on task | (number of hours of usage of the system) |
| | Treatment density | Spaced (span > 1 week) / Packed (≤ 1 week) |
| System | Type of interaction | Task-oriented / Open-ended / System-guided |
| | Type of system | Form-focused system / Goal-oriented system / Reactive system / Narrative system |
| | Meaning constraints | None < Implicit < Explicit < Pre-set |
| | Corrective feedback | None < Implicit < Explicit |
| | Primary modality | Spoken / Written |
| Population | L1 | Chinese / English / Farsi / Korean / Spanish / ... / Mixed |
| | Target language | Arabic / Chinese / English / French / German / ... |
| | L2 proficiency | A1 < A2 < B1 < B2 |
| | Age group | 6-11 < 12-17 < 18+ |
| | Age mean | (if reported; otherwise extrapolated from given range) |
| | Context | School / University / Laboratory |
| Outcome | Outcome type | Production // Comprehension // Knowledge test |
| | Outcome variable | Proficiency / Accuracy / Complexity / Fluency // Listening / Reading // Grammar / Vocabulary |
| | Type of instrument | Meta-linguistic judgment / Selected response / Constrained response / Free response |
| | Outcome modality | Spoken / Written |
| | Outcome temporality | Short-term (immediate) / Long-term (delayed posttest) |

other for within-group effects, as it is common meta-analytic practice in language learning and CALL (Plonsky & Oswald, 2014): they can be transformed into a comparable metric, aggregated together, and thus offer a stronger estimate of the true effect.

In our pool of effect sizes, 92 studies follow an RM design and 8 follow an IGRM design; no IG design is represented. To compute a comparable effect size across study designs, we used the normalised *raw* metric ($d_{IG}$) proposed by Morris and DeShon (2002), which is aligned on the between-group effect that Cohen's $d$ measures (see Appendix A for discussion of raw and change metrics). We used their formulas to compute $d_{IG}$ for the RM and IGRM studies present in our data set, and applied Hedges' correction factor $J$ for small sample bias (Hedges & Olkin, 1985, chap. 5, eq. 7).[2] We use $d$ hereafter as the general notation of this standardised mean difference. For the RM design, the mean change ($M_{post}-M_{pre}$) is normalised by the standard deviation of the pretest scores ($SD_{pre}$), which is considered to be more consistent across studies:

$$d = J(df_{RM})\left(\frac{M_{post} - M_{pre}}{SD_{pre}}\right) \quad (1)$$

For IGRM design, the standardized change in the control group (C) is subtracted from the change in the experimental group (E):

$$d = J(df_{IGRM})\left(\frac{M_{post,E} - M_{pre,E}}{SD_{pre,E}} - \frac{M_{post,C} - M_{pre,C}}{SD_{pre,C}}\right) \quad (2)$$

## Multiple effect sizes and multilevel modelling

The computing of an overall effect requires that the meta-analyst decides on a statistical modelling approach. A *fixed-effects model* assumes that

[2] $J(df)$ is based on the degrees of freedom of the design, calculated from the sub-sample sizes ($n$) in each study as $df_{RM} = n_E - 1$ and $df_{IGRM} = n_E + n_C - 2$.

all effect sizes are estimates of a constant true effect of "dialogue-based CALL on L2 development" and that the observed variation can only be accounted to *within-study* sampling variance. However, most recent meta-analyses do not make this assumption and use a *random-effects model*. This model assumes that, beyond sampling variance, studies have been observing different population effects, due to different study designs and characteristics, and takes this additional *between-studies* variation into account.

Traditional meta-analytic fixed-effects and random-effects techniques are meant to aggregate independent effect sizes estimates. However, in our pool of studies as elsewhere, the analysed publications rarely report only one effect size: they may report effects from distinct instances of a system, on samples from populations with distinct characteristics, and often through multiple tests and outcome measurements. These multiple effect sizes from the same publication cannot be considered independent, as they share certain sources of random variation, such as specificities of the population sampled from, a specific experimental procedure, or certain tendencies in rating non-objective tests. Various solutions have been used to avoid this dependency, usually involving selecting or averaging dependent effect sizes, with the drawback of losing part of the information they convey (Plonsky, 2011).

To avoid the problem of dependence and the loss of information or power, we opted for a multilevel meta-analytic modelling, as described by Van den Noortgate et al. (2013). Whereas a fixed-effects model assumes effect sizes vary only on one level (within studies, due to sampling), and a traditional random-effects model assumes that effect sizes can vary on two levels (at the sampling level and at the study level), the multilevel approach adds a third, intermediate layer of potentially unexplained variation: within a single study, several population effect sizes may be estimated. The information that effect sizes from the same study share (e.g., they usually evaluate the same system with similar sampling, testing and rating procedures) is still taken into account at the third, between-studies level. Table 3 summarises the three layers of aggregation of the model, with their respective number of units.

The major advantage of using a multilevel model is that it allows one to include as many fine-grained effect sizes as possible from the original studies. For instance, Wilske (2015) reports 20 distinct effect sizes, studying various versions of a system with multiple outcome variables and tests. By adding each effect size individually, we maintain the comparative information between a form-focused and an unconstrained input system, with or without corrective feedback, on written accuracy or speaking fluency, etc. This information is particularly valuable for our moderator analyses, but it would have been lost if combined into a single per-study effect.

Table 3
*Levels of multilevel meta-analytic model*

| Level | Number of elements | Source of variance |
|---|---|---|
| 1 Samples | $k_1 = 100$ ($N = 803$) | Random sampling variance |
| 2 Effects sizes | $k_2 = 100$ | Within-study variation (e.g., varying effect measurements) |
| 3 Studies | $k_3 = 17$ | Between-studies variation (e.g., varying systems, populations, designs) |

$k$ = number of effect sizes at levels 1 (sampling variance), 2 (within-study effects) and 3 (number of individual studies/publications). $N$ = total number of unique individuals tested in the various samples.

The multilevel models, with or without moderator variables, were fitted with the *metafor* package (Viechtbauer, 2010), in *R*, using the `rma.mv()` function for multilevel modelling and the restricted maximum-likelihood (`REML`) method.

## Results

As detailed previously, after the inclusion and exclusion process, we retrieved 17 publications, reporting 100 effect sizes on a total of 803 participants. Figure 2 presents a forest plot of the effects for each of the 17 studies. A complete list of all individual effects, with corresponding variables, can be found in Appendix B.

*Figure 2.* Forest plot of study-level effect sizes. $k$ = number of within-study effect sizes.[3]

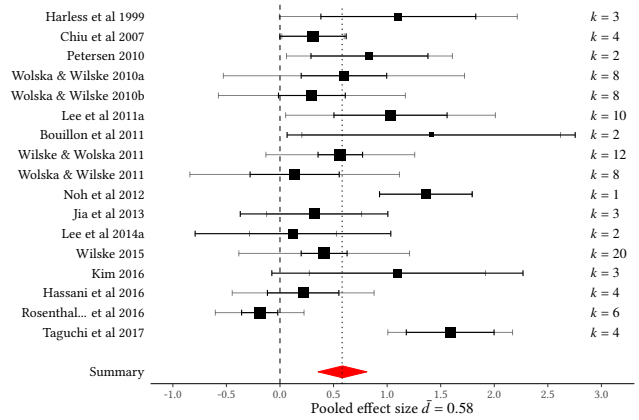
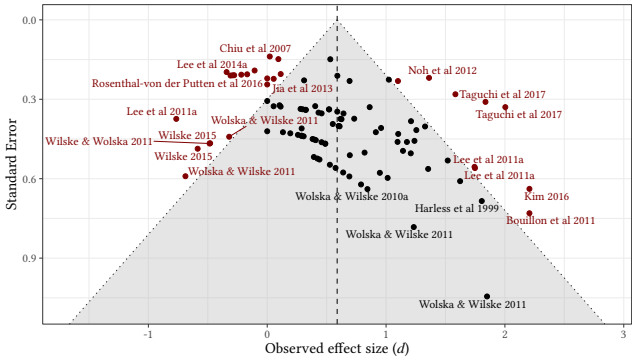



*Figure 3.* Funnel plot of effect sizes against study precision




## Overall effect

The summary effect established by the three-level random-effects model for all studies is $d = 0.59$, with a 95% confidence interval of $[0.35, 0.82]$. It confirms that dialogue-based CALL has globally a highly significant *medium* effect on L2 development ($p < .001$).

## Heterogeneity and publication bias

It is important to note that the observed outcomes vary substantially across studies. A $Q$-test for heterogeneity (Higgins & Green, 2008) confirms that there is substantial residual heterogeneity in the effect sizes at the second and third levels, $Q(df = 99) = 311.1$, $p < .001$, $I^2 = 68.02\%$. The variance is relatively higher between studies ($k_3 = 17$, $\sigma_3^2 = 0.18$), indicating potentially multiple true effects of dialogue-based CALL, than within studies ($k_2 = 100$, $\sigma_2^2 = 0.08$), with also a high sampling variance ($k_1 = 100$, $N = 803$, $Md(\sigma_1^2) = 0.17$), possibly imputable to less precise outcome measurements.

It thus seems clear that we are in the presence of different types of pedagogical interventions, with varying degrees of effectiveness on different outcomes and target groups. This supports our decision to use a random-effects model and especially incentivises moderator analyses, to be able to disentangle the covariates of the observed effects and potential subgroups that can cause these varying effects.

***Publication bias.*** The funnel plot in Figure 3 reveals a potential publication bias, considering the absence of strong negative effects in the lower-left side of the triangle: it is reasonable to assume that highly negative effects in underpowered studies might not have been reported. However, the sample size is not a significant moderator of the effect ($b = 0.00$, 95% CI $[-0.01, 0.01]$, $p = .497$) and including it does not improve the model fit, thus eliminating the possibility that more precise studies could bring less favourable results.

