

# “Hey Siri, can I learn English by talking to you?”

## A meta-analysis of dialogue-based CALL

Serge BIBAUW<sup>1,2,4</sup>, with Wim VAN DEN NOORTGATE<sup>1</sup>, Thomas FRANÇOIS<sup>2,3</sup> and Piet DESMET<sup>1</sup>

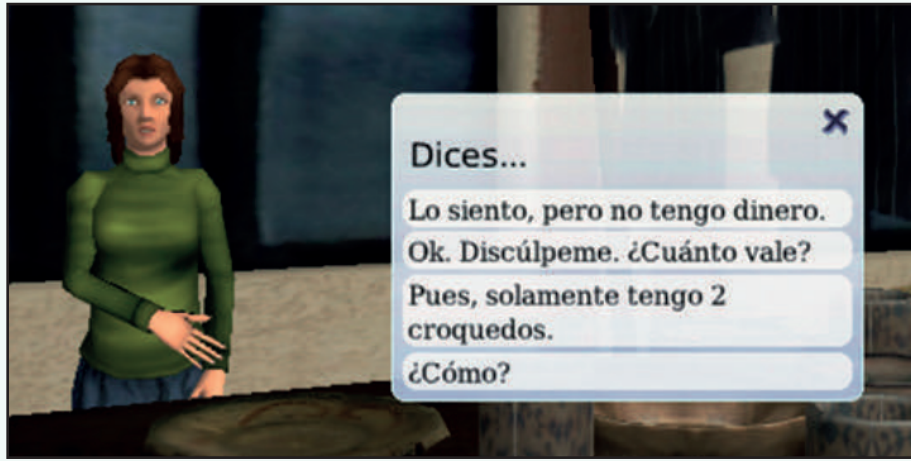
<sup>1</sup>ITEC, KU LEUVEN, ALSO AT IMEC <sup>2</sup>CENTAL, ILC/PLIN, UNIVERSITÉ CATHOLIQUE DE LOUVAIN <sup>3</sup>FONDS DE LA RECHERCHE SCIENTIFIQUE - FNRS <sup>4</sup>UNIVERSIDAD CENTRAL DEL ECUADOR

### Dialogue-based CALL

Dialogue-based CALL systems involve (Bibauw, François & Desmet, *forthcoming*)

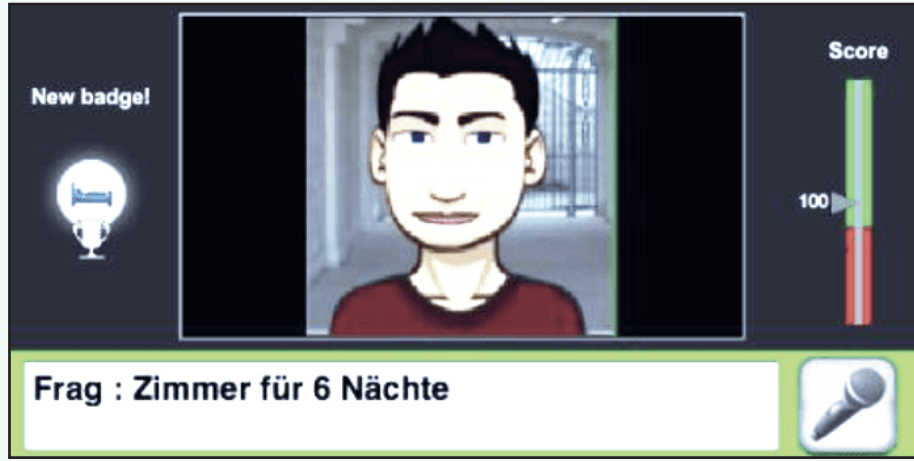
- a **dialogue** (i.e., sequence of conversational turns)
- with an **automated agent** (chatbot, robot, voice assistant, non player character...)
- as a **language learning task** (=scaffolding).

