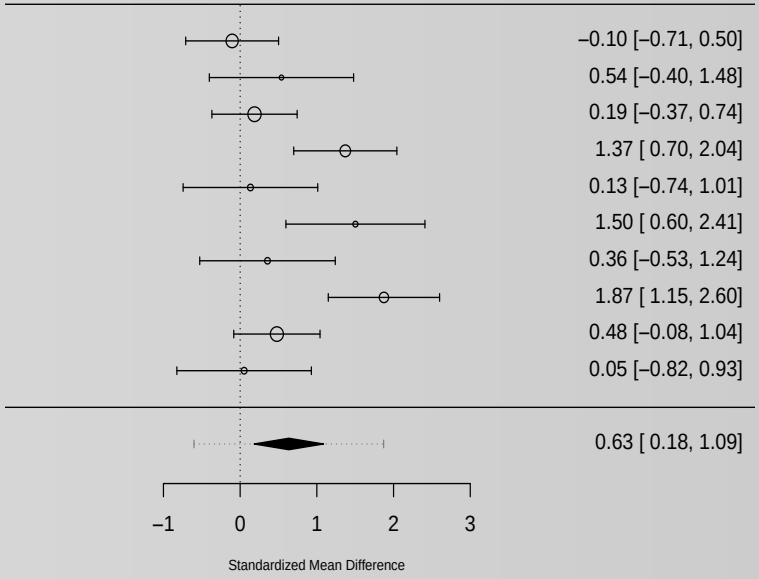
Effectiveness of **dialogue-based CALL** on L2 proficiency development: a **meta-analysis**



Serge **Bibauw**

Thomas **François**

Piet **Desmet**

CALICO Conference
May 18, 2017

KU LEUVEN · imec · UCL Université catholique de Louvain

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, *in prep.*

Dialogue-based CALL
A recent example

Duolingo Bots
(Oct. 2016)

Dialogue-based CALL
Three main types of systems

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
see Chatbots

Bibauw, François & Desmet, *in prep*

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.*

Effectiveness of **dialogue-based CALL** on L2 proficiency development: a **meta-analysis**

Methodology of meta-analysis

Corpus collection, inclusion/exclusion, effect sizes calculation and multilevel modeling

Main results

A random effects multilevel model to summarize the effect of dialogue-based CALL

Moderator variables

Moderators analysis: effects of specific design choices and experimental context

Effectiveness of **dialogue-based CALL** on L2 proficiency development: a **meta-analysis**

Methodology of meta-analysis

Corpus collection, inclusion/exclusion, effect sizes calculation and multilevel modeling