## Moderator analyses

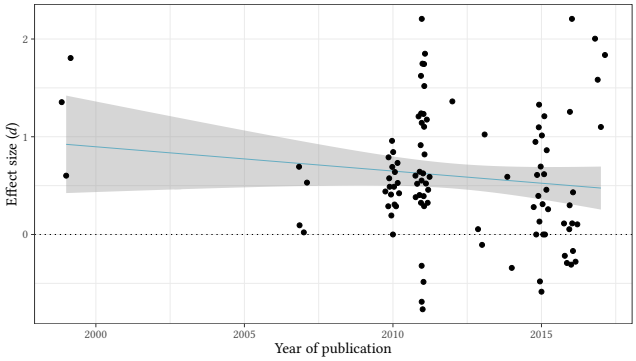
As stated in our second research question, the ambition of this meta-analysis of dialogue-based CALL is also to get insights into the conditions under which the approach produces better outcomes. In particular, we will review the moderator effect of publication and experimental design variables, target population variables, system characteristics, and outcome measurement variables.

We control for the significance of the differences between moderators by reporting $Q$-tests, which are equivalent to ANOVA $F$-tests on categorical variables. For categorical variables, we report the estimated mean effect size ($d$), which includes the intercept, for each possible value. For continuous variables, we report the regression weight ($b$) from the meta-regression model, i.e., how each additional unit influences the effect. Nevertheless, these multiple tests are not meant to confirm a pre-established hypothesis, and should mostly be interpreted as exploratory.[4]

[4]Given this exploratory purpose, we do not apply Bonferroni corrections for multiple comparisons.

**Publication and experiment moderators.** The *type of publication* does not make a significant difference, even if the mean effect of journal articles in our sample tends to be higher than conference papers and doctoral dissertations. This tendency can be explained by field traditions: most conference papers are authored by specialists in natural language processing (NLP) rather than applied linguistics, with stronger technical evaluation procedures (e.g., recognition rate) and only peripheral effectiveness evaluations, whose instruments might not always bring the necessary power to reveal learning effects. There also does not appear to be any *chronological evolution* of observed effects across studies, as shown in Figure 4.

*Figure 4.* Effect sizes against year of publication.




Regarding the *experimental design*, we obtain very similar effect size estimates for within (RM) and within-and-between (IGRM) designs, as presented in Table 4. It seems that the slightly stronger bias to which the within-group design is susceptible does not heavily affect the results.

The *treatment length* deserves special attention. As any pedagogical intervention, the effect of dialogue-based CALL practice is a function of the time the participants spent using the system, and the way it was distributed. Looking at each treatment duration variable in isolation, none achieve significance, probably because of inconsistencies in reporting and accounting for these variables, but, in the analysed studies, the number of sessions ($b = 0.02$) and the time on task ($b = 0.02$) did influence the outcome, while the total span of the experiment did not ($b = 0.00$). Counter-intuitively perhaps, studies using packed practice, operationalised as an intervention lasting for 1 week or less, seemed to present higher outcomes ($d = 0.97$) than the ones using spaced practice ($d = 0.53$).

**Population moderators.** The *L2 proficiency level* of the learners seems to have some influence on the learning gains from dialogue-based CALL. While not reaching significance level, probably due in part to an unequal distribution of studies across levels (with very few conducted on beginners and advanced learners), the moderator, when considered as simply categorical, shows a downward trend from studies involving A1 learners ($d = 0.68$) to studies involving B2 learners ($d = -0.33$). This downward trend is also visible if we use proficiency level as a continuous variable with a linear effect ($b = -0.33$), in which case the moderator is closer to significance, $Q(1) = 3.24$, $p = .072$. This phenomenon can probably in part be explained by the increasing cost of learning gains along with the increase in proficiency.

By looking at each level in isolation, as presented in Table 5, it arises that the most noticeable effects seem to be observed on beginner (A1) and lower-intermediate (A2) learners, while the average effect on upper-intermediate (B1) fails to pass the significance threshold and the gain for advanced learners (B2) could very well be non-existent.

*Age* does not seem to influence the results. While it could be due to the limited scope of the included studies, and the large majority of them being on adults, effect sizes for the three age groups are relatively similar and using the mean age of the sample leads to equally non-significant and low effects. Similarly, experiments conducted in school and higher education *contexts* present indistinguishable effects. The few "laboratory" studies conducted in isolated contexts, however, report weaker effects.

**System moderators.** In terms of *type of interaction*, the system-guided and task-oriented types produce significant effect sizes, with a potentially stronger effect for the former. The *type of dialogue-based CALL system* is not a significantly differential moderator, as shown in Table 6, but certain types present on their own results significantly different from a null effect: form-focused systems and goal-oriented systems. Even though the small number of narrative systems does not allow us to establish their effects, it is interesting to note that the low effect estimate ($d = 0.31$) is coherent with the fact that these systems offer limited opportunities for productive practice, and would

Table 4
*Moderator analyses for experiment variables, including ANOVA-like Q-test of moderators and estimated effect size for each level.*

| Variable | *df* | *Q* | *p* | Values | *k* | *d*/*b* | *SE* | 95% CI |
|---|---|---|---|---|---|---|---|---|
| Type of publication | 2 | 1.69 | .431 | Conference paper | 31 | 0.36 | 0.22 | [−0.07, 0.79] |
| | | | | Dissertation | 22 | 0.58 | 0.34 | [−0.09, 1.24] |
| | | | | Journal article*** | 47 | 0.70 | 0.15 | [0.40, 1.00] |
| Experimental design | 1 | 0.04 | .836 | IGRM* | 8 | 0.65 | 0.30 | [0.06, 1.24] |
| | | | | RM*** | 92 | 0.58 | 0.13 | [0.32, 0.84] |
| Group assignment | 1 | 1.78 | .182 | Intact groups | 3 | 0.32 | 0.34 | [−0.36, 0.99] |
| | | | | Random** | 5 | 0.94 | 0.31 | [0.32, 1.55] |
| Treatm. distribution | 1 | 2.61 | .106 | Packed*** | 11 | 0.97 | 0.24 | [0.49, 1.44] |
| | | | | Spaced*** | 81 | 0.53 | 0.12 | [0.31, 0.76] |
| Duration (weeks) | 1 | 0.02 | .896 | +1 week (*b*) | 92 | 0.00 | 0.03 | [−0.06, 0.06] |
| Sessions | 1 | 0.84 | .361 | +1 session (*b*) | 100 | 0.02 | 0.02 | [−0.02, 0.07] |
| Time on task | 1 | 1.02 | .314 | +1 hour (*b*) | 100 | 0.02 | 0.02 | [−0.02, 0.05] |

*df* = degrees of freedom. *Q* = statistic from *Q*-test for moderator effect. *p* = significance of the *Q*-test. *k* = number of effect sizes corresponding to each value. *d* = mean effect size when including only effects matching this moderator value. *b* = regression weight (relative effect size increment for every increment of 1 unit in the moderator). *SE* = standard error of *d* or *b*. CI = 95% confidence interval of *d* or *b*. ***significant effect at $p < .001$. **significant effect at $p < .01$. *significant effect at $p < .05$.

Table 5
*Moderator analyses for population variables.*

| Variable | *df* | *Q* | *p* | Values | *k* | *d*/*b* | *SE* | 95% CI |
|---|---|---|---|---|---|---|---|---|
| L2 proficiency | 3 | 3.74 | .443 | A1* | 15 | 0.68 | 0.33 | [0.03, 1.32] |
| | | | | A2** | 93 | 0.70 | 0.25 | [0.22, 1.18] |
| | | | | B1 | 83 | 0.36 | 0.35 | [−0.33, 1.05] |
| | | | | B2 | 15 | −0.33 | 0.41 | [−1.12, 0.47] |
| Age group | 2 | 0.44 | .802 | 6–11* | 13 | 0.77 | 0.30 | [0.18, 1.36] |
| | | | | 12–17 | 5 | 0.54 | 0.37 | [−0.19, 1.27] |
| | | | | 18+*** | 82 | 0.56 | 0.15 | [0.27, 0.84] |
| Age (mean) | 1 | 2.05 | .152 | +1 year (*b*) | 97 | −0.03 | 0.02 | [−0.07, 0.01] |
| Context | 2 | 0.69 | .707 | Laboratory | 9 | 0.33 | 0.36 | [−0.37, 1.03] |
| | | | | School ** | 18 | 0.68 | 0.23 | [ 0.23, 1.13] |
| | | | | University *** | 73 | 0.60 | 0.16 | [ 0.29, 0.91] |

therefore have a lower effect on proficiency. The effect of reactive systems remains to be demonstrated, as it fails to reach significance. Similarly, the type of *meaning constraints* imposed on learner production could affect the effectiveness, with the most constrained type, when the meaning to express is pre-set and the learner only has liberty to modify the form of her utterances, presenting the strongest effect sizes.

The absence or presence of *corrective feedback* in the system does not make a statistically significant difference in terms of effect, but the size of the observed effects of each type (with feedback: $d = 0.70$; without feedback: $d = 0.38$) fits the well-documented positive impact of corrective feedback on learning. The effects of implicit and explicit types of feedback are, on the other hand, very similar, with a slight potential advantage for explicit ones.

**Outcome moderators.** There is a significant difference between the effects on the *type of learning outcome*, as shown in Table 7: dialogue-based CALL seems to have the highest impact on production outcomes and knowledge tests, while there was no significant effect on comprehension outcomes—but it may also be due to their under-representation ($k = 4$). It does however seem logical that active practice in dialogue-based CALL has a higher impact on productive skills. When considering a finer classification of outcomes in terms of L2 proficiency *dimension* being tested, there still is a significant difference, with most notable effects on lexical development, holistic proficiency and accuracy in production.

The type of *testing instrument* also significantly influences the results. It seems that tests asking for constrained and free constructed responses (hence, more open-ended), and meta-linguistic judgement, are more sensitive to the effects than selected responses (which, again, is consistent with the mentioned focus on production).

