"Hey Siri, can I learn English by talking to you?"

Insights from a multilevel **meta-analysis** on the effectiveness of **dialogue-based CALL**

Serge **Bibauw**

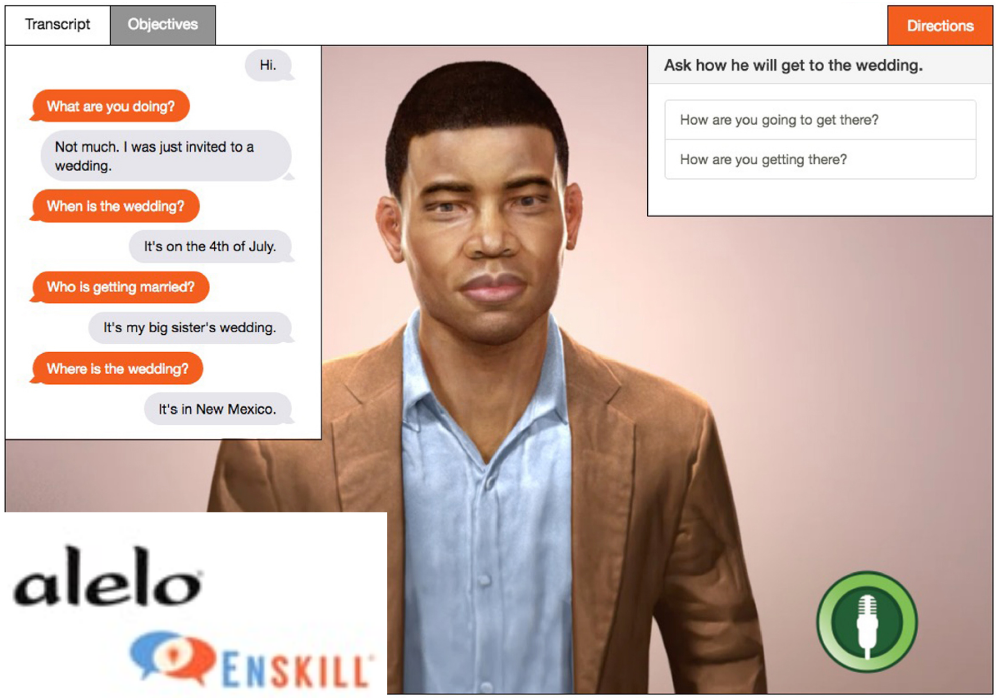
Wim **Van den Noortgate**

Thomas **François**

Piet **Desmet**

CALL 2018 Conference
July 5, 2018

Dialogue-based CALL

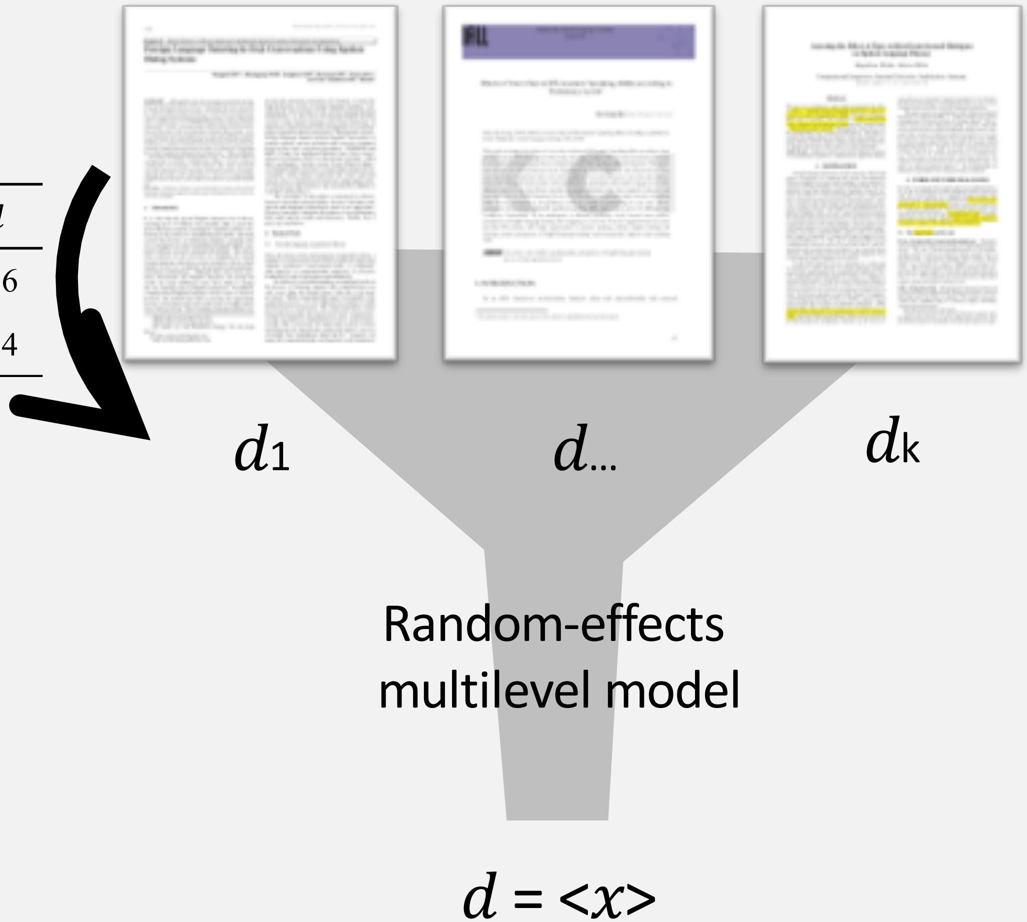


duolingo bots

Apple
Siri

Meta-analysis of effectiveness studies

	M	sd
Pre	56	4.56
Post	61	6.54



Insights from a multilevel meta-analysis on the effectiveness of dialogue-based CALL

Object: dialogue-based CALL

Dialogue systems, chatbots, agents

Methods: meta-analysis

Studies collection and selection, effect sizes calculation and multilevel modeling

Results: effectiveness for L2 learning

General effectiveness

Relative effects per population, treatment characteristics and outcome variables

Insights from a multilevel meta-analysis on the effectiveness of dialogue-based CALL

Object: dialogue-based CALL
Dialogue systems, chatbots, agents

Methods: meta-analysis
Studies collection and selection, effect sizes calculation and multilevel modeling

Results: effectiveness for L2 learning
General effectiveness
Relative effects per population, treatment characteristics and outcome variables

Dialogue-based CALL

Dialogue-based CALL refers to any application or system allowing,

to maintain a dialogue

[immediate, synchronous interaction]
[written or spoken]

with an automated agent

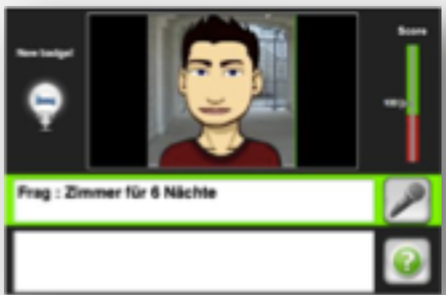
[tutorial CALL (≠ CMC)]

for language learning purposes.

Bibauw, François & Desmet, 2015 (EUROCALL Proceedings); Bibauw, François & Desmet, *under review*

Dialogue-based CALL
Typology of systems

(Bibauw *et al, under review*)



Form-focused dialogue systems

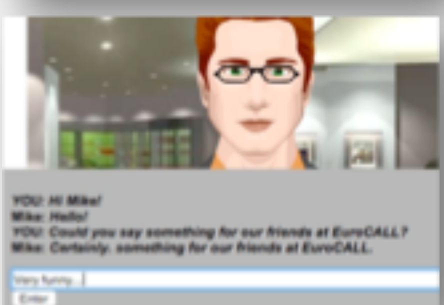
Explicit constraints on meaning,
focus on form/forms

e.g., ICALL intelligent language tutors, and Computer-assisted pronunciation training (CAPT) systems

Goal-oriented dialogue systems

Contextual constraints (task, situated conversation...),
mostly focus on meaning and interaction

e.g., Conversational agents in virtual worlds



Reactive dialogue systems

Free, user-initiated, open-ended dialogue

e.g., Chatbots, and personal assistants

Here, simplified typology (left out *Narrative systems*)

Dialogue-based CALL
Recent evolutions

Rich history of studies & systems:

- First attempts in the 80s (Underwood 1982, 1984)
- *Intelligent Language Tutors* developed in the 90s (Holland et al, 1995)
- Efforts with speech and dialogue in the 2000s (Raux & Eskenazi, 2004; Seneff et al, 2007; Morton et al 2012)
- Principled technological convergence more recently (Petersen, 2010; Wilske, 2015)

But nearly all systems remained internal, research-only prototypes, never made accessible to the public.