Main results

A random effects multilevel model to summarize the effect of dialogue-based CALL

Moderator variables

Moderators analysis: effects of specific design choices and experimental context

Corpus collection
Search methodology

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)
   ```

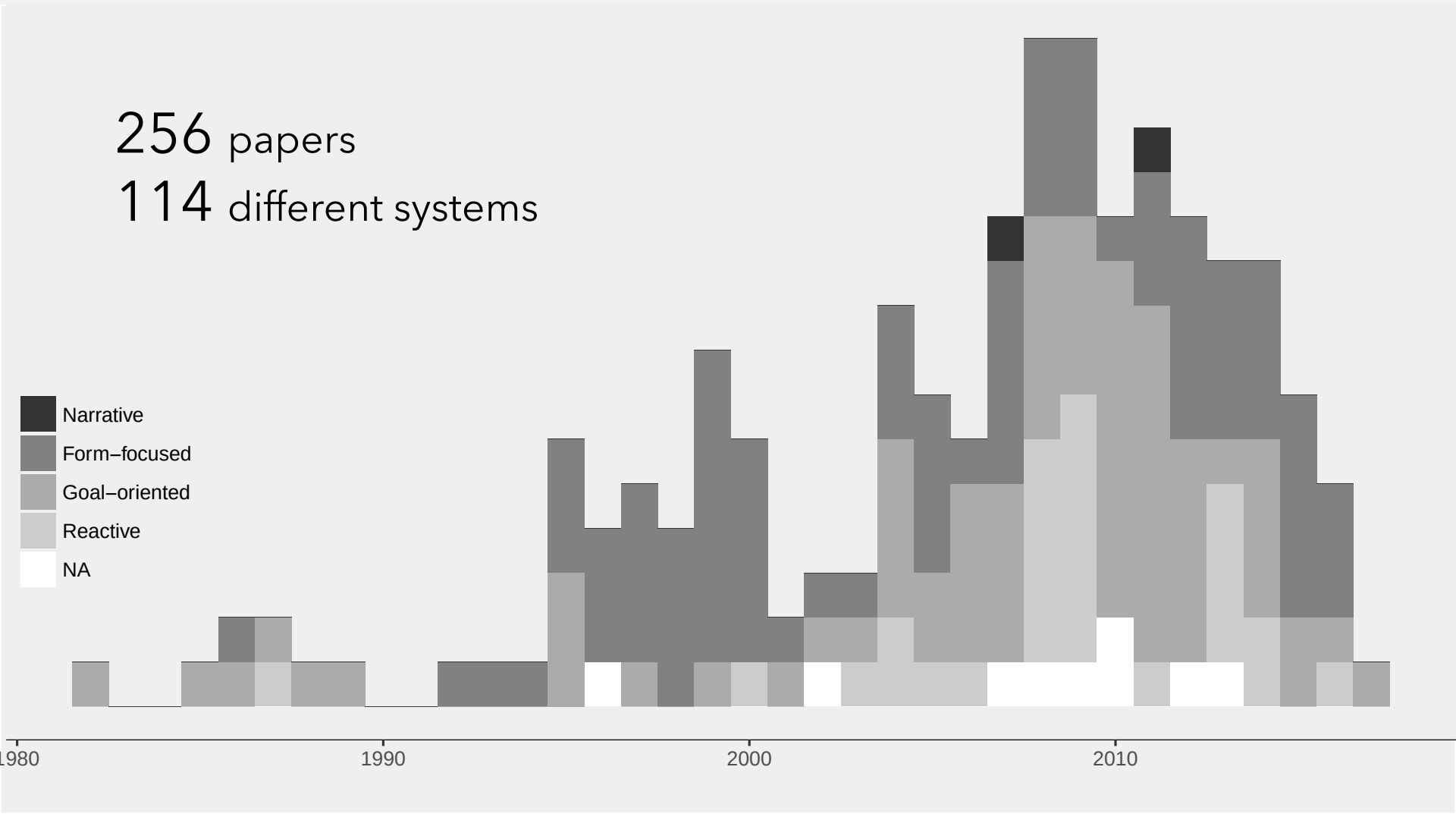
2. **Ancestry** search
   Older publications cited by ref

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

Note on journal search: 32/183 publications from the 4 major CALL journals (13 *CALL*, 12 *CALICO J.*, 4 *ReCALL*, 3 *LL&T*)

Corpus of studies

256 papers
114 different systems



Corpus collection
Domain definition and inclusion criteria

Based on our **operational definition of dialogue-based CALL** (**dialogue**, as a **task**/main activity, with a system/**computer agent**, for **language learning** purposes)

Peer-reviewed publications (journal articles, conference papers, book chapters, dissertations) only