The *modality* (oral vs. written) of the system and the modality of the test have, per se, no significant influence on the effect size. It is impressive how written and spoken systems have statistically identical effects in this data set. But interestingly, while the

Table 6
*Moderator analyses for system variables.*

| Variable | *df* | *Q* | *p* | Values | *k* | *d* | *SE* | 95% CI |
|---|---|---|---|---|---|---|---|---|
| Type of interaction | 2 | 2.39 | .303 | System-guided*** | 11 | 0.97 | 0.27 | [0.43, 1.51] |
| | | | | Open-ended | 6 | 0.57 | 0.34 | [−0.10, 1.24] |
| | | | | Task-oriented*** | 83 | 0.50 | 0.14 | [0.23, 0.76] |
| Type of system | 3 | 1.49 | .685 | Narrative | 4 | 0.31 | 0.49 | [−0.65, 1.27] |
| | | | | Form-focused** | 15 | 0.87 | 0.27 | [0.33, 1.40] |
| | | | | Goal-oriented** | 75 | 0.53 | 0.16 | [0.21, 0.85] |
| | | | | Reactive | 6 | 0.57 | 0.37 | [−0.16, 1.30] |
| Meaning constraints | 3 | 6.93 | .074 | None | 6 | 0.56 | 0.30 | [−0.03, 1.15] |
| | | | | Implicit*** | 75 | 0.52 | 0.13 | [0.27, 0.78] |
| | | | | Explicit* | 15 | 0.44 | 0.22 | [0.01, 0.86] |
| | | | | Pre-set*** | 4 | 1.59 | 0.40 | [0.80, 2.37] |
| Modality | 1 | 0.00 | 1.000 | Spoken*** | 35 | 0.59 | 0.17 | [0.25, 0.93] |
| | | | | Written*** | 65 | 0.59 | 0.17 | [0.25, 0.93] |
| Corrective feedback | 2 | 2.08 | .354 | No* | 23 | 0.38 | 0.19 | [0.01, 0.75] |
| | | | | Implicit*** | 39 | 0.68 | 0.15 | [0.38, 0.98] |
| | | | | Explicit*** | 38 | 0.73 | 0.16 | [0.42, 1.05] |
| Embodied agent | 1 | 0.97 | .325 | No*** | 83 | 0.53 | 0.13 | [0.28, 0.78] |
| | | | | Yes*** | 17 | 0.73 | 0.19 | [0.37, 1.10] |
| Gamification* | 1 | 4.93 | .026 | No*** | 83 | 0.45 | 0.12 | [0.22, 0.68] |
| | | | | Yes*** | 17 | 0.99 | 0.21 | [0.57, 1.41] |

Table 7
*Moderator analyses for outcome variables.*

| Variable | *df* | *Q* | *p* | Values | *k* | *d* | *SE* | 95% CI |
|---|---|---|---|---|---|---|---|---|
| Outcome type** | 2 | 13.43 | .001 | Knowledge test*** | 36 | 0.65 | 0.16 | [0.34, 0.97] |
| | | | | Comprehension | 4 | −0.49 | 0.33 | [−1.14, 0.16] |
| | | | | Production*** | 60 | 0.67 | 0.15 | [0.37, 0.97] |
| Outcome dimension** | 7 | 18.52 | .010 | Grammar* | 21 | 0.50 | 0.20 | [0.11, 0.90] |
| | | | | Vocabulary** | 15 | 0.84 | 0.27 | [0.32, 1.36] |
| | | | | Reading | 1 | 0.63 | 0.71 | [−0.77, 2.03] |
| | | | | Listening | 3 | −0.58 | 0.35 | [−1.28, 0.11] |
| | | | | Complexity | 2 | 0.82 | 0.47 | [−0.11, 1.75] |
| | | | | Accuracy** | 28 | 0.60 | 0.18 | [0.24, 0.96] |
| | | | | Fluency | 18 | 0.39 | 0.22 | [−0.04, 0.81] |
| | | | | Holistic proficiency** | 12 | 0.73 | 0.25 | [0.25, 1.22] |
| Test type* | 3 | 10.23 | .017 | Constrained resp.*** | 32 | 0.85 | 0.19 | [0.48, 1.22] |
| | | | | Free response*** | 38 | 0.66 | 0.19 | [0.29, 1.02] |
| | | | | Metaling. judgment*** | 20 | 0.74 | 0.21 | [0.33, 1.15] |
| | | | | Selected response | 10 | 0.08 | 0.24 | [−0.38, 0.54] |
| Test modality | 1 | 0.03 | .872 | Spoken*** | 39 | 0.57 | 0.15 | [0.28, 0.87] |
| | | | | Written*** | 61 | 0.60 | 0.13 | [0.34, 0.86] |
| Matching modality** | 1 | 7.91 | .005 | No | 28 | 0.27 | 0.17 | [−0.06, 0.59] |
| | | | | Yes*** | 72 | 0.71 | 0.13 | [0.46, 0.96] |
| Temporality | 1 | 0.06 | .805 | Short-term*** | 77 | 0.59 | 0.12 | [0.36, 0.83] |
| | | | | Long-term*** | 23 | 0.56 | 0.16 | [0.24, 0.88] |

modalities themselves do not matter, their interaction, i.e., the fact that the test targets the same modality as the one practised in the system, had a significant influence on the effects: studies with matching modality between system and test had more than twice the effect size of the others. This fact provides insights on the question of transfer of ability across modalities: while some transfer of gains could be happening from written practice into speaking skills ($d$ = 0.29, 95% CI [−0.21, 0.79]) and, vice versa, from oral practice to writing ($d$ = 0.19, [−0.31, 0.70]), it looks like this transfer is quite limited in comparison with skill practice and acquisition in the same modality, either writing ($d$ = 0.65, [0.27, 1.04]) or speaking ($d$ = 0.84, [0.42, 1.26]).

In terms of *temporality of effects*, we do not observe in this meta-analysis a clear difference between the effects on immediate posttests ($d$ = 0.59, [0.36, 0.83]) and delayed posttests ($d$ = 0.56, [0.24, 0.88]), which could indicate that the effects of dialogue-based CALL practice are generally well sustained in the long-term.

## Discussion and conclusion

This meta-analysis is, to the best of our knowledge, the first to summarise the effectiveness of dialogue-based CALL systems, including dialogue systems, chatbots, and conversational agents, on L2 proficiency development. Methodologically, in applied linguistics, it is also one of the first meta-analyses to use a multilevel modelling approach to allow the inclusion of multiple effect sizes per study and the use of effect size conversion formulas to use a single metric across research designs. These methodological innovations allowed us to draw more insight into the relative effectiveness of dialogue systems for language learning, which we summarise and discuss hereafter.

### How effective is dialogue-based CALL?

The results obtained from this multilevel meta-analysis indicate that dialogue-based CALL has a significant, medium effect of $d_{IG}$ = 0.59, [0.35, 0.82], on L2 proficiency development, if expressed in between-subjects metric. This overall effect is comparable to what some other meta-analyses have obtained for CALL interventions, such as $d$ = 0.53 for game-based learning (Y.-h. Chiu, Kao, & Reynolds, 2012) or $d$ = 0.44 for CMC (Lin, 2015a). It is however smaller than the averaged effect size ($d$ = 0.84) calculated by Plonsky and Ziegler (2016) in their second-order synthesis of CALL.

When comparing this overall effect with estimates of the effects of other forms of interaction for language learning, dialogue-based CALL compares fairly well and maintains reasonable outcomes. For instance, Mackey and Goo (2007) evaluated the overall effect of interaction at $d$ = 0.75. It also stands within the range of observed effects of text-based chat on L2 proficiency measured by various meta-analyses ($d$ = 0.44 in Lin, 2015b, $d$ = 1.13 in Ziegler, 2016). It is however quite low in comparison to effects of interactional interventions on more focused grammatical and lexical acquisition, measured via knowledge tests rather than on L2 performance, as synthesised by Keck, Iberri-Shea, Tracy-Ventura, and Wa-Mbaleka (2006) at $d_{RM}$ = 1.17.

In the current state of technology, dialogue systems can do their best to emulate interactions with human interlocutors and possibly systematise certain features such as corrective feedback, but there are still many shortcomings preventing them from being entirely up to the task. Such applications are thus made to compensate for a lack of real interactional opportunities, not to replace them, and can only hope to achieve close-enough effectiveness.

### How do different implementations of dialogue-based CALL compare in terms of effectiveness?

The results of our moderator analyses should be interpreted at a very different degree of evidence in contrast with the previous conclusions regarding overall effectiveness, as most moderators do not achieve significance in $Q$-tests. The relative immaturity of the field, with still few effectiveness studies and most being done on small samples, does not allow us to draw firm conclusions here. Most of our observations here are strictly exploratory and should be regarded only as hinting at new hypotheses that remain to be tested.

In general, this meta-analysis provides supportive evidence for the claim that "the differences between human-computer interaction and human-only interaction do not bring about vastly different conditions for language learning" (Wilske, 2015, p. 244). In this sense,

moderators which are known to affect the effectiveness of traditional forms of L2 interaction, such as corrective feedback (as previously demonstrated by Petersen, 2010; Wilske, 2015), treatment length or sessions spacing, seem to behave similarly in dialogue-based CALL.

**Which systems perform best?** In terms of general architecture of dialogue-based CALL systems, form-focused systems ($d$ = 0.87, [0.33, 1.40]), as in Taguchi, Li, and Tang (2017) and Harless et al. (1999), and goal-oriented systems ($d$ = 0.56, [0.21, 0.85]), such as *POMY* (Noh et al., 2012), both achieve significant effects on their own, while effects from reactive ($d$ = 0.57, [−0.16, 1.30]) and narrative systems ($d$ = 0.31, [−0.65, 1.27]) are unclear due to the limited number of effectiveness studies for these two types. In any case, it is not confirmed yet whether dialogue-based CALL effectiveness would follow the distinction observed by Y.-h. Chiu et al. (2012) in their meta-analysis of games for language learning, that meaningful and engaging applications might have a much stronger effect than drill-and-practice ones.

Focusing on the interactional design of the dialogue management, system-guided interactions present potentially the strongest effect ($d$ = 0.97, [0.43, 1.51]), before task-oriented interactions ($d$ = 0.50, [0.23, 0.76]). It is in line with the effects of the type of meaning constraints on learner production, with fixed meaning producing very high effect sizes ($d$ = 1.59, [0.80, 2.37]). This seems to favour system-guided, i.e., highly constrained, less interactive dialogues, to the detriment of much more complex task-oriented interactions. A possible explanation is that, because the technological cost and the unpredictability of system-guided interactions are low, more attention can be dedicated to conversation design, complexity adaptation and progressive introduction of target structures. In other words, trading off technological design for instructional design could be beneficial. If this difference were to be confirmed, it could discourage the development of complex dialogue management systems in favour of more constrained scripted dialogues. However, this might also be due to system-guided interactions being typically used in form-focused contexts, which assess learning on narrower and more achievable outcomes.