→ No comparability, no replicability

But, recently, **major advances towards publicly available tools** (Duolingo Bots, Alelo Enskill, ETS HALEF) and **joint efforts between industry and researchers** to compare the systems and establish common ground (Sydorenko et al, 2018)

Insights from a multilevel meta-analysis on the effectiveness of dialogue-based CALL

Object: dialogue-based CALL

Dialogue systems, chatbots, agents

Methods: meta-analysis

Studies collection and selection, effect sizes calculation and multilevel modeling

Results: effectiveness for L2 learning

General effectiveness

Relative effects per population, treatment characteristics and outcome variables

Meta-analysis of effectiveness studies

Aggregate results from multiple experimental studies

Treat each study as a subject

Get a more powerful, generalizable, stable and precise idea of the effectiveness of dialogue-based CALL on language learning

Analyzing certain moderator variables to identify tendencies inside the data

Bibauw, François & Desmet, 2015 (EUROCALL Proceedings); Bibauw, François & Desmet, *in prep.*

Meta-analysis
Search & collection process

1. **Database** search
   in Web of Science, Scopus, ProQuest…

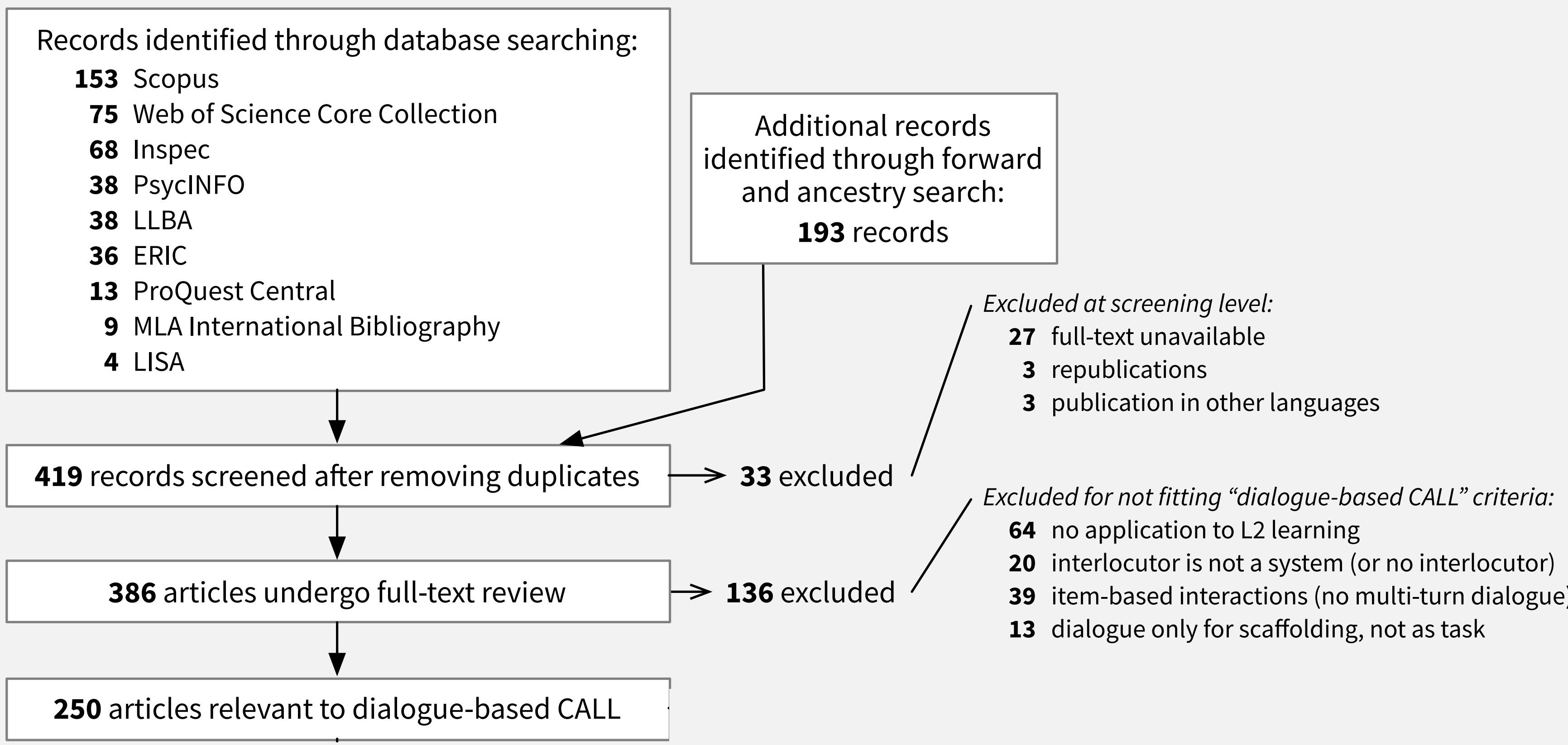
   Search syntax:
   ```
   (chatbot / chat bot / chatterbot /
   conversational agent / conversational companion
   / conversational system / dialog* system /
   dialog* agent / dialog* game / pedagogical agent
   / human-computer dialog* / dialog*-based) +
   ((language / English) (learning / teaching /
   acquisition) / (second / foreign) language / L2
   / EFL / ESL / ICALL)
   ```

2. **Ancestry** search
   Older publications cited by ref

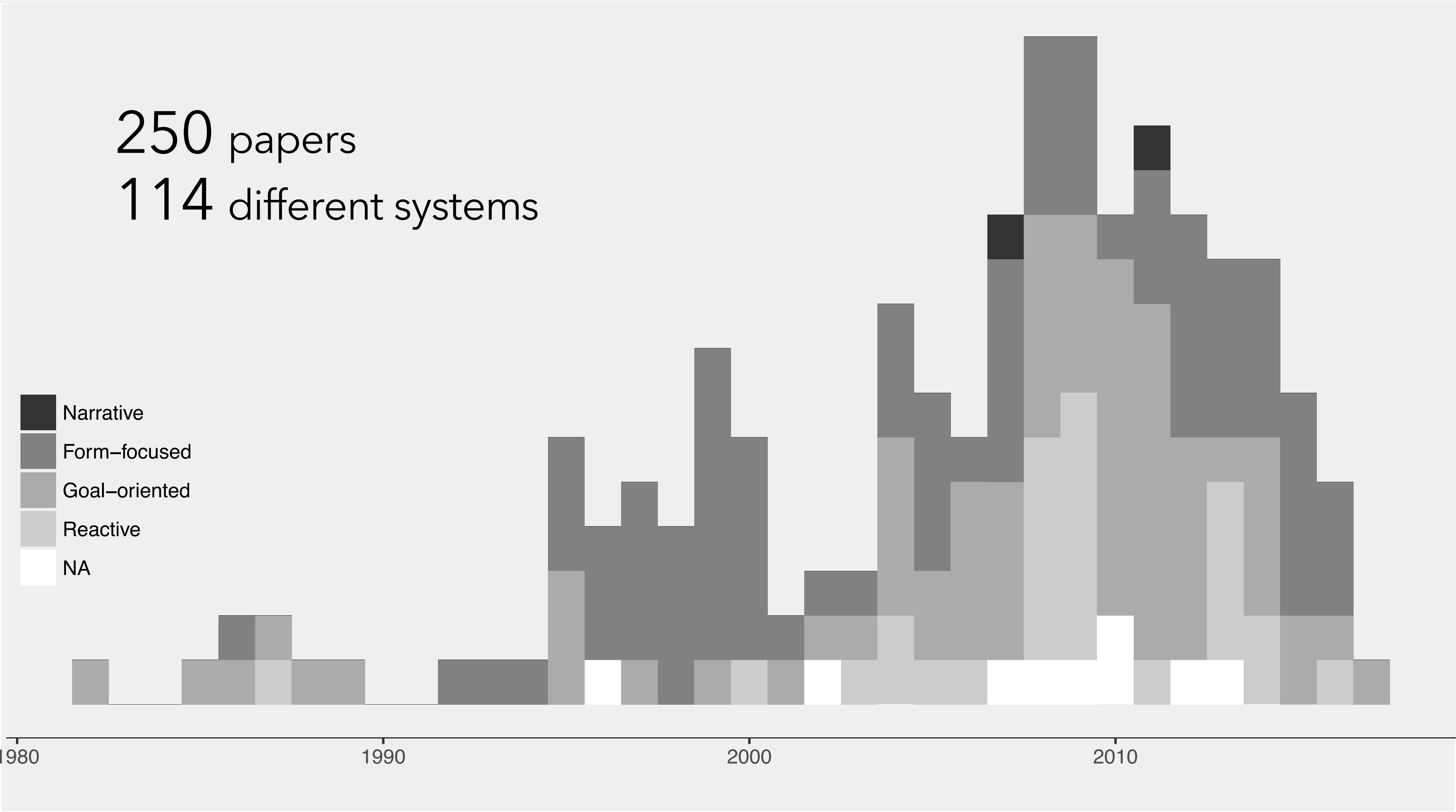
3. **Forward** citations
   New publications citing ref

Note on journal search: 40/250 publications from the 4 major CALL journals (19 *CALL*, 13 *CALICO J.*, 4 *ReCALL*, 4 *LL&T*)

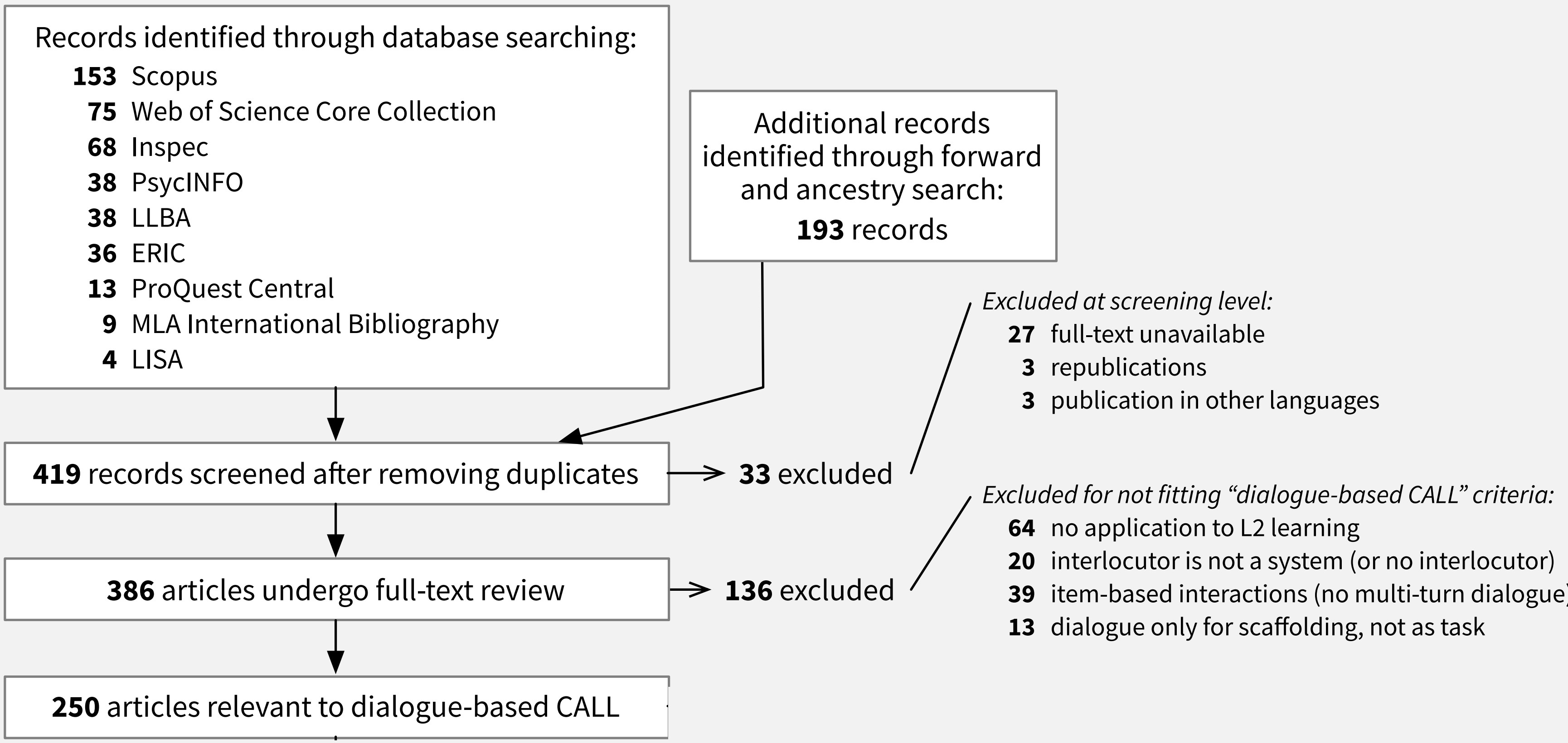
Meta-analysis
Inclusion/exclusion process