⇒ 183 papers

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

Study identification

author, year, team_id, sample_id, study_type...

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)

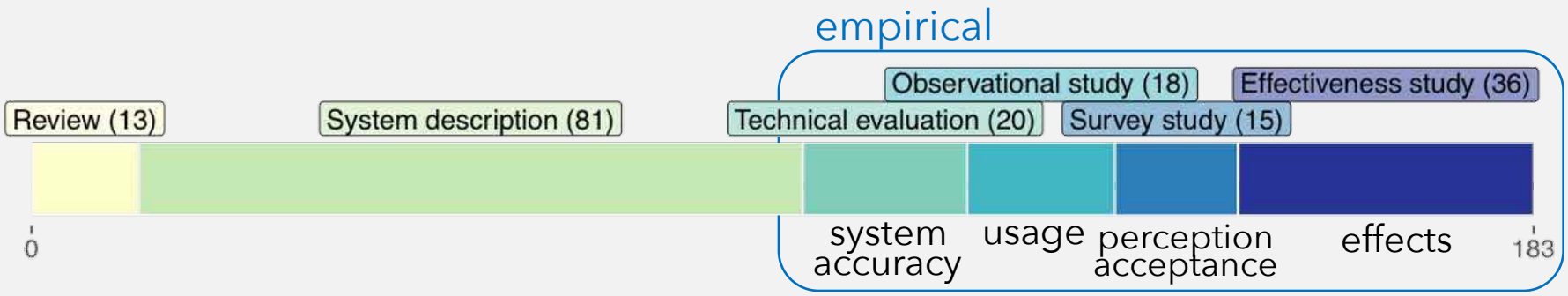
Instruments/outcome variables

proficiency/complexity/accuracy/fluency/vocabulary, speaking/writing, specific test

Quantitative results

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

Studies selection
Effectiveness studies



Effectiveness studies (36 papers)

- Experimental (or quasi-exp.) design
- At least two measurements (pre-post, experimental-control...)
- Motivational or cognitive effects measured
- Few experiments with a *real* control group

Studies selection
Computable effect sizes

Effect size: quantitative measure of the difference between two groups

Typically needs

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

for each group/measurement point

Not available for all studies (especially older studies) (6 studies excluded)