Regarding particular instructional features, corrective feedback seems to allow for higher learning gains, although not significantly in this meta-analysis ($b$ = 0.33, [−0.14, 0.79], $p$ = .169). There was no visible difference here between implicit (recasts, mostly) and explicit forms of feedback. These results are in line with previous evidence on the effects of corrective feedback in SLA (S. Li, 2010) and confirm conclusions from Petersen (2010) and Wilske (2015) that corrective feedback in human-computer interactions could "be as effective at promoting L2 development as in an oral, dyadic context" (Petersen, 2010, p. 188). The lower relative effect of feedback here (in comparison with $d$ = 0.64 in S. Li, 2010 meta-analysis) can probably be understood through the fact that, in dialogue-based CALL, even in the absence of corrective feedback, there are always multiple forms of interactional and communicative feedback, through the agent's responses, and thus the control condition is not the same as in SLA feedback studies. On the other hand, it is striking that the overall effect of dialogue-based *with* corrective feedback ($d$ = 0.70) is even closer to the mean effects of CALL or L2 interaction encountered in the above-mentioned meta-analyses (Mackey & Goo, 2007; Plonsky & Ziegler, 2016), as these interventions typically do include feedback.

Dialogue-based CALL applications that used some form of gamification had a significantly stronger impact on L2 development ($b$ = 0.54, [0.06, 1.01], $p$ = .026). These results advocate for the integration of game-based elements and for motivational considerations in the design of future dialogue systems.

On the other hand, the embodiment of the agent in the learning environment, as a virtual avatar or a physical robot, did not bring about significant changes in comparison with speech-only interfaces, even though the included studies that used agents with a visible representation had slightly higher effects ($b$ = 0.20, [−0.20, 0.60], $p$ = .325). The lack of significant difference is in line with the results of Rosenthal-von der Pütten, Straßmann, and Krämer (2016), who did not find any effect of the type of embodiment, not even on perception of the system by the participants, but contradictory to the review of J. Li (2015), which concluded that the physical presence of robots led to improved user perception and performance.

**For whom is it most effective?** In the past, some studies have reported tendencies towards higher effectiveness of dialogue-based CALL for low to mod-

erate proficiency (Kaplan, Sabol, Wisher, & Seidel, 1998) or low-achieving learners (Huang, Lin, Yang, & Wu, 2008), while others have hypothesised that, because of the possible communication breaks and lack of adaptivity in open-ended systems, it might be more adequate for advanced learners (Fryer & Carpenter, 2006). This meta-analysis tends to support the idea that the learning gains might diminish for higher proficiency users, although the evidence for confirming this claim is still insufficient. Our results could not verify statistically significant effects for B1 learners, but more strikingly, there are no signs of positive learning gains at all for advanced learners (B2). In contrast, the positive effects on beginner and lower-intermediate proficiency learners (A1 and A2) are established by the moderator analysis. We hypothesise that the meaningful practice of the target language allowed by dialogue-based CALL is especially fruitful in the consolidation stages of the learning process, when some explicit linguistic knowledge foundations have been laid but production skills, in particular spoken exchanges with other speakers, might still be hindered by L2 anxiety and lack of practice.

There does not seem to be any significant effect of age on the effectiveness of these systems, and there is no significant difference between school, university and lab-based experiments. These results accord with observations of Jia (2009), who found no difference across age or educational context, and with what has been corroborated about CALL interventions in general (Grgurović, Chapelle, & Shelley, 2013).

**For what is it most effective?** What language learning outcomes are best impacted by dialogue-based CALL practice? Generally, the main research claim on intelligent tutors, as summarised by Golonka, Bowles, Frank, Richardson, and Freynik (2014, p. 89), that "learners demonstrate pretest-posttest gains in different areas, including speaking, reading comprehension, vocabulary, grammar, fluency" holds with an updated and more quantitative evaluation of empirical evidence. More precisely, statistically significant effect sizes are established for vocabulary and morphosyntactic outcomes in knowledge tests, and on holistic proficiency and accuracy measures on production. Effects on fluency could be less important, and remain to be demonstrated, and effects on complexity, reading and listening comprehension have been insufficiently studied to advance any clear pattern.

On the question of transfer of learning across modalities, this meta-analysis provides new insights about the quality of this transfer. First, it is noteworthy primarily spoken and primarily written interface systems have virtually identical effect sizes, and that the effects on spoken tests and written tests are extremely close, statistically indistinguishable. But, while modalities of practice and outcome do not seem to matter in isolation, their interaction does make a statistically significant difference: effect sizes increase threefold when practice and test modalities are the same ($b = 0.44$, $p = .005$). While it does not invalidate previous evidence that written practice, in particular in computer-mediated communication, could promote the development of oral proficiency (e.g., Lin, 2015b), as effect size for non-matching modalities is not null, this finding puts into perspective this transfer as possibly partial and not equally effective as practising in the same oral modality (Ziegler, 2016).

### Limitations

This study is not without limitations. As most meta-analyses, despite our rigorous selection process, our data set suffers from biases. The most important limitation here is probably an issue of power: we could only include a small number of independent studies, which themselves have on average very small sample sizes. The total number of participants ($N = 803$) remains relatively low in comparison with other meta-analyses. Therefore it should be emphasized that dialogue-based CALL strongly needs larger experimental studies to test most research questions on the matter.

We tried to avoid a publication bias by not restricting our inclusion process to peer-reviewed publications only. However, apart from the two included dissertations, we could not find unpublished effectiveness data, and it appears, from the funnel plot presented in Figure 3, that some studies could have produced negative effects that the researchers decided not to publish.

This is linked to the fact that nearly all researchers who conducted the effectiveness evaluations were also the designers of each evaluated system or part of the same team. And even for the only one who

evaluated an external system (Kim, 2016, evaluating *Indigo*), as the system was a general-purpose chatbot, the instructions built around it to transform the tool into a pedagogical task were designed by the same researcher. Hence, there is a high risk that any negative or inconclusive findings on the effectiveness of these systems may have been ignored, or have simply not made it to a publication (a very acute publication bias in fact). This is somewhat indicated, in our meta-analysis, by the absence of any clearly negative effect, as can be observed in Figure 3. Obviously, this self-evaluation bias is in great part explained by the relative novelty of the object, and by the extremely limited availability of previously developed systems, that usually remain at the level of internal prototypes and are rarely available to the public (Sydorenko, Smits, Evanini, & Ramanarayanan, 2019).

Finally, the relatively high heterogeneity of the included studies could be regarded as problematic. Although we believe that all these studies share a common rationale and supporting theory—that practising an L2 through dialogue, including with an artificial conversation partner, leads to improvements in the learner's ability to use the language—, and that their heterogeneity is also an opportunity to learn more in details how different variables impact the learning process, the differences between, for instance, form-focused and goal-oriented dialogue systems are important, as is the variation in learning outcomes and testing procedures. Added to the limited number of independent studies, and the even smaller number of research teams (11) represented in our meta-analysis, this fact could lead to strong biases in the moderator analyses.

Because of these shortcomings, our global effect size should be taken with caution, and, as mentioned above, our moderator analyses should only be regarded as exploratory and indicating potential hypotheses to test on new data. More generally, for the advancement of the field, more external effectiveness evaluations of systems, conducted by independent researchers, should be encouraged.

### Maturity of the field and avenues for research

The research domain of dialogue-based CALL is gradually entering a more mature phase, where systematic experiments are conducted to verify the main claims that have been at the foundations of developments in the field since its inception. It is still early, and the number of meta-analysable studies remains limited. In particular, the lack of independent evaluations of systems, i.e., experimental studies conducted by teams independent of their designers, is certainly limiting the strength of claims of usefulness. This fact is intimately connected to the unavailability to the public of previously developed systems, most of which remained at a prototype level (Bibauw et al., 2019).

However, research and industry have recently shown encouraging signs of a change on this matter, with major commercial players, such as Duolingo, ETS and Alelo releasing or planning to release public dialogue-based CALL applications, and incipient collaborations between industry and academia to compare the systems and establish common ground (Sydorenko et al., 2019). Such efforts could open the field both to a large audience of language learners and to many research opportunities.

We can also hope for future technological advances in natural language understanding and dialogue management making their path into dialogue-based CALL systems. To date, dialogue systems have not yet witnessed the breakthroughs that deep learning has brought to other NLP tasks, at least not with the same magnitude (Serban, Lowe, Henderson, Charlin, & Pineau, 2018). While research on dialogue systems is actively pursuing fully data-driven end-to-end approaches, systems used in production tend to opt for rule-based and hybrid approaches, combining ad hoc and handcrafted subsystems to achieve satisfactory results (Harms, Kucherbaev, Bozzon, & Houben, 2019). Currently, these approaches require very intensive manual work and offer limited scalability and adaptability, but hopefully, probabilistic solutions will soon be adaptable for final-user applications.

From the available experimental studies to date, this meta-analysis has demonstrated that, overall, the effectiveness of dialogue-based CALL is comparable to other CALL or instructed SLA interventions, in particular when dialogue systems provide corrective feedback. Future research should thus probably focus more on which affordances and implementations of such systems provide better results, rather than comparing dialogue systems in general to other CALL or

traditional instruction methods. As Chun (2016) reminds us, "a primary research question is not whether technology-based instruction is effective, but rather under what conditions and for whom" (p. 107).