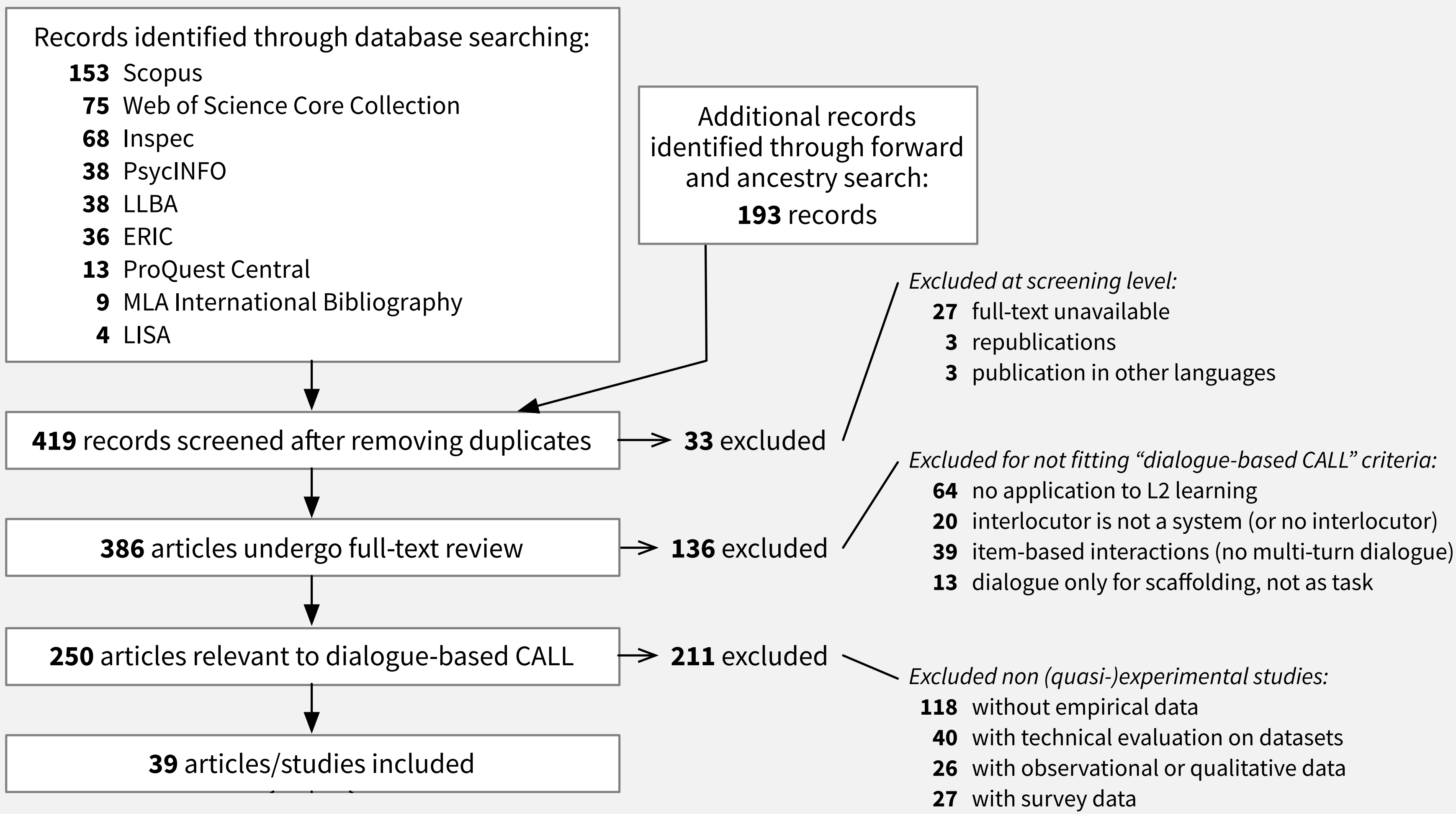
Studies on dialogue-based CALL



Meta-analysis
Inclusion/exclusion process



Meta-analysis
Inclusion/exclusion process



Coding scheme

ref	system	dep_var	proficiency_level	n_treatment	m_t_pre	m_t_post
Lee et al 2012	POMY	Comprehension	A1	21	10.9500000	10.6700000
Harless et al 1999	Conversim	Comprehension	<NA>	9	73.0000000	75.0000000
Lee et al 2014	POMY	Accuracy	mixed	25	-0.3081438	-0.2611765
Lee et al 2012	POMY	Accuracy	A1	21	31.6200000	40.6200000
Hassani et al 2016	IVELL	Accuracy	A2	10	-0.0670000	-0.0360000
Rayner & Tsourakis 2013	CALL-SLT	Accuracy	A1	12	0.0000000	22.8876200
Hassani et al 2016	IVELL	Complexity	A2	10	0.4160000	0.6920000
Lee et al 2012	POMY	Fluency	A1	21	33.5700000	47.4800000
Lee et al 2014	POMY	Fluency	mixed	25	136.3000000	170.0000000
Hassani et al 2016	IVELL	Fluency	A2	10	-0.4180000	-0.2620000
Wolska & Wilske 2011	[Wilske2]	Fluency	mixed	6	0.5700000	0.6800000
Wilske 2014	[Wilske2]	Fluency	mixed	7	0.8200000	0.8600000
Wolska & Wilske 2011	[Wilske2]	Fluency	mixed	6	2.0500000	2.1900000
Wilske 2014	[Wilske2]	Fluency	mixed	7	2.3900000	2.4600000
Kim 2016	Indigo	Proficiency	A1	20	64.5000000	112.5000000

Publication variables
author, year, publication type, source, sample…

Population variables
context, age, L1, L2 proficiency level

Treatment variables
experimental design, treatment duration (weeks), time on task (hours), number of sessions, treatment density (packed vs. spaced)

System variables
system, target L2, system_type, dialogue_type, primary_modality, corrective_feedback, initiative, embodied_agent, gamified…

Instruments/outcome variables
proficiency/complexity/accuracy/fluency/vocabulary, speaking/writing, specific test

Quantitative results
n, mean, sd (pre/post, experimental/control)

Meta-analysis
Computable effect sizes

Effect size: standardized measure of the observed (here, learning) effect

Effect size (d) typically computed over:

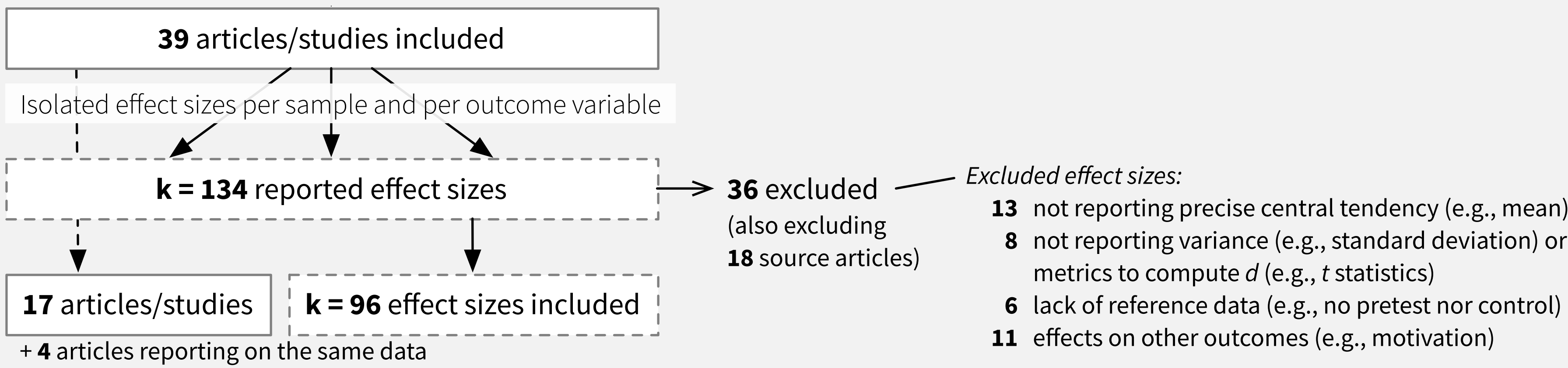
- **mean**
- **standard deviation**
- **n** (subjects)

for each group/measurement point
(or alternate: t-score, etc.)

Not available for all studies (especially older studies)

Asked the authors for raw data
(worked for some – thanks to them!)