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

Studies selection
Instruments

Language learning tests

- Speaking/writing tests, measuring holistic proficiency or some of its components (complexity, accuracy, fluency)
- Accuracy tests (grammar/syntax/morphology)
- Vocabulary tests

Excluding motivational effects (1 study)

Excluding non-related cognitive effects (1 study)

Excluding uncontrolled teacher-made exams (1 study)

Meta-analysis
Effect sizes computation

Z-scores
⇒ Allows to compare results across various measurements scales

Standardized Mean Difference (SMD)

$Mean_{post} - Mean_{pre} / SD_{pooled}$

Cohen's d

Hedges' g

$$\hat{g} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)SD_1^2 + (n_2-1)SD_2^2}{(n_{total}-2)}}} \times 1 - \left(\frac{3}{4(n_1-n_2)-9}\right)$$

When control group:

Controlled Standardized Mean Difference

$SMD_{experimental} - SMD_{control}$

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
Random effects modeling

Fixed effects (FE) vs. **Random effects** (RE)

- **FE** assumes a single common effect size among the studies
  (All variance is due to random and sampling errors)
- **RE** assumes an inherent variance between studies

Considering the variability among systems and outcome measures studied:
Random effects model

- Confirmed by heterogeneity test
  ($Q = 246$, $p < .0001$)

Meta-analysis
Multilevel modeling

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

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

Multilevel modelling

Here, 3 levels: **team / sample / study**

K = 11 21 40

Weights accordingly, as dependent measures

Effectiveness of **dialogue-based CALL** on L2 proficiency development: a **meta-analysis**

Methodology of meta-analysis

Corpus collection, inclusion/exclusion, effect sizes calculation and multilevel modeling

Main results

A random effects multilevel model to summarize the effect of dialogue-based CALL