Our moderator analyses have attempted to clear the path for future system design and evaluation by identifying trends and insights hidden in previous studies regarding the relative effectiveness of certain designs and features for defined populations and learning outcomes. While these findings are essentially exploratory, they ask many questions that could be addressed in future investigations. Do relatively free, task-oriented dialogue systems provide better learning opportunities than more constrained, possibly fully scripted, guided interactions? Is it possible, as our findings could suggest, that the major technological complexity and development efforts required by having to manage free user input in the former type do not necessarily lead to increases in learning outcomes and that this development time might be more beneficial if invested in instructional content design? Is the incompleteness of transfer across modalities confirmed? Is there a significant effect on fluency, and is it indeed lower than on other dimensions of proficiency? Many more questions regarding the optimal technological and instructional design choices, the most useful features to implement and the most benefited outcomes and types of learners are still in need of empirical responses. We hope that, in the future, SLA and CALL researchers, NLP and AI developers and language learning content creators will be able to join their efforts to answer them.

## References

**Publications included in the meta-analysis are preceded by an asterisk.*


Arispe, K. (2014). What's in a bot? L2 lexical development mediated through ICALL. *Open Journal of Modern Linguistics*, *4*(1), 150–165. doi:10.4236/ojml.2014.41013

Basiron, H. B. (2008). Corrective feedback in dialogue-based computer assisted language learning. In *Proceedings of NZCSRSC 2008* (pp. 192–195). Christchurch. Retrieved from http://eprints.utem.edu.my/42/

Bibauw, S., François, T., & Desmet, P. (2019). Discussing with a computer to practice a foreign language: Research synthesis and conceptual framework of dialogue-based CALL. *Computer Assisted Language Learning*. doi:10.1080/09588221.2018.1535508

*Bouillon, P., Rayner, E., Tsourakis, N., & Qinglu, Z. (2011). A student-centered evaluation of a web-based spoken translation game. In *Proceedings of SLaTE 2011*, Venice. Retrieved from https://www.isca-speech.org/archive/slate_2011/sl11_033.html

Chiu, Y.-h., Kao, C.-w., & Reynolds, B. L. (2012). The relative effectiveness of digital game-based learning types in English as a foreign language setting: A meta-analysis. *British Journal of Educational Technology*, *43*(4), E104–E107. doi:10.1111/j.1467-8535.2012.01295.x

*Chiu, T.-L., Liou, H.-C., & Yeh, Y. (2007). A study of web-based oral activities enhanced by automatic speech recognition for EFL college learning. *Computer Assisted Language Learning*, *20*(3), 209–233. doi:10.1080/09588220701489374

Chun, D. M. (2016). The role of technology in SLA research. *Language Learning & Technology*, *20*(2), 98–115. doi:10125/44463

Ellis, N. C., & Bogart, P. S. H. (2007). Speech and language technology in education: The perspective from SLA research and practice. In *Proceedings of SLaTE 2007* (pp. 1–8). Farmington: ISCA. Retrieved from https://www.isca-speech.org/archive_open/slate_2007/sle7_001.html

Fryer, L., & Carpenter, R. (2006). Bots as language learning tools. *Language Learning & Technology*, *10*(3), 8–14. doi:10125/44068

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, *27*(1), 70–105. doi:10.1080/09588221.2012.700315

Grgurović, M., Chapelle, C. A., & Shelley, M. C. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL*, *25*(02), 165–198. doi:10.1017/S0958344013000013

*Harless, W. G., Zier, M. A., & Duncan, R. C. (1999). Virtual dialogues with native speakers: The evaluation of an interactive multimedia method. *CALICO Journal*, *16*(3), 313–337. doi:10.1558/cj.v16i3.313-337

Harms, J.-G., Kucherbaev, P., Bozzon, A., & Houben, G.-J. (2019). Approaches for dialog management in conversational agents. *IEEE Internet Computing*, *23*(2), 13–22. doi:10.1109/MIC.2018.2881519

*Hassani, K., Nahvi, A., & Ahmadi, A. (2016). Design and implementation of an intelligent virtual environment for improving speaking and listening skills. *Interactive Learning Environments*, *24*(1), 252–271. doi:10.1080/10494820.2013.846265

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando: Academic Press.

Higgins, J. P. T., & Green, S. (Eds.). (2008). *Cochrane handbook for systematic reviews of interventions*. Chichester: Wiley-Blackwell.

Huang, Y.-T., Lin, Y.-L., Yang, J.-C., & Wu, Y.-C. (2008). An English dialogue companion system for supporting conversation practice. In *Proceedings of ICCE 2008* (pp. 43–48). Taipei: APSCE.

Jia, J. (2009). An AI framework to teach English as a Foreign Language: CSIEC. *AI Magazine*, *30*(2), 59–71. doi:10.1609/aimag.v30i2.2232

Kaplan, J. D., Sabol, M. A., Wisher, R. A., & Seidel, R. J. (1998). The Military Language Tutor (MILT) program: An advanced authoring system. *Computer Assisted Language Learning*, *11*(3), 265–87. doi:10.1076/call.11.3.265.5679

Keck, C., Iberri-Shea, G., Tracy-Ventura, N., & Wa-Mbaleka, S. (2006). Investigating the empirical link between task-based interaction and acquisition: A meta-analysis. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 91–131). Amsterdam: John Benjamins.

*Kim, N.-Y. (2016). Effects of voice chat on EFL learners' speaking ability according to proficiency levels. *Multimedia-Assisted Language Learning*, *19*(4), 63–88. doi:10.15702/mall.2016.19.4.63

*Lee, K., Kweon, S.-o., Lee, S., Noh, H., & Lee, G. G. (2014). POSTECH Immersive English Study (POMY): Dialog-based language learning game. *IEICE Transactions on Information and Systems*, *E97-D*, 1830–1841. doi:10.1587/transinf.E97.D.1830

*Lee, S., Noh, H., Lee, J., Lee, K., Lee, G. G., Sagong, S., & Kim, M. (2011). On the effectiveness of robot-assisted language learning. *ReCALL*, *23*(1), 25–58. doi:10.1017/s0958344010000273

Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, *77*, 23–37. doi:10.1016/j.ijhcs.2015.01.001

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*(2), 309–365. doi:10.1111/j.1467-9922.2010.00561.x

Lin, H. (2015a). A meta-synthesis of empirical research on the effectiveness of computer-mediated communication (CMC) in SLA. *Language Learning & Technology*, *19*(2), 85–117. doi:10125/44419

Lin, H. (2015b). Computer-mediated communication (CMC) in L2 oral proficiency development: A meta-analysis. *ReCALL*, *27*(3), 261–287. doi:10.1017/S095834401400041X

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 407–452). Oxford: Oxford University Press.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105–125. doi:10.1037/1082-989x.7.1.105

*Noh, H., Ryu, S., Lee, D., Lee, K., Lee, C., & Lee, G. G. (2012). An example-based approach to ranking multiple dialog states for flexible dialog management. *IEEE Journal of Selected Topics in Signal Processing*, *6*(8), 943–958. doi:10.1109/JSTSP.2012.2229692

*Petersen, K. A. (2010). *Implicit corrective feedback in computer-guided interaction: Does mode matter?* (Doctoral dissertation, Georgetown University, Washington). doi:10822/553155

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning*, *61*(4), 993–1038. doi:10.1111/j.1467-9922.2011.00663.x

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. doi:10.1111/lang.12079

Plonsky, L., & Oswald, F. L. (2015). Meta-analysing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106–128). New York: Routledge.

Plonsky, L., & Ziegler, N. (2016). The CALL-SLA interface: Insights from a second-order synthesis. *Language Learning & Technology*, *20*(2), 17–37. doi:10125/44459

*Rosenthal-von der Pütten, A. M., Straßmann, C., & Krämer, N. C. (2016). Robots or agents – Neither helps you more or less during second language acquisition. In D. Traum, W. Swartout, P. Khooshabeh, S. Kopp, S. Scherer, & A. Leuski (Eds.), *Intelligent Virtual Agents* (pp. 256–268). doi:10.1007/978-3-319-47665-0_23

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., & Pineau, J. (2018). A survey of available corpora for building data-driven dialogue systems. *Dialogue & Discourse*, *9*(1), 1–49. Retrieved from http://dad.uni-bielefeld.de/index.php/dad/article/view/3690

Sydorenko, T., Daurio, P., & L. Thorne, S. (2018). Refining pragmatically-appropriate oral communication via computer-simulated conversations. *Computer Assisted Language Learning*, *31*(1-2), 157–180. doi:10.1080/09588221.2017.1394326

Sydorenko, T., Smits, T. F. H., Evanini, K., & Ramanarayanan, V. (2019). Simulated speaking environments for language learning: Insights from three cases. *Computer Assisted Language Learning*, *32*(1-2), 17–48. doi:10.1080/09588221.2018.1466811

*Taguchi, N., Li, Q., & Tang, X. (2017). Learning Chinese formulaic expressions in a scenario-based interactive environment. *Foreign Language Annals*, *50*(4), 641–660. doi:10.1111/flan.12292

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*, 576–594. doi:10.3758/s13428-012-0261-6

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. doi:10.18637/jss.v036.i03

Wang, N., & Johnson, W. L. (2008). The politeness effect in an intelligent foreign language tutoring system. In B. P. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of ITS 2008* (pp. 270–280). doi:10.1007/978-3-540-69132-7_31

*Wilske, S. (2015). *Form and meaning in dialog-based computer-assisted language learning* (Doctoral dissertation, Universität des Saarlandes, Saarbrücken). doi:10.22028/D291-23654

*Wilske, S., & Wolska, M. (2011). Meaning versus form in computer-assisted task-based language learning: A case study on the German dative. *Journal for Language Technology and Computational Linguistics*, *26*(1), 23–37. Retrieved from https://jlcl.org/content/2-allissues/13-Heft1-2011/3.pdf

*Wolska, M., & Wilske, S. (2010). Form-focused task-oriented dialogues for computer assisted language learning: A pilot study on german dative. In *Proceedings of L2WS 2010* (P1–6). Retrieved from https://www.isca-speech.org/archive/L2WS_2010/papers/lw10_P1-6.pdf

*Wolska, M., & Wilske, S. (2011). Assessing the effect of type-written form-focused dialogues on spoken language fluency. In *Proceedings of SLaTE 2011* (pp. 57–60). Venice: ISCA. Retrieved from https://dblp.org/rec/bib/conf/slte/2011