Meta-analysis
Inclusion of individual effect sizes



k = 96 effect sizes

Meta-analysis

Effect size calculation

Effect size: standardized measure of the observed (here, learning) effect

Usually, in SLA/CALL:

Standardized Mean Difference

Cohen's d (M_{post} – M_{pre} / SD_{pooled})

Hedge's g

	Exp. Grp M (sd)	Control M (sd)
Post	61 (6.2)	57 (7.4)

EC

	M (sd)
Pre	56 (4.3)
Post	61 (6.2)

PP

	Exp. Grp M (sd)	Control M (sd)
Pre	56 (4.3)	54 (5.6)
Post	61 (6.2)	57 (7.4)

ECPP

Standardized Mean Change

Meta-analysis
A comparable effect size metrics

Morris & DeShon (2002) offer a solution: comparable metrics across experimental designs (EC / PP / ECPP)

- *change* metric (aligned on *within*-group effect)
- *raw* metric (aligned on *between*-groups effect)

We selected the *raw* metric formula:

$$d_{\mathrm{PP}} = J(df_{\mathrm{PP}}) \left(\frac{M_{\mathrm{post,E}} - M_{\mathrm{pre,E}}}{SD_{\mathrm{pre,E}}} \right)$$

$$d_{\mathrm{ECPP}} = J(df_{\mathrm{ECPP}}) \left(\frac{M_{\mathrm{post,E}} - M_{\mathrm{pre,E}}}{SD_{\mathrm{pre,E}}} - \frac{M_{\mathrm{post,C}} - M_{\mathrm{pre,C}}}{SD_{\mathrm{pre,C}}} \right)$$

Meta-analysis
Summary effect size

Model computes a **summary effect** by aggregating all the single study effect sizes

Weighting according to sample size and precision

→ More powerful, more stable, more precise and generalizable than the individual study effect sizes

Meta-analysis
Multilevel modeling

Publications report multiple outcome measures (e.g., vocabulary and morphology tests) or multiple sampling groups (e.g., proficiency levels)

Traditional meta-analysis techniques allow only one (independent) effect size per study, but loosing thus all the information on distinct implementations

⇒ Including all the variation without "fooling" the model with non-independent measures:

Multilevel modelling

> Aggregates **multiple effects per study**, by adding an intermediate level of *within*-study variation.

Table 1: Levels of multilevel meta-analytic model

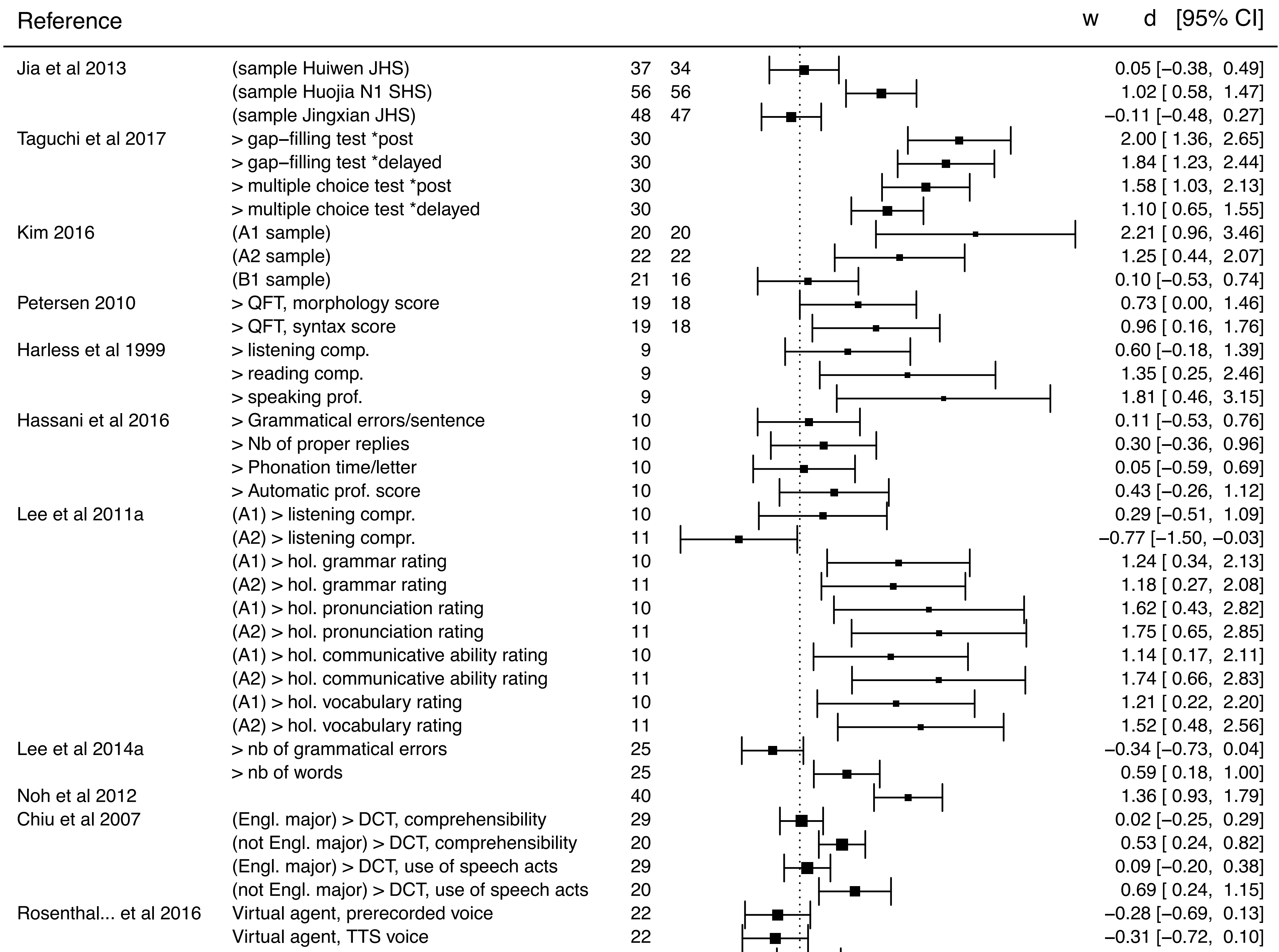
	Level	Number of clusters/items	Source of variance
1	Samples	$k = 96$ ($n = 803$)	Random sampling variance
2	Effects sizes	$k = 96$	Variation within study
3	Studies	$l = 17$	Variation between studies

Insights from a multilevel meta-analysis on the effectiveness of dialogue-based CALL

Object: dialogue-based CALL
Dialogue systems, chatbots, agents

Methods: meta-analysis
Studies collection and selection, effect sizes calculation and multilevel modeling

Results: effectiveness for L2 learning
General effectiveness
Relative effects per population, treatment characteristics and outcome variables


Reference
w
d [95% CI]
Jia et al 2013 (sample Huiwen JHS) 37 34 0.05 [−0.38, 0.49]
(sample Huojia N1 SHS) 56 56 1.02 [0.58, 1.47]
(sample Jingxian JHS) 48 47 −0.11 [−0.48, 0.27]
Taguchi et al 2017 > gap−filling test *post 30 2.00 [1.36, 2.65]
> gap−filling test *delayed 30 1.84 [1.23, 2.44]
> multiple choice test *post 30 1.58 [1.03, 2.13]
> multiple choice test *delayed 30 1.10 [0.65, 1.55]
Kim 2016 (A1 sample) 20 20 2.21 [0.96, 3.46]
(A2 sample) 22 22 1.25 [0.44, 2.07]
(B1 sample) 21 16 0.10 [−0.53, 0.74]
Petersen 2010 > QFT, morphology score 19 18 0.73 [0.00, 1.46]
> QFT, syntax score 19 18 0.96 [0.16, 1.76]
Harless et al 1999 > listening comp. 9 0.60 [−0.18, 1.39]
> reading comp. 9 1.35 [0.25, 2.46]
> speaking prof. 9 1.81 [0.46, 3.15]
Hassani et al 2016 > Grammatical errors/sentence 10 0.11 [−0.53, 0.76]
> Nb of proper replies 10 0.30 [−0.36, 0.96]
> Phonation time/letter 10 0.05 [−0.59, 0.69]
> Automatic prof. score 10 0.43 [−0.26, 1.12]
Lee et al 2011a (A1) > listening compr. 10 0.29 [−0.51, 1.09]
(A2) > listening compr. 11 −0.77 [−1.50, −0.03]
(A1) > hol. grammar rating 10 1.24 [0.34, 2.13]
(A2) > hol. grammar rating 11 1.18 [0.27, 2.08]
(A1) > hol. pronunciation rating 10 1.62 [0.43, 2.82]
(A2) > hol. pronunciation rating 11 1.75 [0.65, 2.85]
(A1) > hol. communicative ability rating 10 1.14 [0.17, 2.11]
(A2) > hol. communicative ability rating 11 1.74 [0.66, 2.83]
(A1) > hol. vocabulary rating 10 1.21 [0.22, 2.20]
(A2) > hol. vocabulary rating 11 1.52 [0.48, 2.56]
Lee et al 2014a > nb of grammatical errors 25 −0.34 [−0.73, 0.04]
> nb of words 25 0.59 [0.18, 1.00]
Noh et al 2012 40 1.36 [0.93, 1.79]
Chiu et al 2007 (Engl. major) > DCT, comprehensibility 29 0.02 [−0.25, 0.29]
(not Engl. major) > DCT, comprehensibility 20 0.53 [0.24, 0.82]
(Engl. major) > DCT, use of speech acts 29 0.09 [−0.20, 0.38]
(not Engl. major) > DCT, use of speech acts 20 0.69 [0.24, 1.15]
Rosenthal... et al 2016 Virtual agent, prerecorded voice 22 −0.28 [−0.69, 0.13]
Virtual agent, TTS voice 22 −0.31 [−0.72, 0.10]

Results

Summary effect

General effectiveness of dialogue-based CALL for L2 proficiency development (k = 96):

d = 0.605 ***

95% CI = [0.377, 0.833]

= Medium effect (Plonsky & Oswald, 2014)

Results & discussion
Summary effect compared to CALL/SLA

Global effect close to the median of meta-analyses in CALL/SLA (Plonsky & Oswald, 2014)

- ≳ game-based learning (d = .53, Chiu et al, 2012)
- ≲ CALL in general (d = .84, Plonsky & Ziegler, 2016)

Consistent with effect of face-to-face interaction (Mackey & Goo, 2007) and SCMC.