#### Narrative systems



*Croquelandia (Sykes, 2013)*

#### Form-focused systems



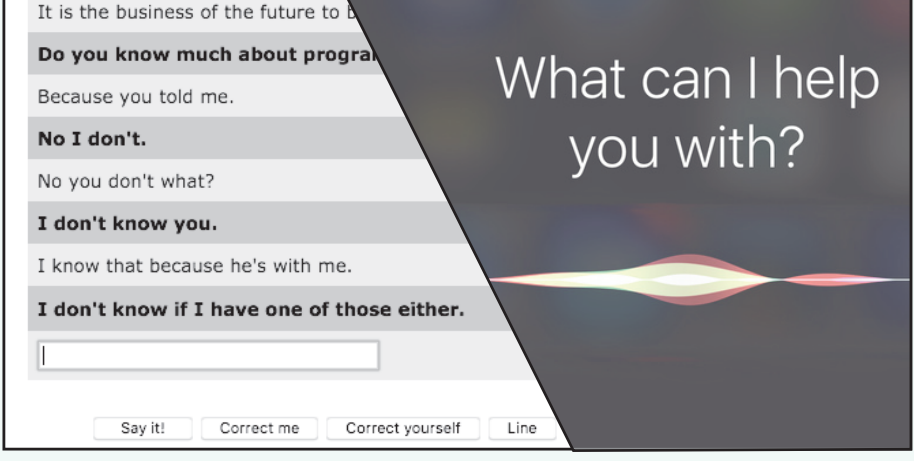
*CALL-SLIT (Baur et al., 2014)*

#### Goal-oriented systems



*HALEF/ETS (Sydorenko et al., 2018)*

#### Reactive systems



*Jabberwacky (Fryer et al., 2017); Siri (Underwood, 2017)*

### Methods

Records identified through database searching:

- 153 Scopus
- 75 Web of Science
- 68 Inspec
- 38 PsycINFO
- 38 LLBA
- 36 ERIC
- (...)

Additional records identified through forward and ancestry search: 193 records

419 records screened after removing duplicates

33 excluded (full-text unavailable, replications, other languages)

386 articles undergo full-text review

136 non "dialogue-based CALL" excluded (no dialogue with agent as task for language learning)

250 articles relevant to dialogue-based CALL

211 non experimental studies excluded (no empirical data, or only observational or survey data)

39 articles/studies included

Isolated effect sizes per sample and per outcome variable

k = 134 reported effect sizes

36 excluded (no measure of central tendency or variance, or effect on other outcomes.)

k = 96 effect sizes included

Standardized Mean Difference (d) computed with single raw metric (Morris & DeShon, 2002):

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

**Meta-analysis:** statistical summary of studies, aggregat-ing all compatible effects to compute a summary effect.

**Multilevel meta-analysis**

- every measurement of effect on each outcome variable for each sample is included as a single effect size;
- lack of independence between effects from the same study taken into account by layer of random variation at the study level;
- allows high granularity in study of moderator variables.

(see Van den Noortgate et al 2012)

| Level of aggregation | Items/clusters            | Remaining variation        |
|----------------------|---------------------------|----------------------------|
| Study                | k <sub>studies</sub> = 17 | Variation between-studies  |
| Effect size          | k = 96                    | Variation between-subjects |
| Subject              | n = 803                   | Random sampling variance   |

**Mixed-effects model**

- random between-studies effect
- random between-subjects effect
- fixed effects for covariates and moderator variables