Ziegler, N. (2016). Synchronous computer-mediated communication and interaction: A meta-analysis. *Studies in Second Language Acquisition*, *38*(3), 553–586. doi:10.1017/S027226311500025X

## Appendix A
## Methods: supplementary information

### Data collection and selection

The database search was conducted on Scopus, Web of Science Core Collection, INSPEC, PsycINFO, Linguistics & Language Behaviour Abstracts (LLBA), ERIC, ProQuest Central, MLA International Bibliography and Library and Information Science Abstracts (LISA), with the following search syntax, applied on title, abstract and keywords indexes:

```
( chatbot* OR "chat bot*" OR chatterbot* OR
"conversational agent*" OR "conversational
companion*" OR "conversational system*" OR
"dialogue system*" OR "dialogue agent*" OR
"dialogue game" OR "human-computer dialogue"
OR "dialogue-based" OR "pedagogical agent*"
OR "dialog system*" OR "dialog agent*" OR
"dialog game" OR "human-computer dialog" OR
"dialog-based" ) AND ( "language learning"
OR "language teaching" OR "language
acquisition" OR "English learning" OR
"English teaching" OR "English acquisition"
OR "second language" OR "foreign language"
OR L2 OR EFL OR ESL OR ICALL )
```

Figure 1 summarises the in- and exclusion criteria and process. Screening the titles and abstracts of the corpus of publications, we excluded records whose full-text was unavailable, as well as a few republications and publications in languages we could not understand (it was only the case for two publications in Chinese and one in Korean; it is clear however that a search process on English keywords, on essentially English-speaking scientific databases, was unlikely to return more publications in foreign languages).

### Coding

***Experimental variables.*** The *experimental design*—single-group pretest-posttest (within-group design), independent-groups (between-groups design) or independent-groups pretest-posttest (between- and within-group design)—and, if so, the *group assignment* procedure—random or not (intact groups)—were coded. As we wanted to establish the effectiveness of dialogue-based CALL per se, we only considered as between-groups studies those where the comparison group was a "true" control group, in an either passive (no treatment) or "business-as-usual" condition. If it was not the case (i.e., alternate intervention as control), we only included the within-group effects for the experimental group. When studies compared various types of dialogue systems (e.g., Rosenthal-von der Pütten et al., 2016; Wilske & Wolska, 2011), we included the effects for each type as autonomous within-group effects.

Considering the importance of treatment duration on its effectiveness, we attempted to code it as precisely as possible considering the reporting in the original publications and distinguished three duration variables: *treatment sessions*, as the number of sessions participants practised with the system on separate days, *treatment span* as the number of weeks between the first and the last session of the intervention, and *time on task*, as the time (in hours) effectively spent by the subjects using the dialogue-based CALL application. Considering the spacing effect and the benefits of distributed practice, it seems relevant to distinguish the number of presentations from the total number of weeks that the treatment lasted. All this information was however rarely reported in a systematic manner in publications, and it was frequently necessary to compute, with some degree of interpretation, one or more of these variables from the reported procedure. Besides, while it is the *actual* duration of the treatment that matters for its effectiveness, some studies only report the number of hours the learners were instructed to practice autonomously, which could be an overestimation of their actual time on task.

The context of the experiment was also categorised broadly as school, university or laboratory (including other non-educational settings, such as military research).

***Population variables.*** We categorised the general L2 proficiency of the participants according to the levels of the Common European Framework of Reference, interpreting "beginner" as A1, "lower-intermediate" as A2, "upper-intermediate" as B1, and "advanced" as B2.[5] As most samples of participants included mixed levels, we coded all represented levels as

[5] As these proficiency levels correspond to the level at the beginning of the intervention, it seems difficult to plan a short-term effectiveness study on students that already have a C level of proficiency, as the room of improvement would be very narrow.

equally present in the study, using proportions[6]. Variables such as the native language (L1) of the learners were also coded, but have too many levels to allow for any meaningful comparison within the limited set of studies.

Considering the importance of the age factor on language learning, we coded the age of participants both as an *age group*, distinguishing children (6-11), adolescents (12-17) and adults (18+), and as the *mean age* (when it was reported), as a more accurate continuous estimate.

***System variables.*** The variables describing the interactional, instructional and technological characteristics of each studied system followed the dialogue-based CALL analysis framework from Bibauw et al. (2019). In particular, variables such as the *type of interaction*, the *type of system*, the *type of form-* and *meaning constraints*, the *modality of user input*—written or spoken—and the presence and type of *corrective feedback* provide the most determining classification of systems. More fine-grained distinctions, such as the type of dialogue (7 types), are however insufficiently represented in this pool of effectiveness studies to allow for valid analyses.

***Outcome variables.*** We coded the characteristics of the dependent variables measured in the original studies by first identifying the *type of outcome* (knowledge test, comprehension task or production task), and as a categorisation, the measured L2 *proficiency dimension*—grammar or vocabulary for knowledge tests; listening or reading for comprehension tasks; holistic proficiency, complexity, accuracy or fluency for production tasks—, their *modality* and the *type of response* demanded by the testing instruments—reusing the categorisation into meta-linguistic judgement, selected response, constrained response and free response proposed by Norris and Ortega (2000).While this categorisation could be discussed, it reached suitable levels of intercoder agreement, at $\kappa$ = .84 for type of outcome and $\kappa$ = .78 for proficiency dimension.

In the case of repeated measures, when studies included more than two time points for outcome measurement (e.g., Wolska & Wilske, 2010), we used the first test before the treatment as pretest, the last one immediately after treatment as posttest, ignoring intermediate measurements. Results from delayed posttests were included as another effect size with a differing temporality of testing (long-term effects versus short-term for immediate posttests).

When, within a study, various versions of a system or treatment were evaluated (e.g., Wilske, 2015; Wilske & Wolska, 2011) with constrained versus free production, and explicit versus implicit corrective feedback), or the system was evaluated on various participants with distinct characteristics (e.g., S. Lee et al., 2011, with various proficiency levels), separate effect sizes for these different variations were calculated. In other words, studies comparing alternate conditions on two different samples of participants are treated in this meta-analysis as yielding two separate effects with distinct parameters. The only effects from the selected studies that could not be included were alternate conditions that did not involve an automated agent (e.g., interactions with peer or native speaker) (Kim, 2016; Petersen, 2010). The question of how dialogue-based CALL compares with interactions with humans is certainly of much interest, but the only two cases represented in our data set are not enough to tackle it in a meta-analysis.

## Comparable effect size metrics across designs

Morris and DeShon (2002) offer two metrics: either a *change* metric ($d_{\text{RM}}$), aligned on the within-group effect, or a *raw* metric ($d_{\text{IG}}$), aligned on the between-group effect that Cohen's $d$ measures. Due to the absence in most publications of the necessary data to compute change scores (e.g., standard deviation of gain scores) and the lack of homogeneity in the data—among other things, the fact that the pretest-posttest correlation ($\rho$) is different across studies, and the high variety of outcome measures and treatment procedures, including the important differences between the tested systems—, it seemed more appropriate for this meta-analysis to use a raw score metric for effect sizes. Morris and DeShon (2002) also recommend using the independent-groups metric in most situations, for familiarity and comparability reasons.

[6]For instance, a study including only advanced learners was coded as `A1=0, A2=0, B1=0, B2=1`, while a study including all but advanced learners was coded as `A1=.33, A2=.33, B1=.33, B2=0`.

**Sampling variance of effect sizes.** The objective of the meta-analysis is to aggregate meaningfully the individual studies' effect sizes. In the final model, effect sizes are weighted according to their respective precision, which is defined by the sampling variance ($\sigma^2$) of the effect size estimate ($d$). We used the formulas for sampling variance of raw metric scores provided by Morris and DeShon (2002), for RM design:

$$\sigma^2 = \left[\frac{2\,(1-\rho)}{n_\mathrm{E}}\right]\left(\frac{df_\mathrm{RM}}{df_\mathrm{RM}-2}\right)\left[1+\frac{n_\mathrm{E}}{2\,(1-\rho)}\delta_\mathrm{IG}^2\right]-\frac{\delta_\mathrm{IG}^2}{J(df_\mathrm{RM})^2} \tag{3}$$

and for IGRM design:

$$\sigma^2 = \left(\frac{n_\mathrm{E}+n_\mathrm{c}}{n_\mathrm{E}\cdot n_\mathrm{C}}\right)\left(\frac{df_\mathrm{IGRM}}{df_\mathrm{IGRM}-2}\right)\left(1+\frac{n_\mathrm{E}\cdot n_\mathrm{C}}{n_\mathrm{E}+n_\mathrm{C}}\delta_\mathrm{IG}^2\right)-\frac{\delta_\mathrm{IG}^2}{J(df_\mathrm{IGRM})^2} \tag{4}$$

where $\delta_\mathrm{IG}$ is the population effect size (which is unknown and will be replaced by our estimate $d_\mathrm{IG}$), and $\rho$ is the population correlation coefficient between pretest and posttest scores (replaced by estimate $r$ based on the available samples).

### Estimating undisclosed parameters

Beyond the fact that not all publications report the essential parameters to estimate the effect they observed (typically, $M$, $SD$ and $n$ for each outcome measurement), certain cases require additional parameters, among other things for computing the sampling variance of the effect size, such as the correlation coefficient between pretest and posttest scores ($r$), or the standard deviation of the *gain* between pretest and posttest. In CALL publications, such parameters are extremely rarely reported. In cases where it was possible, we have thus attempted to estimate the necessary parameters from other available data.