- ≲ F2F interaction (d = .75, Mackey & Goo, 2007)
- ≲ SCMC (Ziegler, 2015; Lin, 2015)

Slightly inferior, but logical:

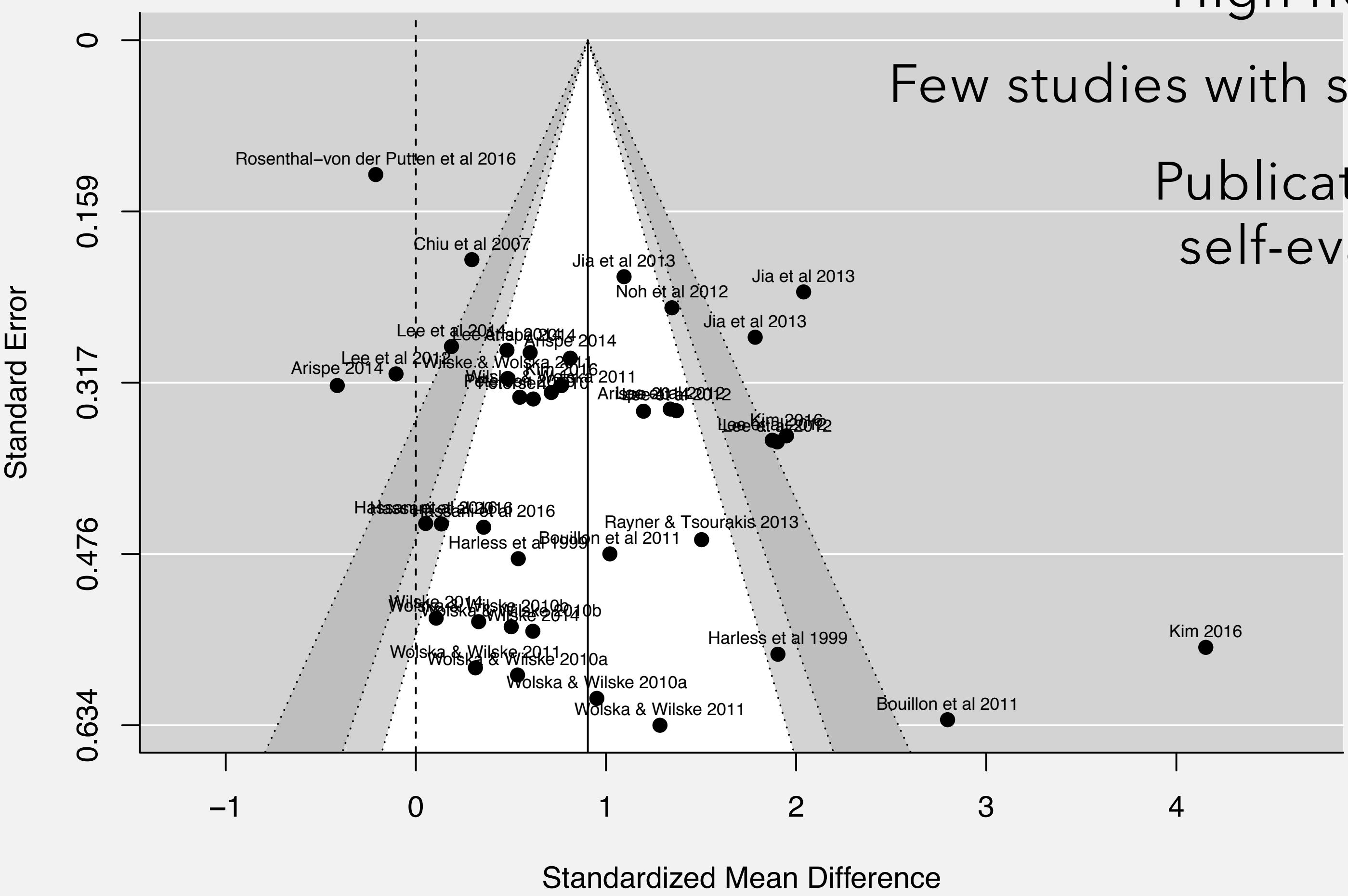
- Human interlocutors remain the gold standard!
- Outcome variables often very ambitious (holistic proficiency…) and treatment duration often very reduced (≤ 3h)

Results & discussion
Limitations

High heterogeneity

Few studies with strong results

Publication bias and
self-evaluation bias



Insights from a multilevel meta-analysis on the effectiveness of dialogue-based CALL

Object: dialogue-based CALL

Dialogue systems, chatbots, agents

Methods: meta-analysis

Studies collection and selection, effect sizes calculation and multilevel modeling

Results: effectiveness for L2 learning

General effectiveness

Relative effects per population, treatment characteristics and outcome variables

Results

Moderator analysis

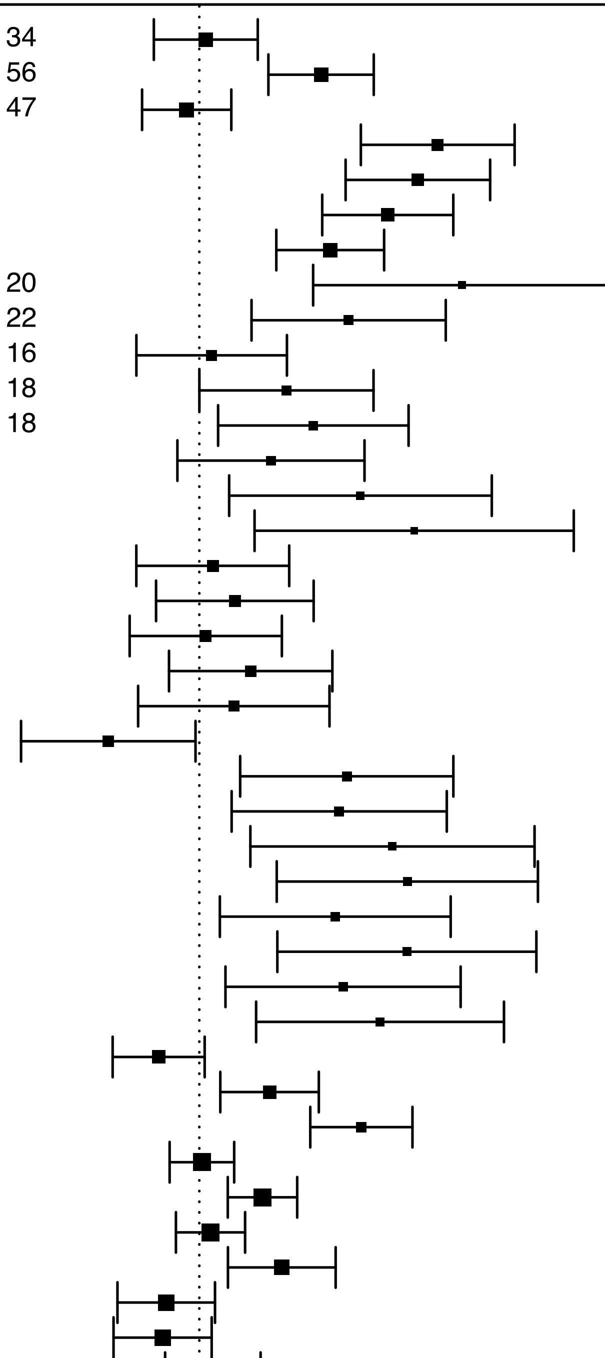
Insights about the influence of some **covariates/moderators**

Sample and context
context, age, L1, L2, proficiency level

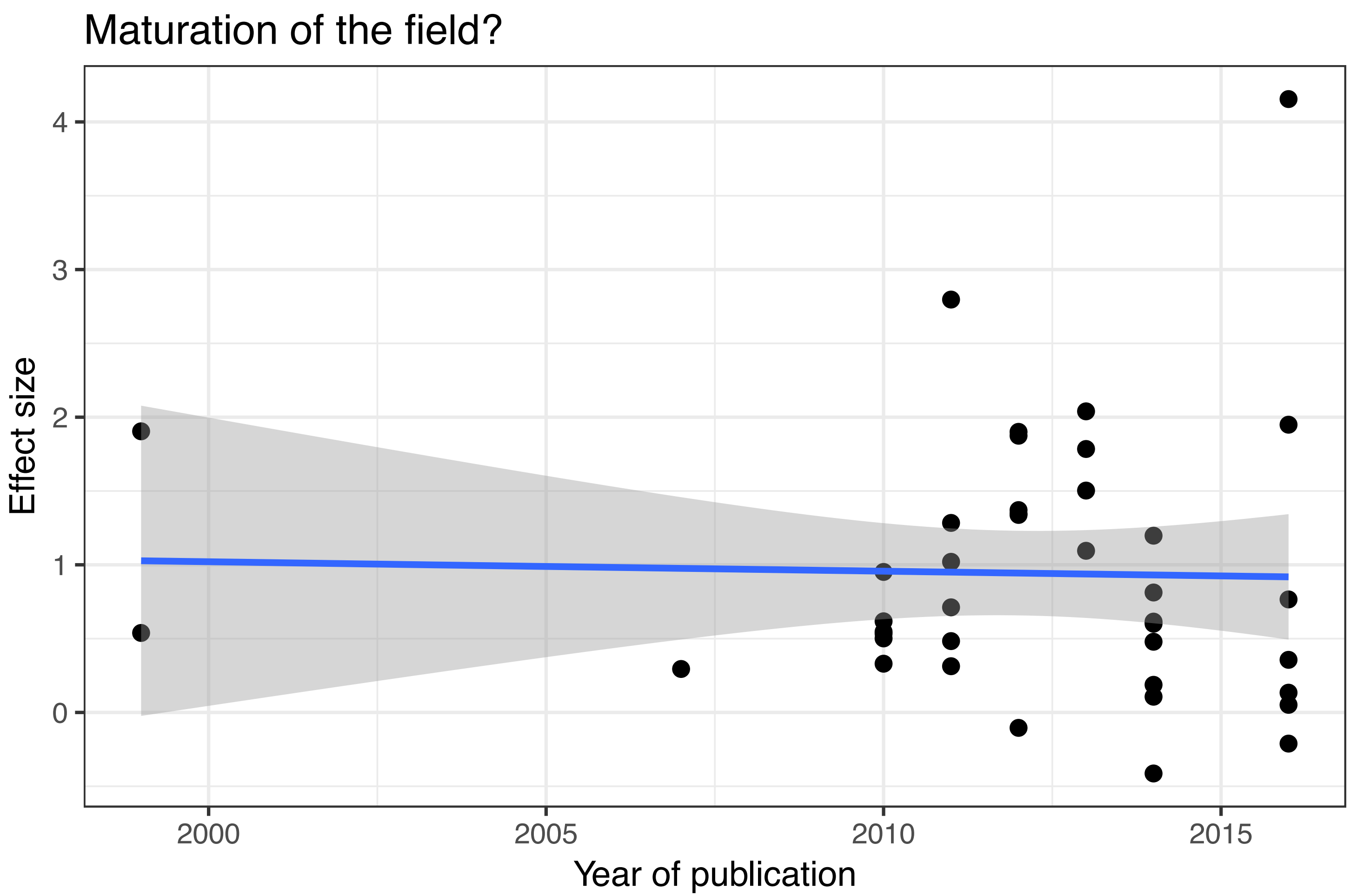
System (treatment) variables
system, system type, dialogue type, primary modality, corrective feedback, initiative, embodied agent, gamified...
treatment duration (in weeks),
time on task (in hours)