Moderator variables

Moderators analysis: effects of specific design choices and experimental context

Results
Summary effect

Within-subjects (pre-post) (k = 40):

d = 0.904 *** (within-subjects)
95% CI = [0.511, 1.298]

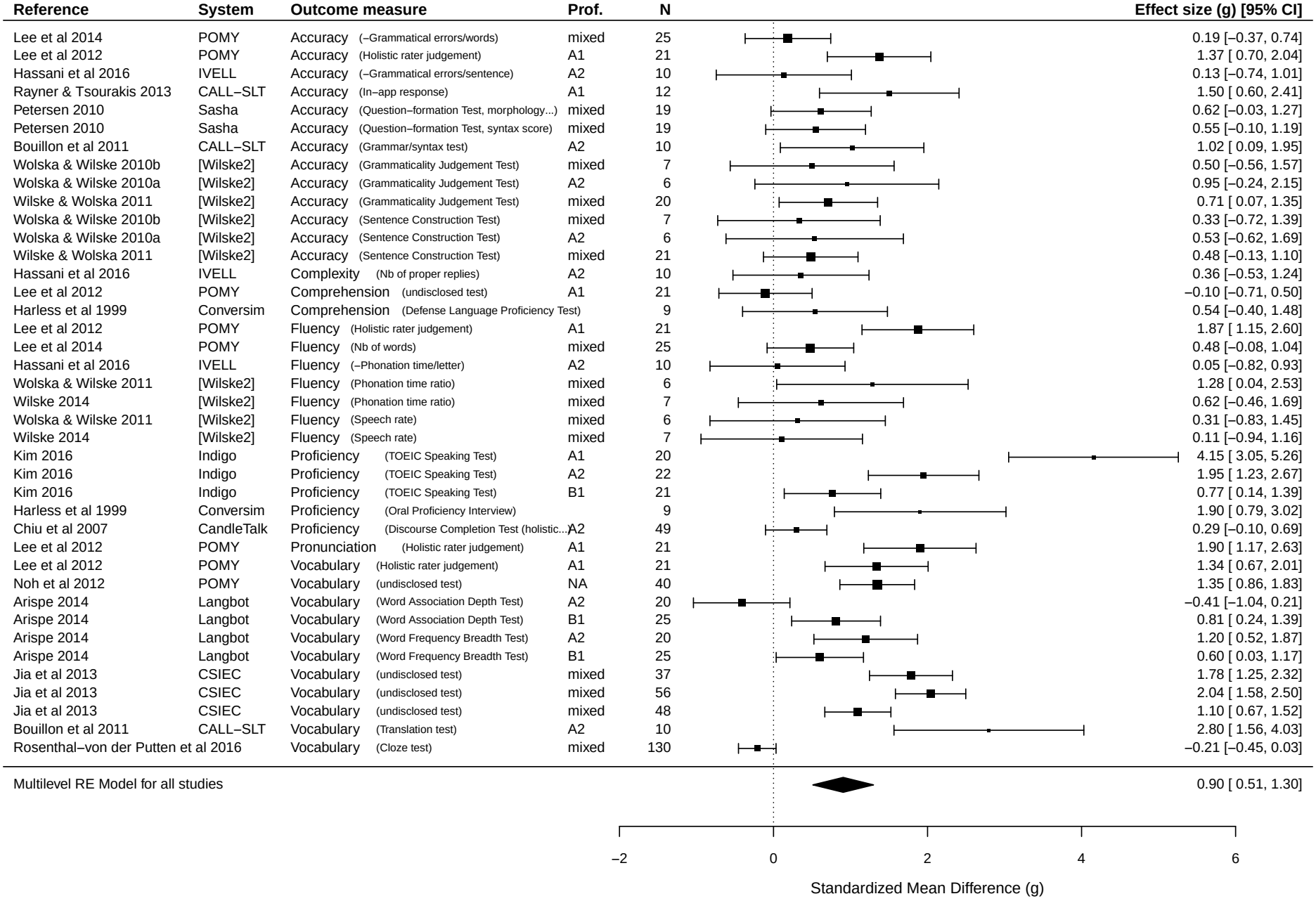
= Large effect (Cohen's "rule of thumb")
= Medium effect (Plonsky & Oswald, 2014, AL/SLA field-specific scale)

Between-subjects (pre-post – control) (k = 12):

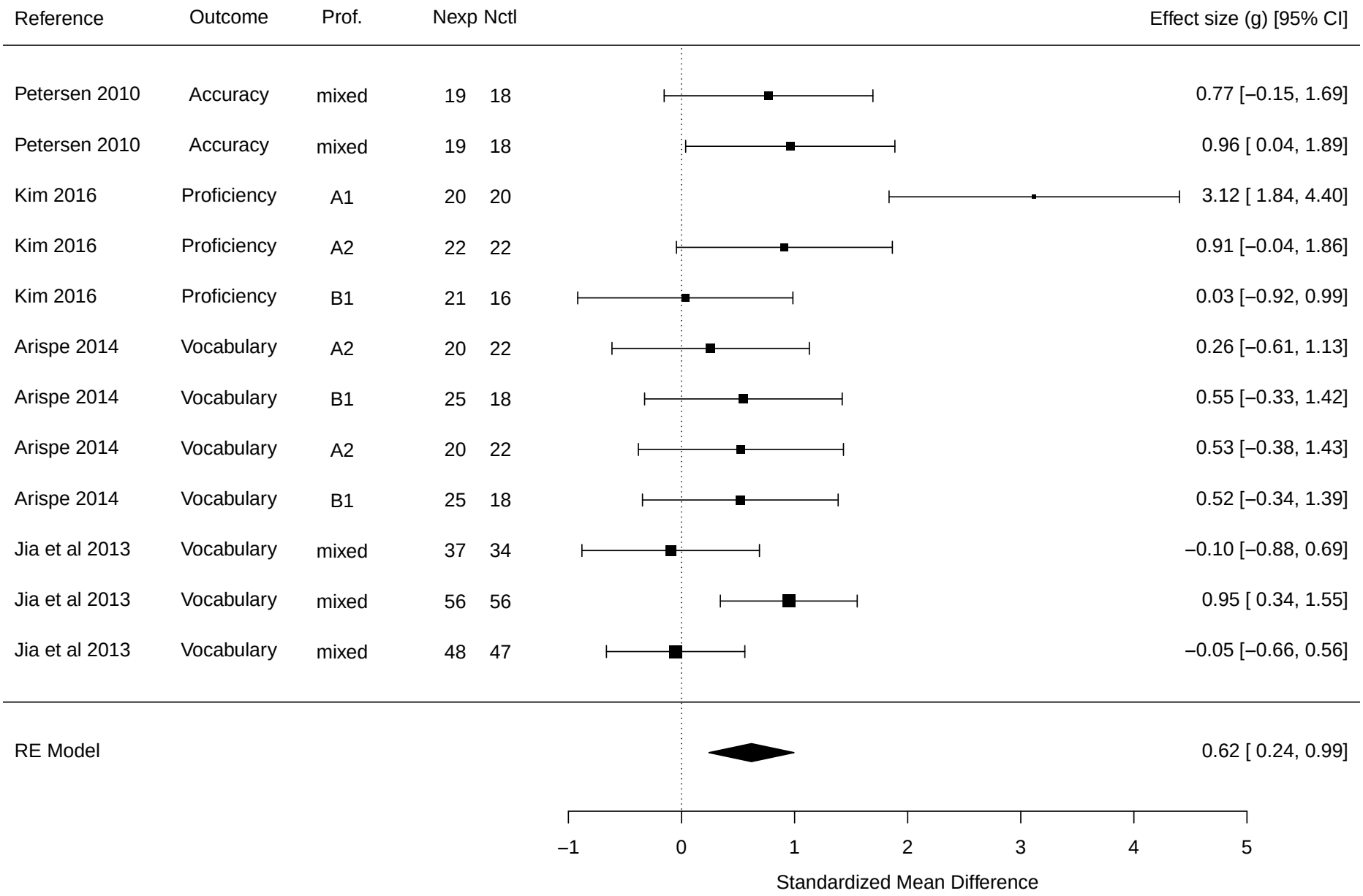
d = 0.618 **
95% CI = [0.243, 0.995]

= Medium effect (Plonsky & Oswald, 2014)

Within-subjects Forest plot

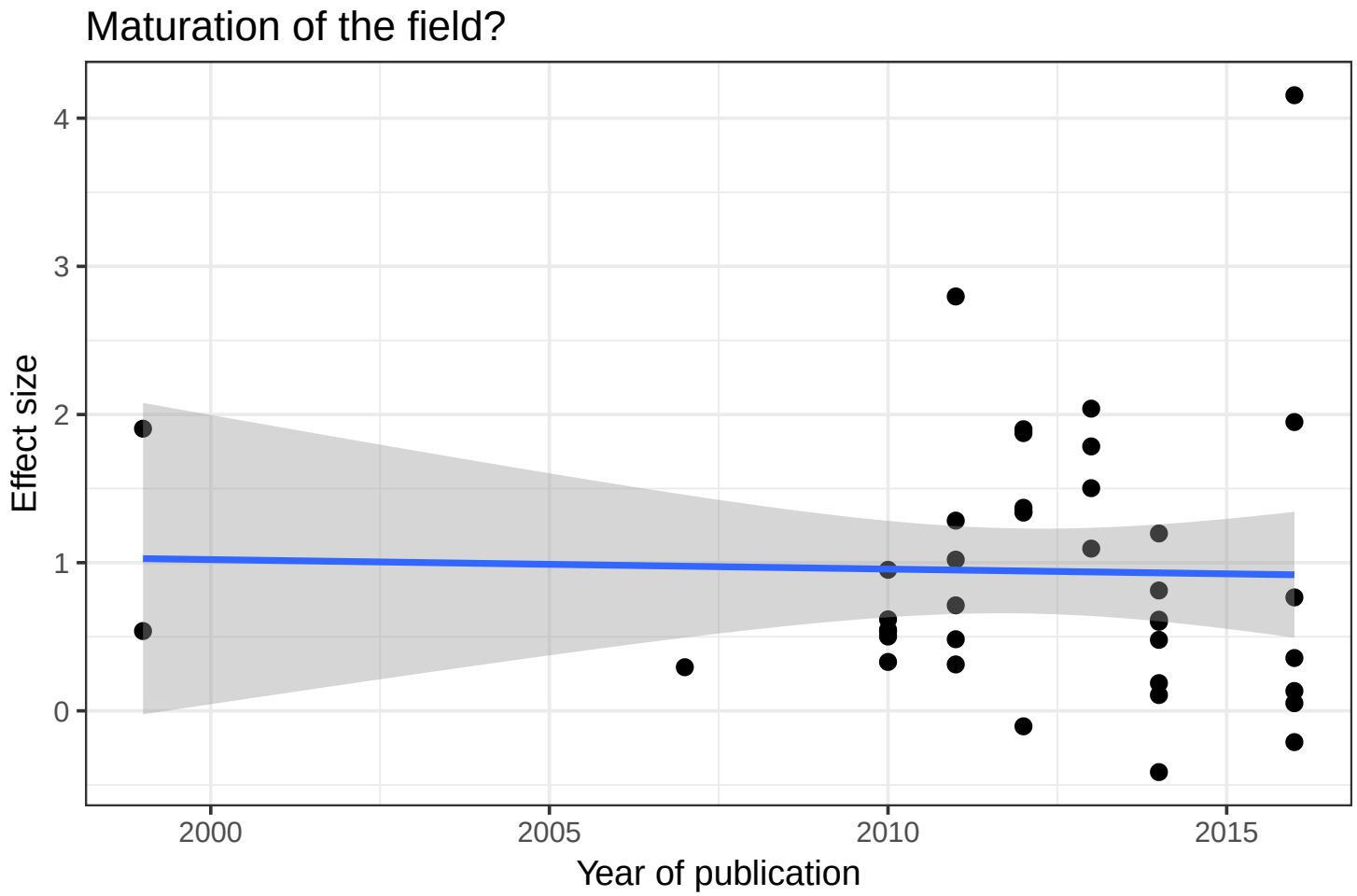


Between-subjects Forest plot

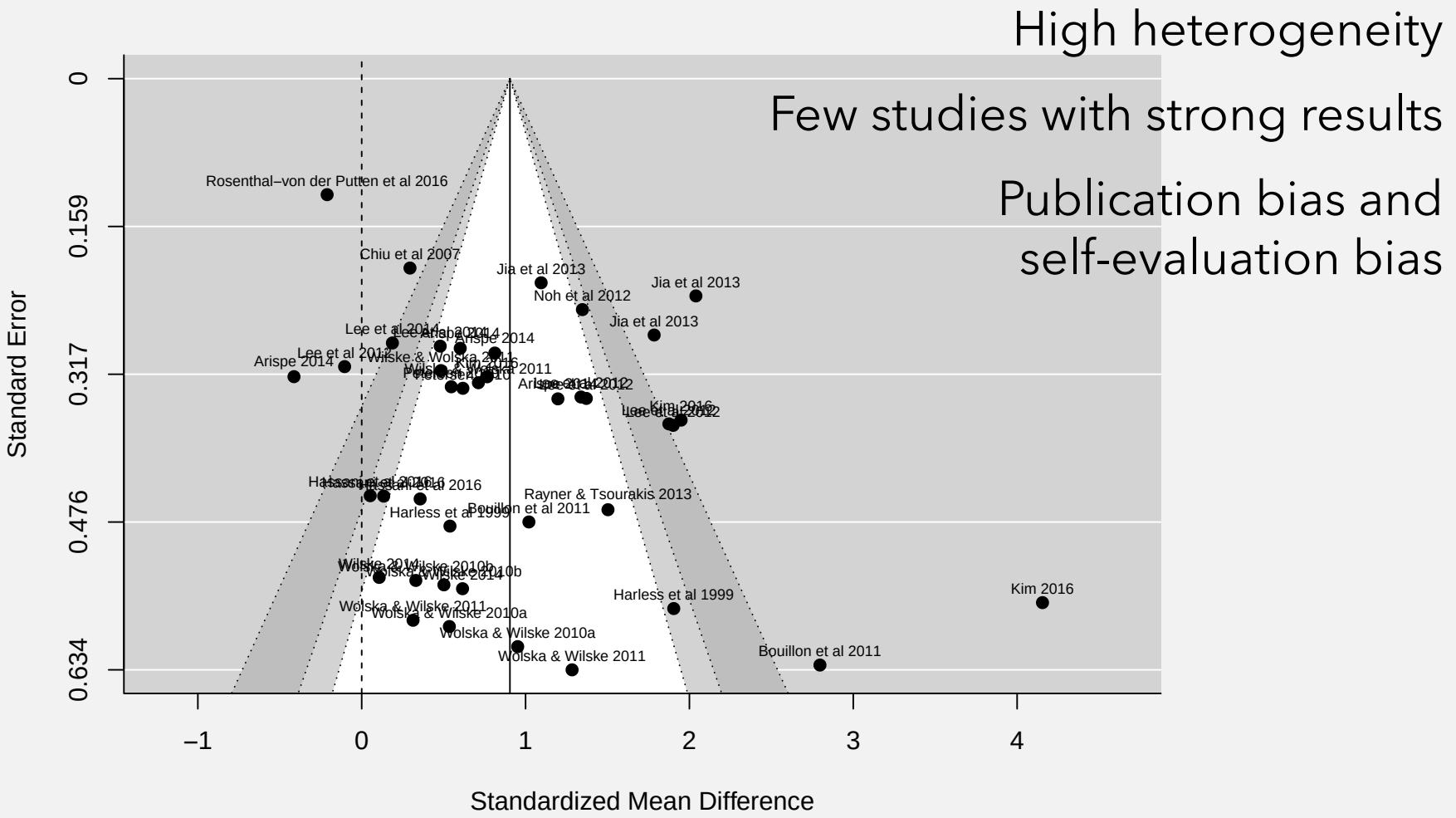


Results

Evolution across time



Discussion
Limitations



High heterogeneity

Few studies with strong results

Publication bias and self-evaluation bias

Effectiveness of **dialogue-based CALL** on L2 proficiency development: a **meta-analysis**

Methodology of meta-analysis