**Standard deviation of change scores.** Two publications ($k = 6$) in our corpus (Bouillon, Rayner, Tsourakis, & Qinglu, 2011; T.-L. Chiu, Liou, & Yeh, 2007) report the standard deviation of the pre-post change ($SD_\mathrm{change} = SD(x_\mathrm{post} - x_\mathrm{pre})$). It is missing from all the others ($k = 94$). However, this parameter is an essential information for RM design, as it is required for $t$-statistics and effect size (change metric). It differs substantially from the standard deviation of each test scores and cannot be obtained from them (unlike $M_\mathrm{RM} = M_\mathrm{post} - M_\mathrm{pre}$). Fortunately, for studies which do not report $SD_\mathrm{change}$, it is possible to estimate it from $t$-scores, if one is reported, with the following formula (Morris & DeShon, 2002):

$$SD_\mathrm{change} = \sqrt{\frac{n\left(M_\mathrm{post}-M_\mathrm{pre}\right)^2}{t_\mathrm{RM}^2}} \tag{5}$$

**Pretest-posttest correlation coefficient.** Not a single publication in our data set reports the correlation between pre- and posttest scores, which is used in certain sampling variance formulas. Fortunately, it is possible to estimate it from standard deviations, if $SD_\mathrm{change}$ is known or possible to estimate (Morris & DeShon, 2002):

$$r = \frac{SD_\mathrm{pre}^2 + SD_\mathrm{post}^2 - SD_\mathrm{change}^2}{2\cdot SD_\mathrm{pre}\cdot SD_\mathrm{post}} \tag{6}$$

For studies where no standard deviation of change scores, nor $t$-score, nor correlation were available, following recommendations from Morris and DeShon (2002), we used a global estimate for $r$ based on the average of known correlation coefficients from the other studies. We used a linear regression on several variables likely to influence the pretest-posttest correlation, such as time on task and number of sessions of treatment, but all covariates failed to achieve significance, probably due to the very small number of accurate coefficients available to train the model. We thus used a common estimate of $r$, averaged at the study-level to avoid bias due to an over-representation of non-independent estimates: $\hat{r} = .57$.

### Adaptation of certain results

**Outcome measurements with different polarity.** As effect sizes were computed as the difference $M_\mathrm{post} - M_\mathrm{pre}$, positive values of $d$ indicate a positive learning gain. However, a few effects are reported as negative outcomes, such as number of errors (K. Lee, Kweon, Lee, Noh, & Lee, 2014) or length of pauses (Wolska & Wilske, 2011), so that a positive learning gain would consist in their decrease. After calculating effect sizes for these outcomes, their sign was changed ($d = -d_{negative}$), so that for all effect sizes, a positive effect size refers to a desirable effect.

**Median and interquartile range.** One publication (Wilske, 2015) reports, for a subset of the experiment's measured outcomes, only the median and interquartile range (*IQR*) of the scores. We transformed them, under the assumption of normality, into parametric statistics by assuming that $M \approx median$ and $SD \approx IQR/1.35$ (following Higgins & Green, 2008).

**Outliers.** Our data set has two outliers, i.e., effect sizes that are so high that they diverge fundamentally from the rest of the studies. As the mean and variance of effect sizes are very sensitive to extreme values, outliers can distort the meta-analysis misleadingly. In the case of Kim (2016), it seems particularly difficult to understand how the speaking proficiency of the beginner group jumped practically 5 standard deviations above the control group ($d = 4.70$) in less than 3 hours of exposure, while the intermediate and advanced learners with the same treatment demonstrated much more reasonable learning effects ($d = 1.28$ and $d = 0.11$). The other outlier, from Bouillon et al. (2011), corresponds to very high learning effects ($d = 3.61$) observed on a focused vocabulary test for a small sample ($N = 10$). We thus decided to winsorise these extreme values, following the procedure from Lipsey and Wilson (2001): we established cut-off points at 2 standard deviations below and above the mean effect size, $[-0.94, 2.21]$, and replaced any extreme value of $d$ by these limits. It is noteworthy however that the summary effect obtained *without* winsorizing these outliers is virtually identical ($d = 0.59$) to the one we report in the study.

## Appendix B
## Individual study results

Effect size and main descriptive variables for all $k = 100$ effects measured in the 17 meta-analysed studies. To distinguish multiple within-study effects, we specify, when relevant: *(Population) System/Treatment ⟶ Outcome measure *timing.*