Instruments/outcome variables
proficiency/complexity/accuracy/fluency/vocabulary, speaking/writing, specific test

Reference				w	d [95% CI]
Jia et al 2013	(sample Huiwen JHS)	37	34		0.05 [−0.38, 0.49]
	(sample Huojia N1 SHS)	56	56		1.02 [0.58, 1.47]
	(sample Jingxian JHS)	48	47		−0.11 [−0.48, 0.27]
Taguchi et al 2017	> gap−filling test *post	30			2.00 [1.36, 2.65]
	> gap−filling test *delayed	30			1.84 [1.23, 2.44]
	> multiple choice test *post	30			1.58 [1.03, 2.13]
	> multiple choice test *delayed	30			1.10 [0.65, 1.55]
Kim 2016	(A1 sample)	20	20		2.21 [0.96, 3.46]
	(A2 sample)	22	22		1.25 [0.44, 2.07]
	(B1 sample)	21	16		0.10 [−0.53, 0.74]
Petersen 2010	> QFT, morphology score	19	18		0.73 [0.00, 1.46]
	> QFT, syntax score	19	18		0.96 [0.16, 1.76]
Harless et al 1999	> listening comp.	9			0.60 [−0.18, 1.39]
	> reading comp.	9			1.35 [0.25, 2.46]
	> speaking prof.	9			1.81 [0.46, 3.15]
Hassani et al 2016	> Grammatical errors/sentence	10			0.11 [−0.53, 0.76]
	> Nb of proper replies	10			0.30 [−0.36, 0.96]
	> Phonation time/letter	10			0.05 [−0.59, 0.69]
	> Automatic prof. score	10			0.43 [−0.26, 1.12]
Lee et al 2011a	(A1) > listening compr.	10			0.29 [−0.51, 1.09]
	(A2) > listening compr.	11			−0.77 [−1.50, −0.03]
	(A1) > hol. grammar rating	10			1.24 [0.34, 2.13]
	(A2) > hol. grammar rating	11			1.18 [0.27, 2.08]
	(A1) > hol. pronunciation rating	10			1.62 [0.43, 2.82]
	(A2) > hol. pronunciation rating	11			1.75 [0.65, 2.85]
	(A1) > hol. communicative ability rating	10			1.14 [0.17, 2.11]
	(A2) > hol. communicative ability rating	11			1.74 [0.66, 2.83]
	(A1) > hol. vocabulary rating	10			1.21 [0.22, 2.20]
	(A2) > hol. vocabulary rating	11			1.52 [0.48, 2.56]
Lee et al 2014a	> nb of grammatical errors	25			−0.34 [−0.73, 0.04]
	> nb of words	25			0.59 [0.18, 1.00]
Noh et al 2012		40			1.36 [0.93, 1.79]
Chiu et al 2007	(Engl. major) > DCT, comprehensibility	29			0.02 [−0.25, 0.29]
	(not Engl. major) > DCT, comprehensibility	20			0.53 [0.24, 0.82]
	(Engl. major) > DCT, use of speech acts	29			0.09 [−0.20, 0.38]
	(not Engl. major) > DCT, use of speech acts	20			0.69 [0.24, 1.15]
Rosenthal... et al 2016	Virtual agent, prerecorded voice	22			−0.28 [−0.69, 0.13]
	Virtual agent, TTS voice	22			−0.31 [−0.72, 0.10]

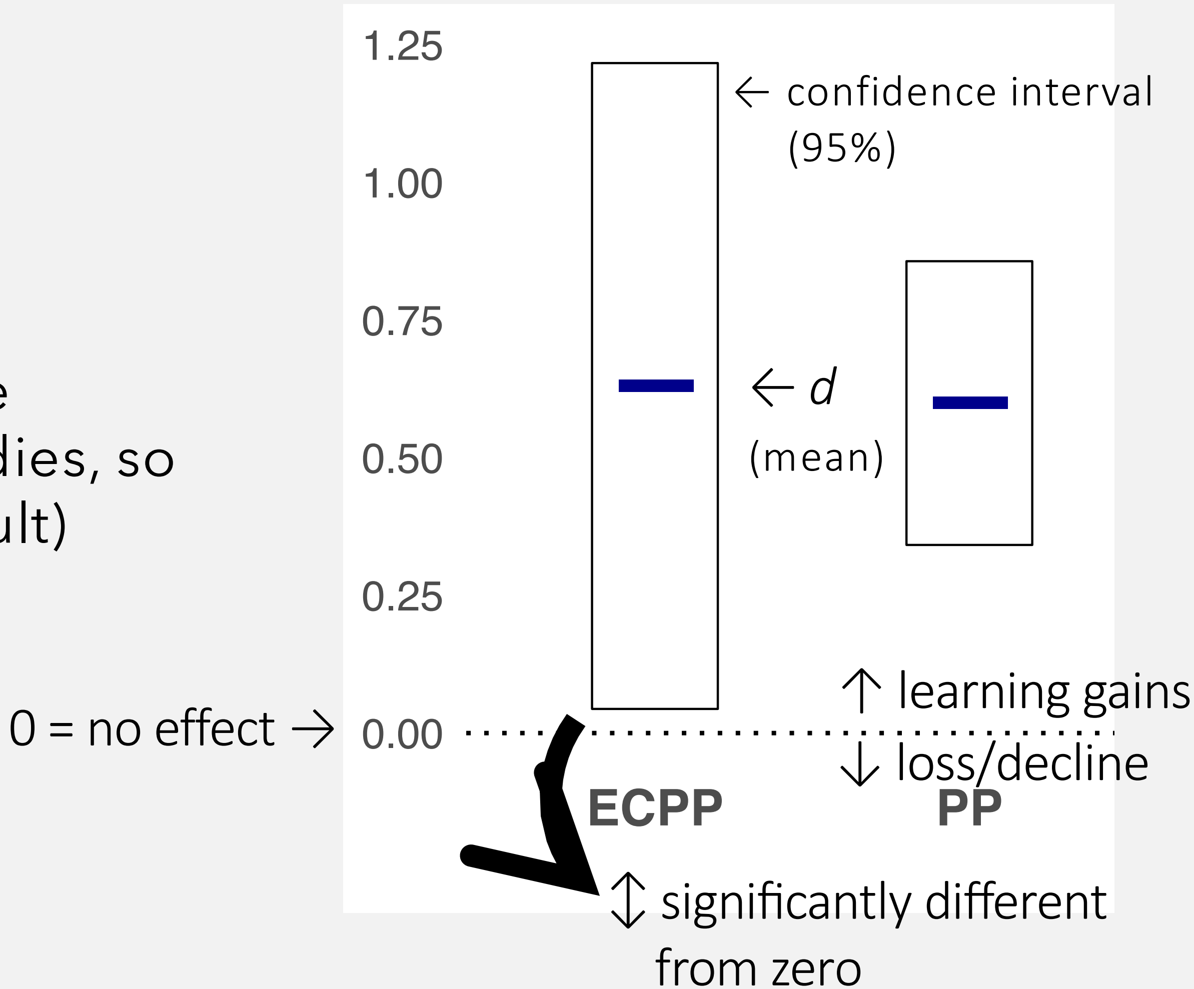


Moderator analysis
Evolution across time



Moderator analysis
Experimental design

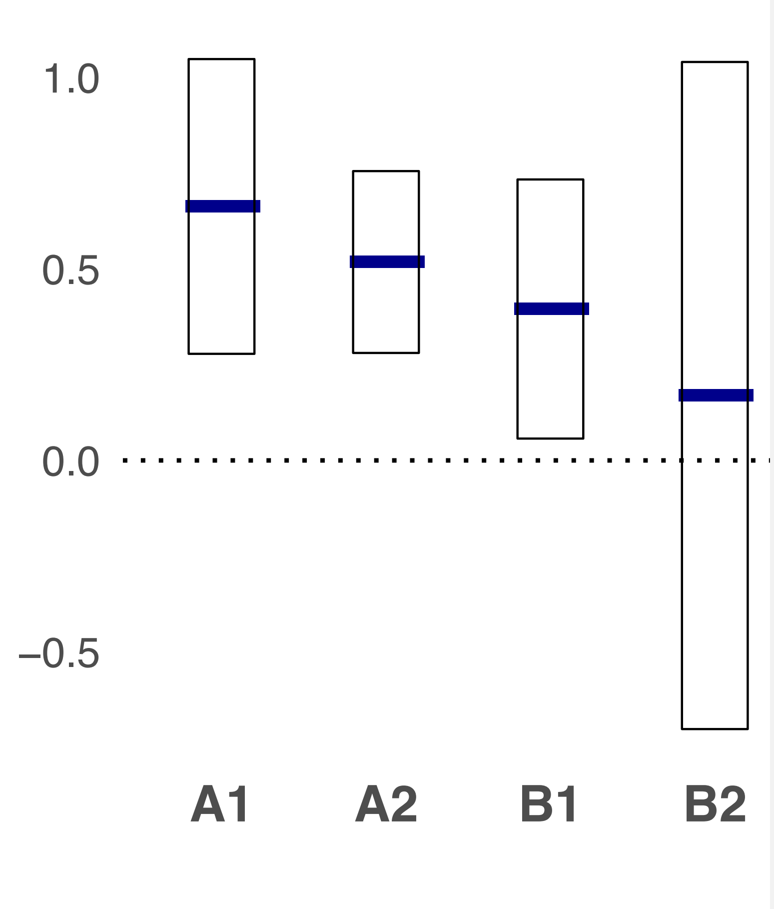
No major difference
(much more PP studies, so
more confident result)