Corpus collection, inclusion/exclusion, effect sizes calculation and multilevel modeling

Main results

A random effects multilevel model to summarize the effect of dialogue-based CALL

Moderator variables

Moderators analysis: effects of specific design choices and experimental context

Moderators analysis

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

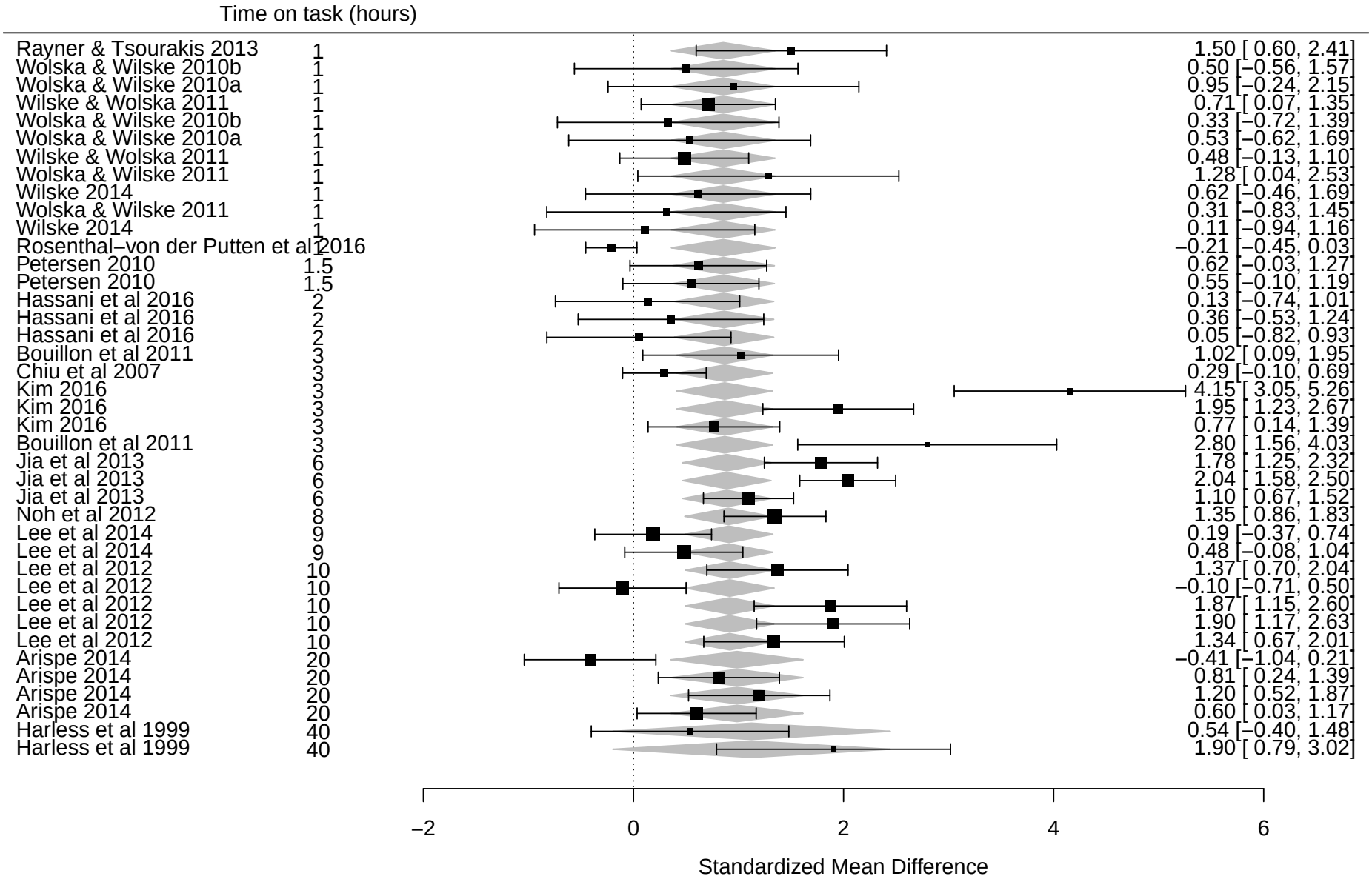
Sample and context
context, age, L1, L2, proficiency level

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

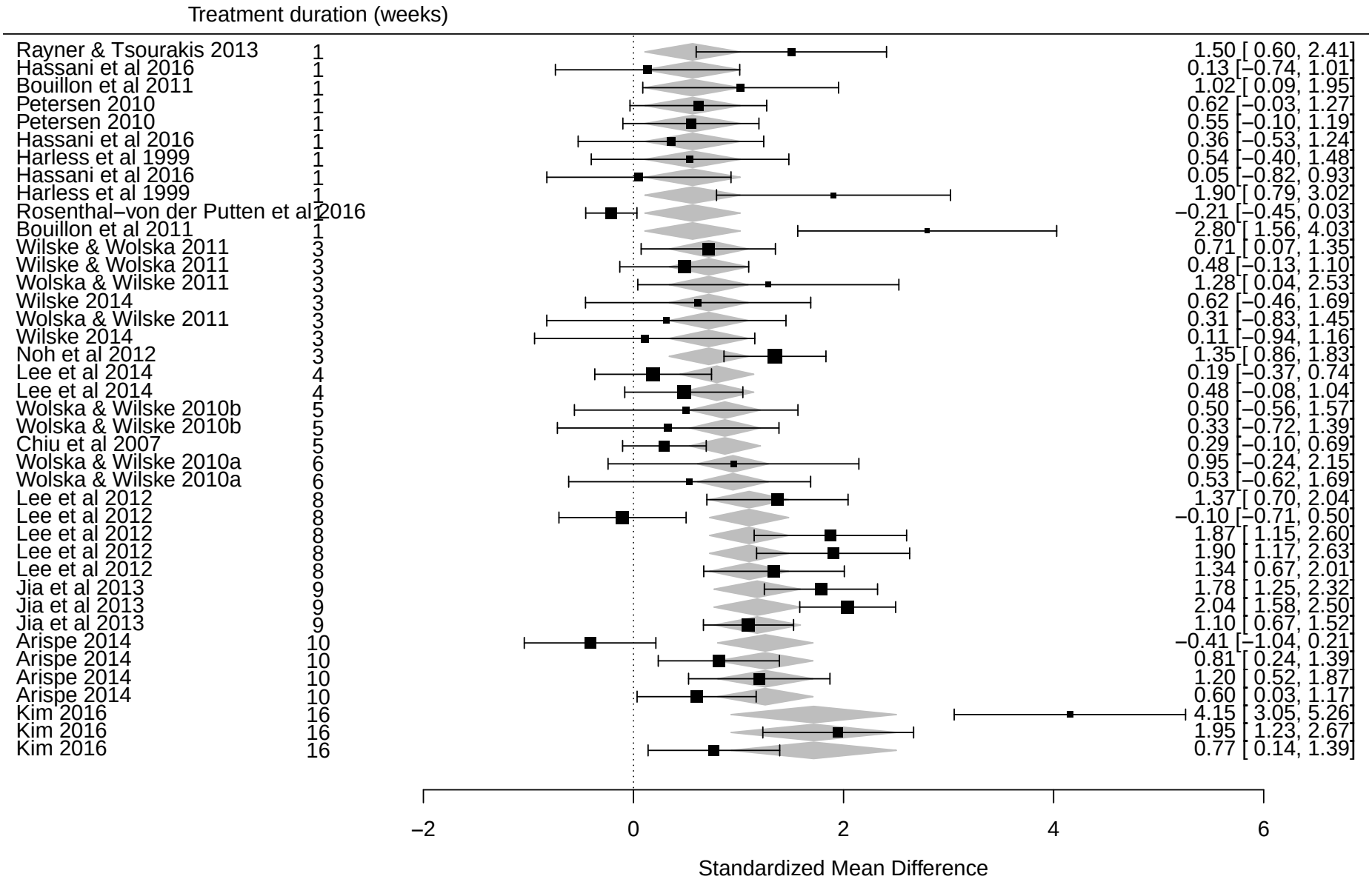
Time on task

$d = +0.007/_{\text{hour}}$ (non significant)



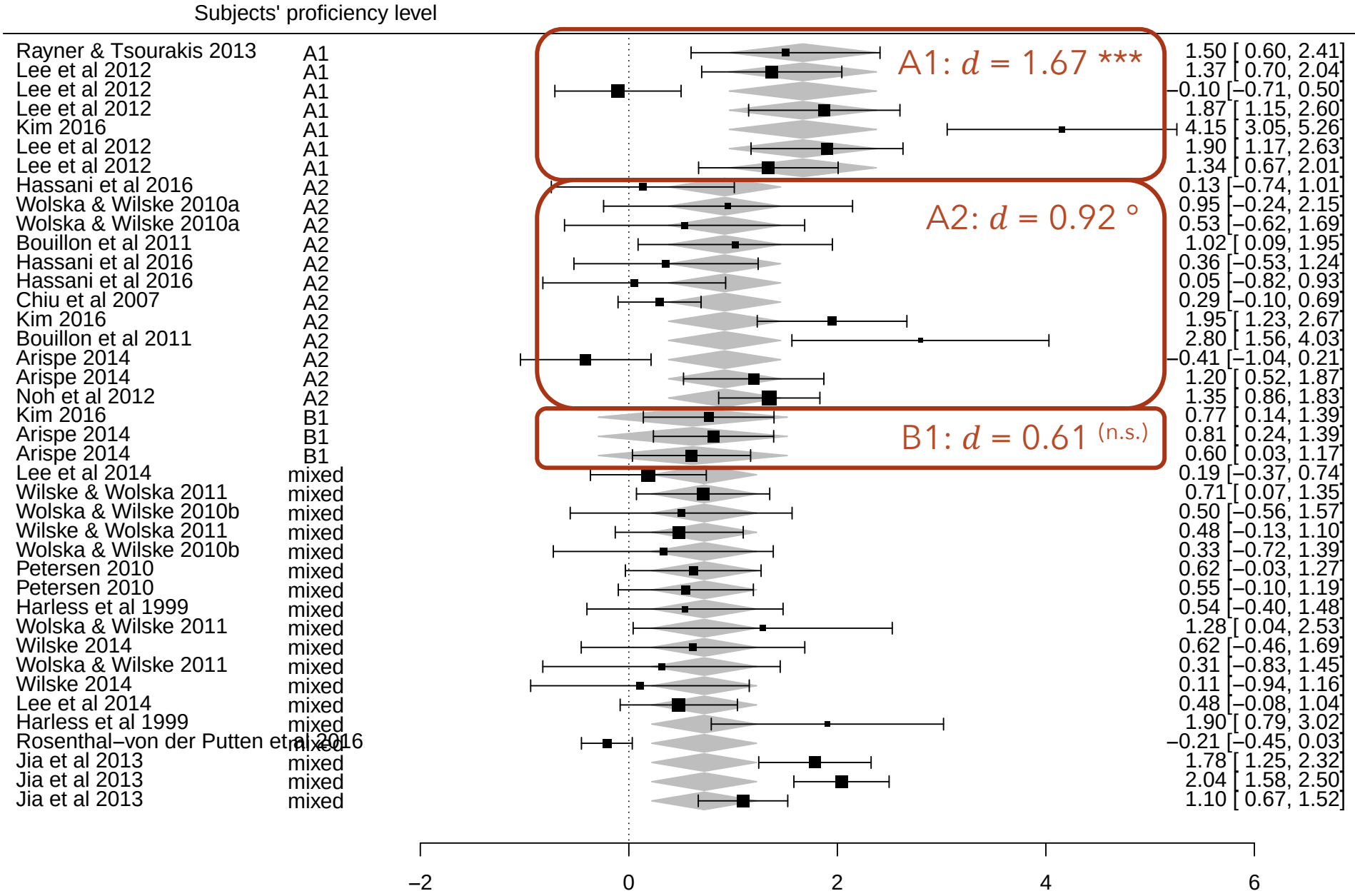
Treatment duration (weeks)

$d = +0.077/_{\text{week}}$ * (p = 0.026)



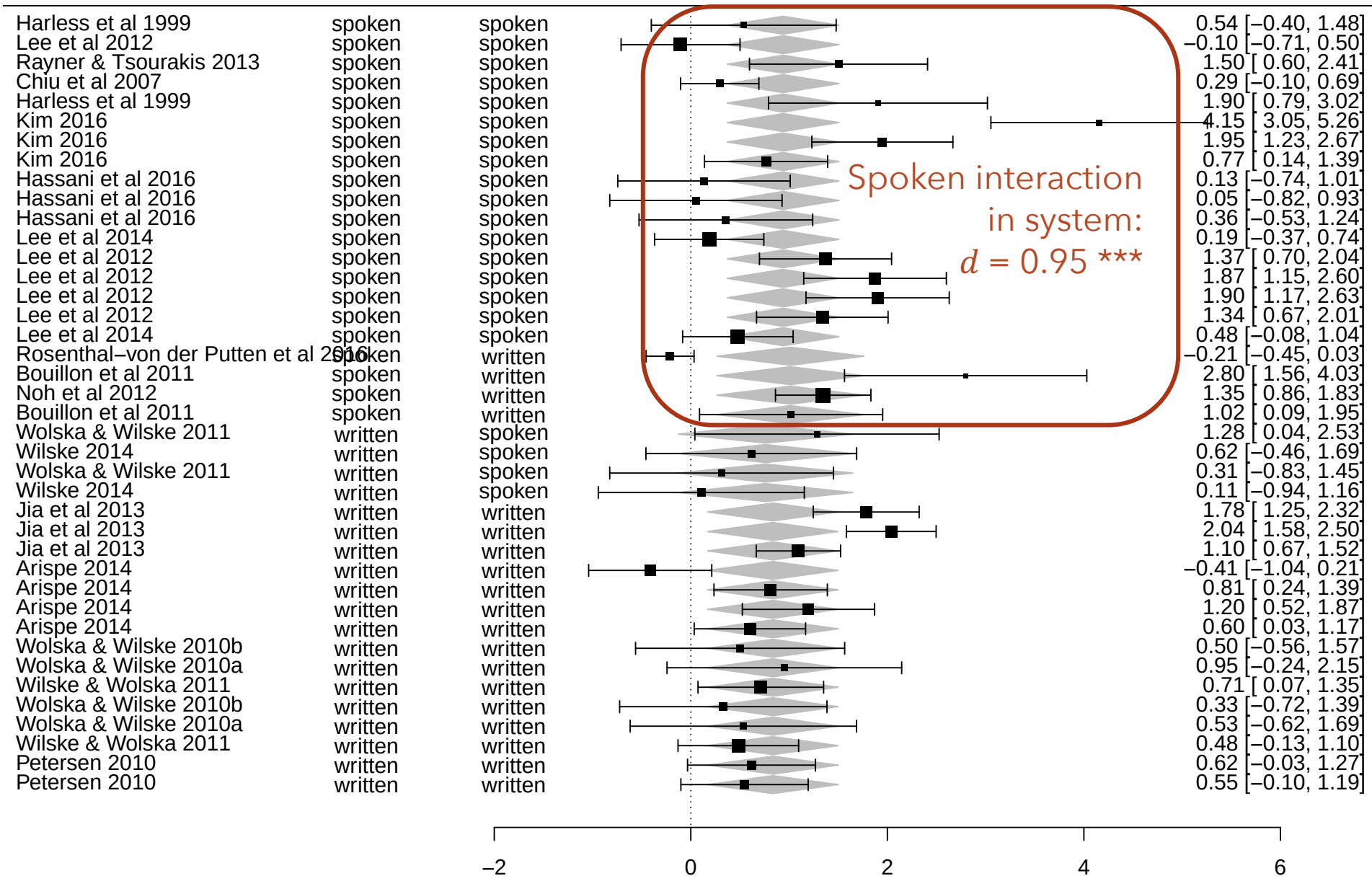
Learners' proficiency level

Test of Moderators:
$QM^{(df=3)} = 7.098$, $p = 0.069$



System & test modality

Test of Moderators (coefficient(s) 2):
$QM^{(df=1)} = 0.086$, $p = 0.769$