| Study | Weeks | Sess. | Hours | $n_E$ | $n_C$ | $d$ | $SE$ | 95% CI |
|---|---|---|---|---|---|---|---|---|
| Bouillon et al 2011 | | | | | | | | |
| CALL-SLT ⟶ grammar test | N/A | 2 | 3.0 | 10 | | 0.82 | 0.25 | [0.66, 0.98] |
| CALL-SLT ⟶ translation test | N/A | 2 | 3.0 | 10 | | 2.21 | 0.53 | [1.88, 2.54] |
| Chiu et al 2007 | | | | | | | | |
| (EM) CandleTalk ⟶ DCT, compr$^{ty}$ | 5 | 6 | 4.5 | 29 | | 0.02 | 0.02 | [0.01, 0.03] |
| (EM) CandleTalk ⟶ DCT, adeq. | 5 | 6 | 4.5 | 29 | | 0.09 | 0.02 | [0.09, 0.10] |
| (non EM) CandleTalk ⟶ DCT, compr$^{ty}$ | 5 | 6 | 4.5 | 20 | | 0.53 | 0.02 | [0.52, 0.54] |
| (non EM) CandleTalk ⟶ DCT, adeq. | 5 | 6 | 4.5 | 20 | | 0.69 | 0.05 | [0.67, 0.72] |
| Harless et al 1999 | | | | | | | | |
| ⟶ listening compr. | 1 | 4 | 32.0 | 9 | | 0.60 | 0.16 | [0.49, 0.71] |
| ⟶ reading compr. | 1 | 4 | 32.0 | 9 | | 1.35 | 0.32 | [1.14, 1.56] |
| ⟶ speaking prof. | 1 | 4 | 32.0 | 9 | | 1.81 | 0.47 | [1.50, 2.11] |
| Hassani et al 2016 | | | | | | | | |
| ⟶ automated prof. score (LPL) | 1 | 1 | 2.0 | 10 | | 0.43 | 0.13 | [0.35, 0.51] |
| ⟶ grammatical errors/sentence | 1 | 1 | 2.0 | 10 | | 0.11 | 0.11 | [0.04, 0.18] |
| ⟶ nb of proper replies | 1 | 1 | 2.0 | 10 | | 0.30 | 0.12 | [0.23, 0.37] |
| ⟶ phonation time/letter | 1 | 1 | 2.0 | 10 | | 0.05 | 0.11 | [−0.01, 0.12] |
| Jia et al 2013 | | | | | | | | |
| (sample Huiwen JHS) | 9 | 9 | 6.8 | 37 | 34 | 0.05 | 0.05 | [0.04, 0.07] |
| (sample Huojia N1 SHS) | 9 | 9 | 6.8 | 56 | 56 | 1.02 | 0.05 | [1.01, 1.03] |
| (sample Jingxian JHS) | 9 | 9 | 6.8 | 48 | 47 | −0.11 | 0.04 | [−0.11, −0.10] |
| Kim 2016 | | | | | | | | |
| (A2 sample) | 16 | 16 | 2.7 | 20 | 20 | 2.21 | 0.41 | [2.08, 2.33] |
| (B1 sample) | 16 | 16 | 2.7 | 22 | 22 | 1.25 | 0.18 | [1.20, 1.31] |
| (B2 sample) | 16 | 16 | 2.7 | 21 | 16 | 0.10 | 0.11 | [0.07, 0.14] |
| Lee et al 2011a | | | | | | | | |
| (A1) ⟶ hol. comm. ability | 8 | 16 | 5.7 | 10 | | 1.14 | 0.24 | [0.99, 1.29] |
| (A1) ⟶ hol. grammar | 8 | 16 | 5.7 | 10 | | 1.24 | 0.21 | [1.11, 1.37] |
| (A1) ⟶ hol. pronunciation | 8 | 16 | 5.7 | 10 | | 1.62 | 0.37 | [1.39, 1.85] |
| (A1) ⟶ hol. vocabulary | 8 | 16 | 5.7 | 10 | | 1.21 | 0.25 | [1.05, 1.37] |
| (A1) ⟶ listening compr. | 8 | 16 | 5.7 | 10 | | 0.29 | 0.17 | [0.19, 0.39] |
| (A2) ⟶ hol. comm. ability | 8 | 16 | 5.7 | 11 | | 1.74 | 0.31 | [1.56, 1.93] |
| (A2) ⟶ hol. grammar | 8 | 16 | 5.7 | 11 | | 1.18 | 0.21 | [1.05, 1.30] |
| (A2) ⟶ hol. pronunciation | 8 | 16 | 5.7 | 11 | | 1.75 | 0.31 | [1.56, 1.93] |
| (A2) ⟶ hol. vocabulary | 8 | 16 | 5.7 | 11 | | 1.52 | 0.28 | [1.35, 1.69] |
| (A2) ⟶ listening compr. | 8 | 16 | 5.7 | 11 | | −0.77 | 0.14 | [−0.85, −0.68] |
| Lee et al 2014a | | | | | | | | |
| ⟶ nb of grammatical errors | 4 | 12 | 9.0 | 25 | | −0.34 | 0.04 | [−0.36, −0.33] |
| ⟶ nb of words | 4 | 12 | 9.0 | 25 | | 0.59 | 0.05 | [0.57, 0.61] |
| Noh et al 2012 | | | | | | | | |
| POMY ⟶ vocabulary test | 3 | 12 | 8.0 | 40 | | 1.36 | 0.05 | [1.35, 1.38] |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Petersen 2010 | | | | | | | | |
| Sasha → QFT, morphology score | 2 | 3 | 1.5 | 19 | 18 | 0.73 | 0.14 | [0.69, 0.78] |
| Sasha → QFT, syntax score | 2 | 3 | 1.5 | 19 | 18 | 0.96 | 0.17 | [0.90, 1.01] |
| Rosenthal-von der Putten et al 2016 | | | | | | | | |
| Physical robot, prerec. → Cloze test | N/A | 1 | 1.0 | 22 | | 0.11 | 0.04 | [0.10, 0.13] |
| Physical robot, TTS → Cloze test | N/A | 1 | 1.0 | 22 | | −0.17 | 0.04 | [−0.19, −0.15] |
| Speech-only, prerec. → Cloze test | N/A | 1 | 1.0 | 22 | | −0.29 | 0.05 | [−0.31, −0.27] |
| Speech-only, TTS → Cloze test | N/A | 1 | 1.0 | 22 | | −0.22 | 0.04 | [−0.24, −0.20] |
| Virtual agent, prerec. → Cloze test | N/A | 1 | 1.0 | 22 | | −0.28 | 0.05 | [−0.30, −0.26] |
| Virtual agent, TTS → Cloze test | N/A | 1 | 1.0 | 22 | | −0.31 | 0.05 | [−0.33, −0.29] |
| Taguchi et al 2017 | | | | | | | | |
| → multiple choice test *immediate | 1 | 2 | 2.0 | 30 | | 1.84 | 0.10 | [1.80, 1.87] |
| → multiple choice test *delayed | 1 | 2 | 2.0 | 30 | | 2.00 | 0.11 | [1.96, 2.04] |
| → gap-filling test *immediate | 1 | 2 | 2.0 | 30 | | 1.10 | 0.05 | [1.08, 1.12] |
| → gap-filling test *delayed | 1 | 2 | 2.0 | 30 | | 1.58 | 0.08 | [1.55, 1.61] |
| Wilske & Wolska 2011 | | | | | | | | |
| Free prod. + ML CF → GJT *immediate | 2 | 2 | 1.7 | 9 | | 0.55 | 0.16 | [0.45, 0.66] |
| Free prod. + ML CF → GJT *delayed | 2 | 2 | 1.7 | 9 | | 0.60 | 0.16 | [0.49, 0.71] |
| Free prod. + ML CF → SCT *immediate | 2 | 2 | 1.7 | 10 | | 0.32 | 0.12 | [0.25, 0.40] |
| Free prod. + ML CF → SCT *delayed | 2 | 2 | 1.7 | 10 | | 0.32 | 0.12 | [0.25, 0.40] |
| Free prod. + recast → GJT *immediate | 2 | 2 | 1.7 | 11 | | 0.64 | 0.13 | [0.57, 0.72] |
| Free prod. + recast → GJT *delayed | 2 | 2 | 1.7 | 11 | | 0.52 | 0.12 | [0.45, 0.59] |
| Free prod. + recast → SCT *immediate | 2 | 2 | 1.7 | 11 | | 0.52 | 0.12 | [0.45, 0.59] |
| Free prod. + recast → SCT *delayed | 2 | 2 | 1.7 | 11 | | 0.40 | 0.11 | [0.34, 0.47] |
| Gap-filling → GJT *immediate | 2 | 2 | 1.7 | 10 | | 0.91 | 0.18 | [0.80, 1.03] |
| Gap-filling → GJT *delayed | 2 | 2 | 1.7 | 10 | | 0.63 | 0.14 | [0.54, 0.71] |
| Gap-filling → SCT *immediate | 2 | 2 | 1.7 | 11 | | 1.10 | 0.19 | [0.99, 1.21] |
| Gap-filling → SCT *delayed | 2 | 2 | 1.7 | 11 | | 0.59 | 0.12 | [0.52, 0.66] |
| Wilske 2015 | | | | | | | | |
| Free prod. → phonation time ratio | 2 | 2 | 1.7 | 7 | | 0.69 | 0.27 | [0.50, 0.89] |
| Free prod. → length of pauses | 2 | 2 | 1.7 | 7 | | 0.39 | 0.21 | [0.24, 0.55] |
| Free prod. → length of runs | 2 | 2 | 1.7 | 7 | | −0.59 | 0.24 | [−0.77, −0.41] |
| Free prod. → speech rate | 2 | 2 | 1.7 | 7 | | 0.13 | 0.19 | [−0.01, 0.27] |
| Gap-filling → length of pauses | 2 | 2 | 1.7 | 7 | | 0.95 | 0.34 | [0.70, 1.20] |
| Gap-filling → length of runs | 2 | 2 | 1.7 | 7 | | −0.48 | 0.22 | [−0.65, −0.32] |
| Gap-filling → phonation time ratio | 2 | 2 | 1.7 | 7 | | 1.01 | 0.36 | [0.75, 1.28] |
| Gap-filling → speech rate | 2 | 2 | 1.7 | 7 | | 0.26 | 0.19 | [0.11, 0.40] |
| Free prod. + ML CF → GJT *immediate | 2 | 2 | 1.7 | 19 | | 0.00 | 0.05 | [−0.02, 0.02] |
| Free prod. + ML CF → GJT *delayed | 2 | 2 | 1.7 | 9 | | 0.61 | 0.17 | [0.50, 0.72] |
| Free prod. + ML CF → SCT *immediate | 2 | 2 | 1.7 | 19 | | 0.31 | 0.05 | [0.29, 0.33] |
| Free prod. + ML CF → SCT *delayed | 2 | 2 | 1.7 | 10 | | 0.62 | 0.14 | [0.53, 0.71] |
| Free prod. + recast → GJT *immediate | 2 | 2 | 1.7 | 16 | | 0.00 | 0.06 | [−0.03, 0.03] |
| Free prod. + recast → GJT *delayed | 2 | 2 | 1.7 | 11 | | 0.00 | 0.10 | [−0.06, 0.06] |
| Free prod. + recast → SCT *immediate | 2 | 2 | 1.7 | 14 | | 1.21 | 0.15 | [1.13, 1.29] |
| Free prod. + recast → SCT *delayed | 2 | 2 | 1.7 | 10 | | 1.10 | 0.22 | [0.96, 1.23] |
| Gap-filling → GJT *immediate | 2 | 2 | 1.7 | 14 | | 1.33 | 0.16 | [1.24, 1.41] |
| Gap-filling → GJT *delayed | 2 | 2 | 1.7 | 10 | | 0.28 | 0.12 | [0.21, 0.35] |
| Gap-filling → SCT *immediate | 2 | 2 | 1.7 | 14 | | 0.86 | 0.11 | [0.80, 0.92] |
| Gap-filling → SCT *delayed | 2 | 2 | 1.7 | 10 | | 0.46 | 0.13 | [0.38, 0.54] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Wolska & Wilske 2010a | | | | | | | |
| Free prod. → GJT *immediate | 2 | 2 | 1.7 | 6 | 0.69 | 0.36 | [0.41, 0.98] |
| Free prod. → GJT *delayed | 2 | 2 | 1.7 | 6 | 0.44 | 0.29 | [0.21, 0.67] |
| Free prod. → SCT *immediate | 2 | 2 | 1.7 | 6 | 0.53 | 0.31 | [0.28, 0.77] |
| Free prod. → SCT *delayed | 2 | 2 | 1.7 | 6 | 0.42 | 0.28 | [0.20, 0.65] |
| Gap-filling → GJT *immediate | 2 | 2 | 1.7 | 6 | 0.79 | 0.39 | [0.47, 1.10] |
| Gap-filling → GJT *delayed | 2 | 2 | 1.7 | 6 | 0.64 | 0.34 | [0.37, 0.91] |
| Gap-filling → SCT *immediate | 2 | 2 | 1.7 | 6 | 0.84 | 0.42 | [0.51, 1.18] |
| Gap-filling → SCT *delayed | 2 | 2 | 1.7 | 6 | 0.57 | 0.32 | [0.32, 0.83] |
| Wolska & Wilske 2010b | | | | | | | |
| Free prod. → GJT *immediate | 2 | 2 | 1.7 | 7 | 0.49 | 0.22 | [0.32, 0.65] |
| Free prod. → GJT *delayed | 2 | 2 | 1.7 | 7 | 0.49 | 0.22 | [0.32, 0.65] |
| Free prod. → SCT *immediate | 2 | 2 | 1.7 | 7 | 0.29 | 0.20 | [0.14, 0.43] |
| Free prod. → SCT *delayed | 2 | 2 | 1.7 | 7 | 0.29 | 0.20 | [0.14, 0.43] |
| Gap-filling → GJT *immediate | 2 | 2 | 1.7 | 7 | 0.19 | 0.19 | [0.05, 0.33] |
| Gap-filling → GJT *delayed | 2 | 2 | 1.7 | 7 | 0.00 | 0.18 | [−0.14, 0.14] |
| Gap-filling → SCT *immediate | 2 | 2 | 1.7 | 7 | 0.41 | 0.21 | [0.25, 0.57] |
| Gap-filling → SCT *delayed | 2 | 2 | 1.7 | 7 | 0.31 | 0.20 | [0.16, 0.45] |
| Wolska & Wilske 2011 | | | | | | | |
| Free prod. → phonation time ratio | 2 | 2 | 1.7 | 6 | 1.85 | 1.10 | [0.97, 2.73] |
| Free prod. → length of pauses | 2 | 2 | 1.7 | 6 | 1.23 | 0.62 | [0.74, 1.73] |
| Free prod. → length of runs | 2 | 2 | 1.7 | 6 | −0.69 | 0.36 | [−0.97, −0.40] |
| Free prod. → speech rate | 2 | 2 | 1.7 | 6 | 0.39 | 0.28 | [0.17, 0.61] |
| Gap-filling → length of pauses | 2 | 2 | 1.7 | 7 | −0.32 | 0.20 | [−0.47, −0.17] |
| Gap-filling → length of runs | 2 | 2 | 1.7 | 7 | −0.48 | 0.22 | [−0.65, −0.32] |
| Gap-filling → phonation time ratio | 2 | 2 | 1.7 | 7 | 0.46 | 0.22 | [0.29, 0.62] |
| Gap-filling → speech rate | 2 | 2 | 1.7 | 7 | 0.38 | 0.21 | [0.23, 0.54] |

*Sess.* = number of treatment sessions. $n_E$ = sample size of experimental group. $n_C$ = sample size of control group. *EM* = English major students. *DCT* = dialogue completion test. *compr*$^{fy}$ = comprehensibility rating. *adeq.* = adequate use of speech acts rating. *compr.* = comprehension test. *prof.* = proficiency rating. *hol.* = holistic rating. *comm.* = communicative. *QFT* = questions formation test. *GJT* = grammatical judgement test. *SCT* = sentence construction test. *immediate* = immediate posttest. *delayed* = delayed posttest. *prerec.* = pre-recorded voice. *TTS* = text-to-speech, synthesized voice. *Free prod.* = system allowing free production. *ML CF* = metalinguistic corrective feedback. N/A = non-reported or inapplicable value.