Moderator analysis
Participants: L2 proficiency

Beginners benefit more from these systems than advanced learners

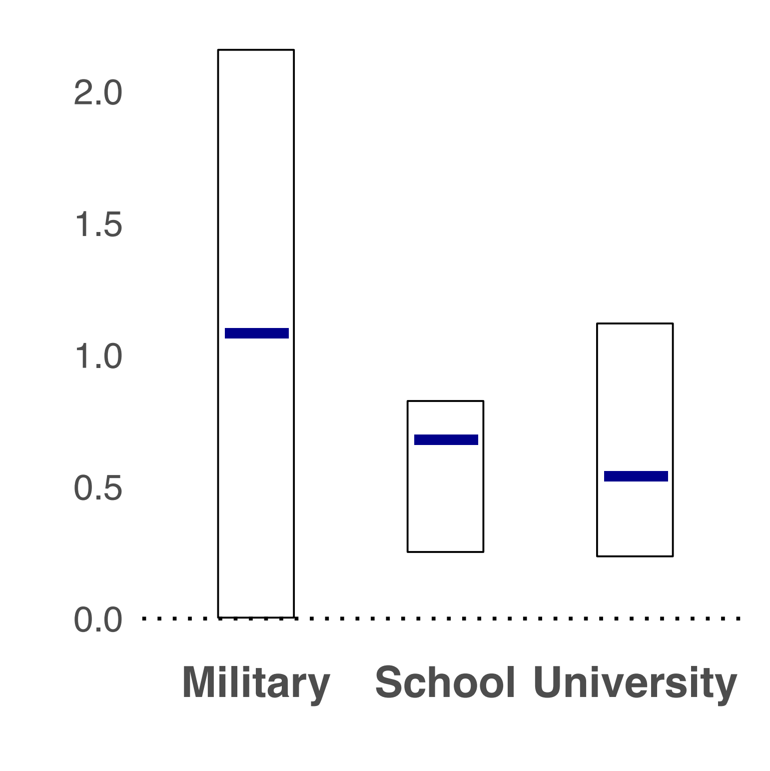
Very significant difference and predictor
($Q(df=3)$ = 10.8, p < .001)



Moderator analysis
Context

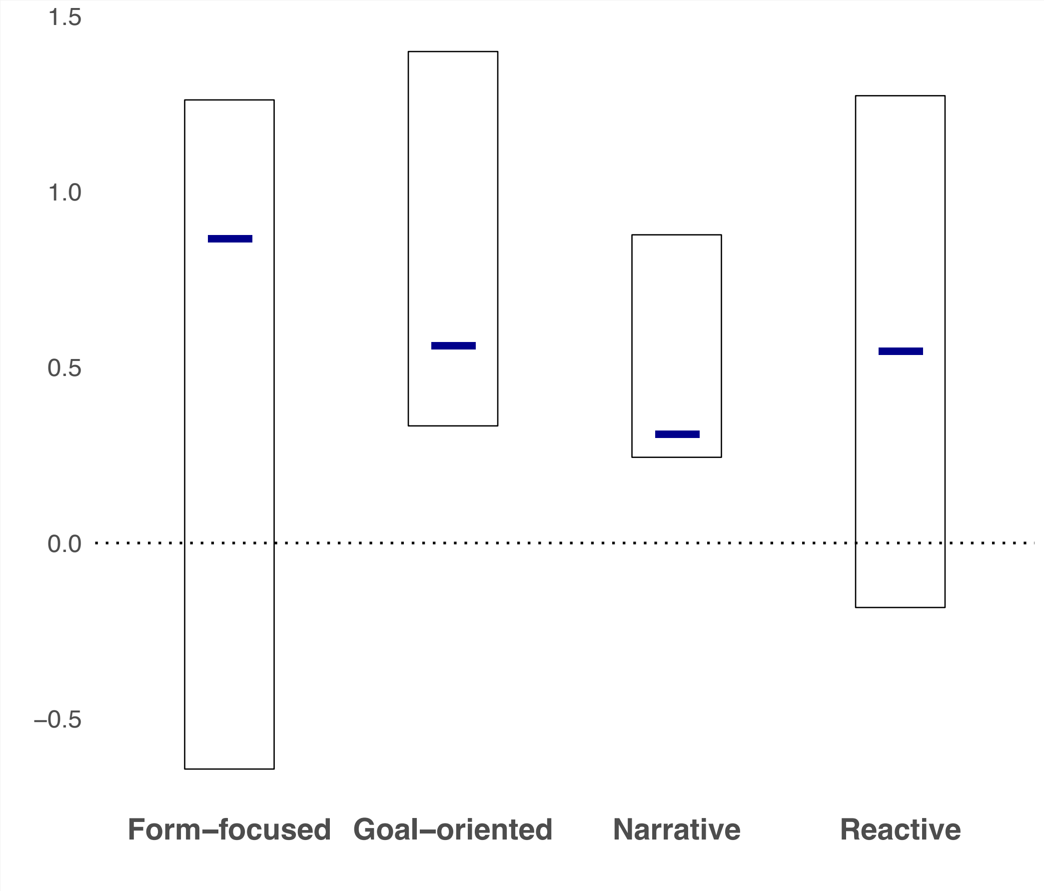
No significant difference
(p = .58)

Seems to be effective both in the school as the university context (+ external, such as military, underrepresented).



Moderator analysis
Type of system

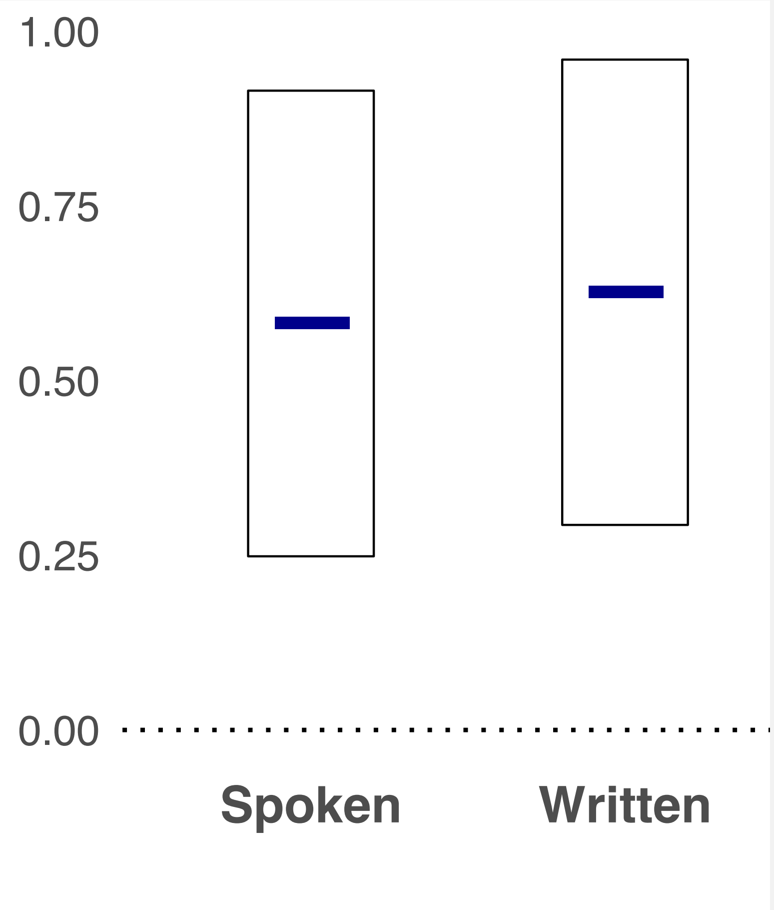
Goal-oriented systems seem to be more effective globally.