Moderators
Other moderators/covariates

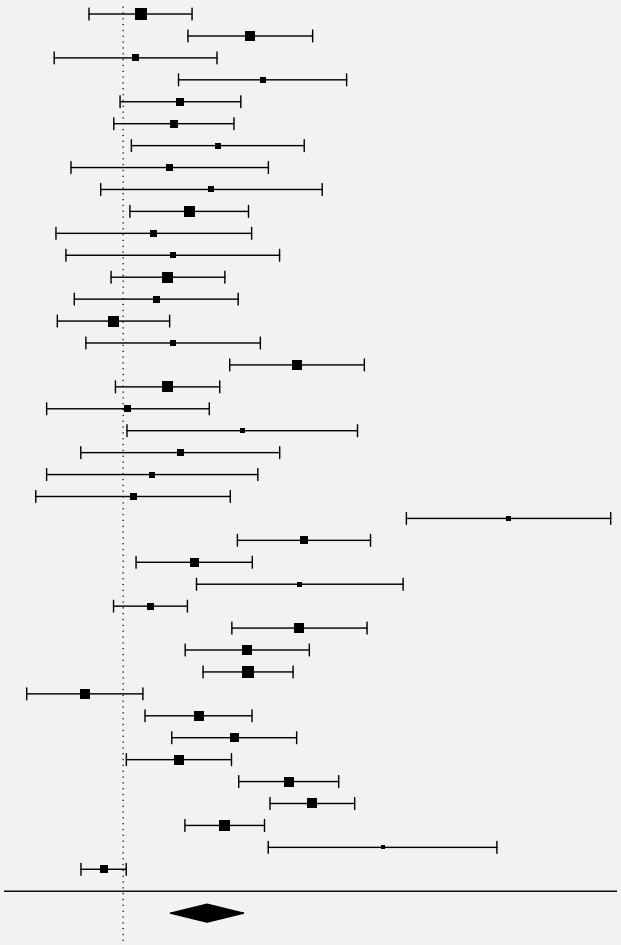
Learners variables: L1, age, context...
⇒ non significant

Instruments/outcomes: outcome measure group (accuracy/complexity/fluency/vocabulary)...
⇒ non significant

System characteristics: target language, dialogue type, system type, initiative, embodied agent, gamified...
⇒ non significant

Most likely due to multiple cases/coefficients and too few studies to establish any significance

Dialogue-based CALL: meta-analysis
Summary



Medium effect of dialogue-based CALL on L2 proficiency development

d = 0.90 (within) / d = 0.62 (between)

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

Thomas **François**
thomas.francois@uclouvain.be

Piet **Desmet**
piet.desmet@kuleuven.be

Download this presentation and the **full dataset** at
http://serge.bibauw.be/calico

KU LEUVEN itec imec UCL Université catholique de Louvain CENTAL centre de traitement automatique du langage