Moderator analysis

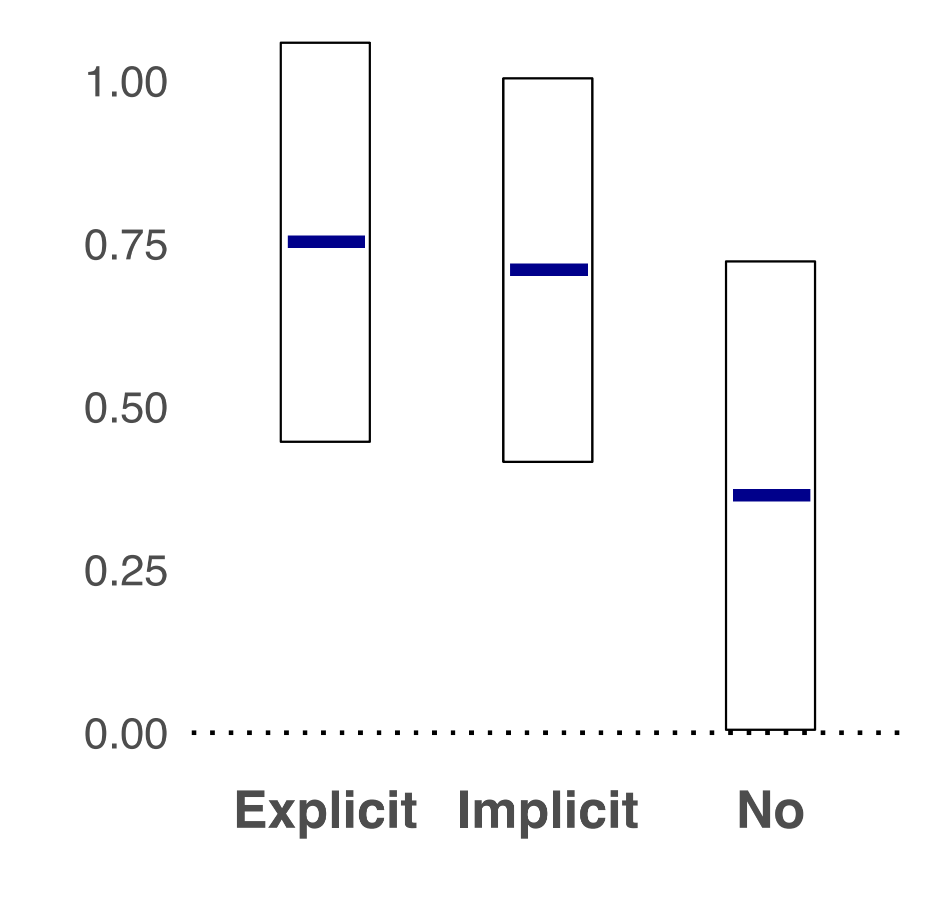
System modality

Very similar effects, in both modalities.



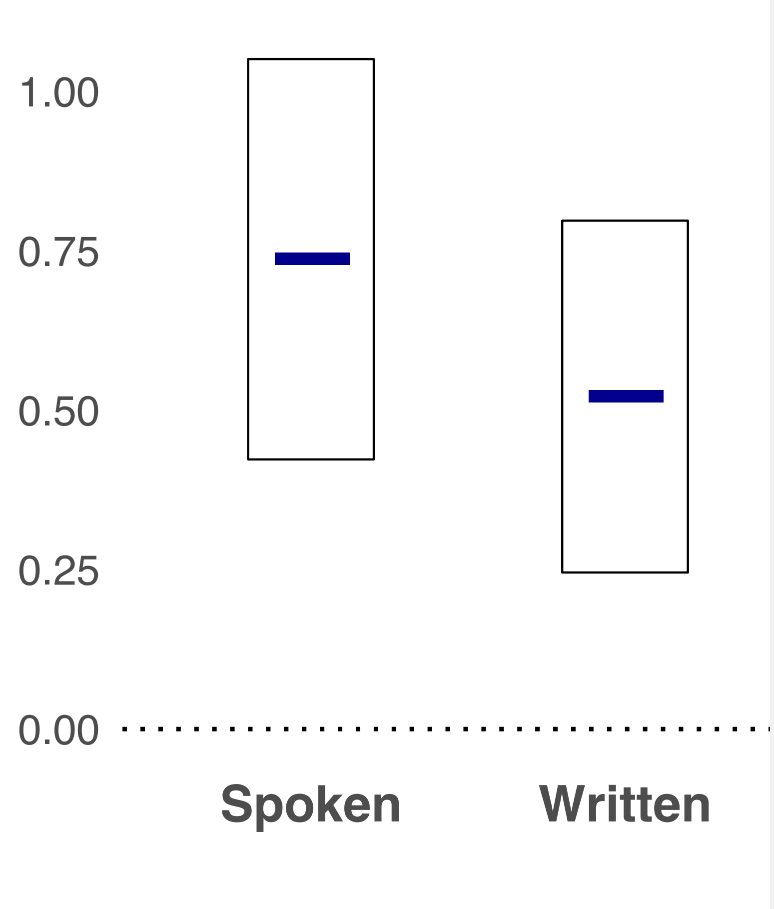
Moderator analysis
System: Corrective feedback

Consistently with what we know about corrective feedback, systems providing feedback are much more effective

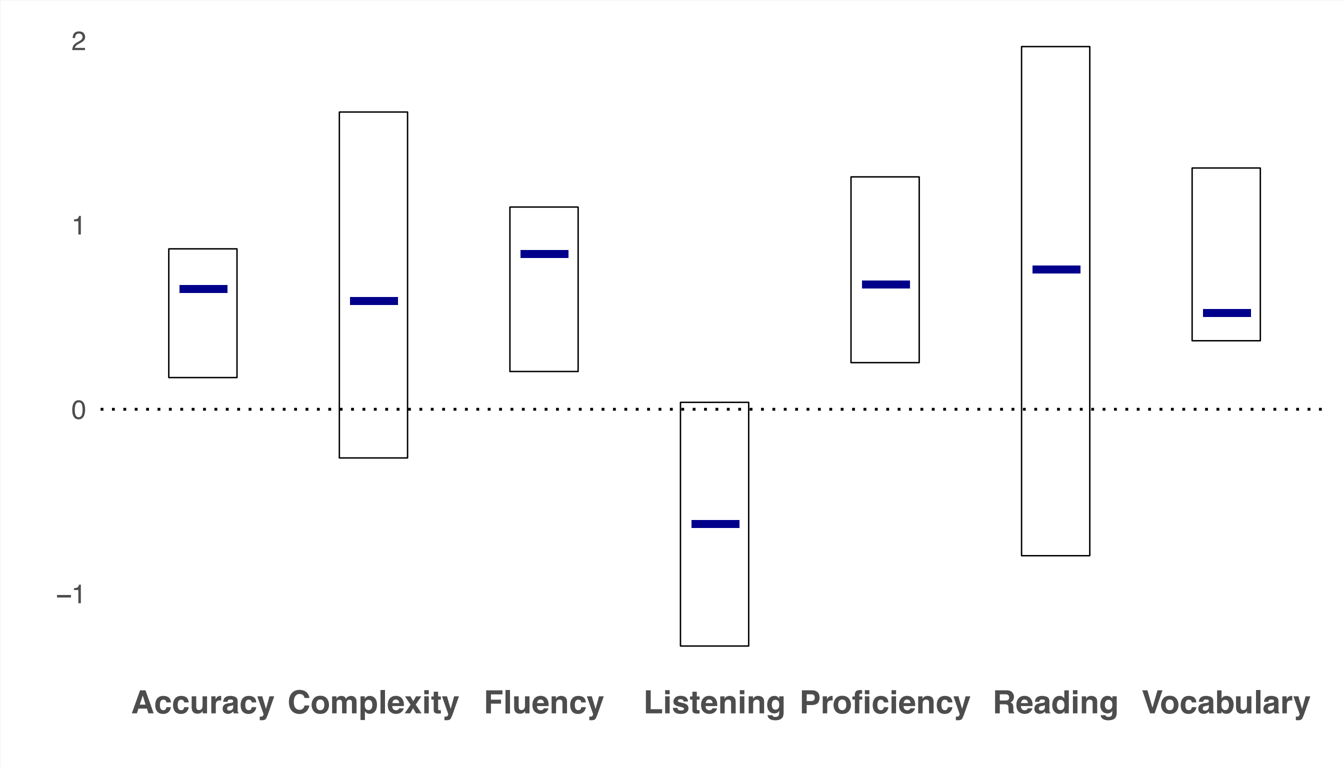


Moderator analysis
Outcome modality

Higher effect on speaking

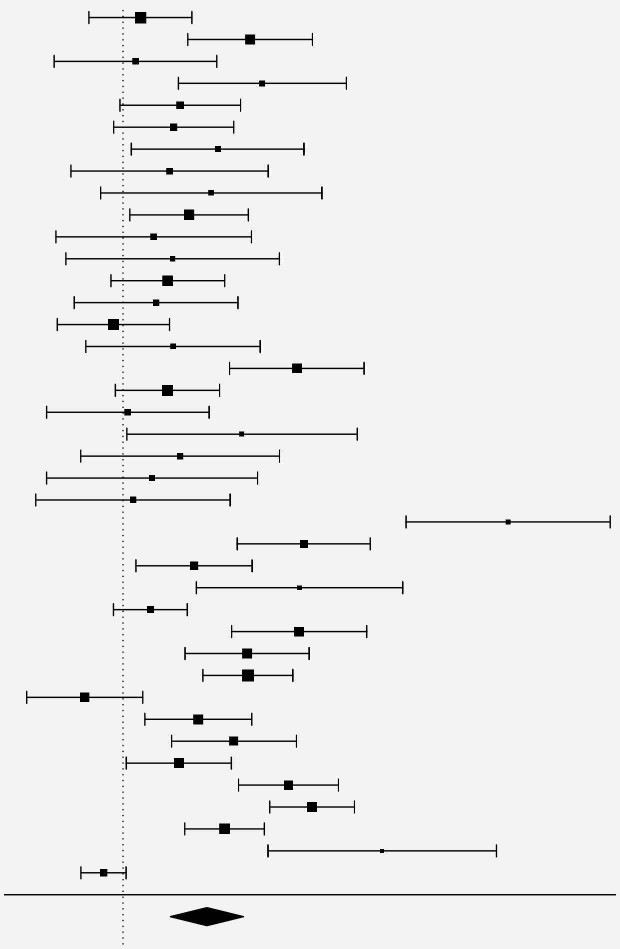


Moderator analysis
Outcome variables



More promising effects on **fluency**

Dialogue-based CALL: meta-analysis
Summary



Medium effect of dialogue-based CALL on L2 proficiency development
d = .605 ***

Possibly differentiated effect depending on **proficiency level**, **system modality** & **test modality**
But these observations still need to be confirmed by other studies

Need for more **comparable designs**, big enough **samples** and **precise instruments**
Future research should inscribe itself in this emerging field and compare its results within the field

Thank you! Merci! Dank u! ¡Gracias!

Serge **Bibauw**

serge.bibauw@kuleuven.be

Wim **Van den Noortgate**

wim.vandennoortgate@kuleuven.be

Thomas **François**

thomas.francois@uclouvain.be

Piet **Desmet**

piet.desmet@kuleuven.be