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

**RQ2** How 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?

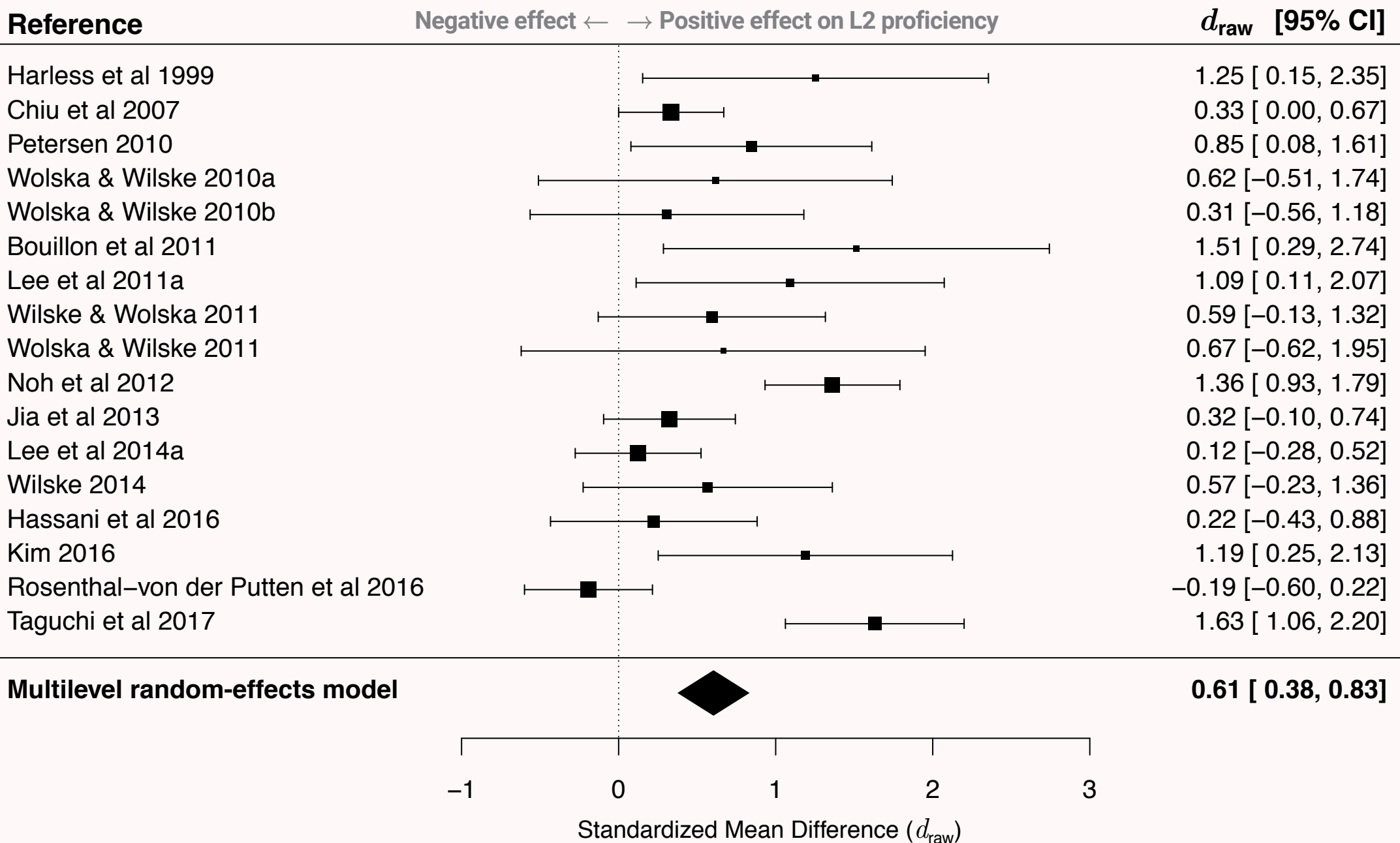
### Findings

Medium-high global effect of DBCALL on L2 proficiency:

$$d_{raw} = .61$$

High heterogeneity and limited statistical power in existing studies:

$$Q(df = 95) = 301.3$$



**Reference**

| Study                               | $d_{raw}$ [95% CI]  |
|-------------------------------------|---------------------|
| Harless et al 1999                  | 1.25 [ 0.15, 2.35]  |
| Chiu et al 2007                     | 0.33 [ 0.00, 0.67]  |
| Petersen 2010                       | 0.85 [ 0.08, 1.61]  |
| Wolska & Wilske 2010a               | 0.62 [-0.51, 1.74]  |
| Wolska & Wilske 2010b               | 0.31 [-0.56, 1.18]  |
| Bouillon et al 2011                 | 1.51 [ 0.29, 2.74]  |
| Lee et al 2011a                     | 1.09 [ 0.11, 2.07]  |
| Wilske & Wolska 2011                | 0.59 [-0.13, 1.32]  |
| Wolska & Wilske 2011                | 0.67 [-0.62, 1.95]  |
| Noh et al 2012                      | 1.36 [ 0.93, 1.79]  |
| Jia et al 2013                      | 0.32 [-0.10, 0.74]  |
| Lee et al 2014a                     | 0.12 [-0.28, 0.52]  |
| Wilske 2014                         | 0.57 [-0.23, 1.36]  |
| Hassani et al 2016                  | 0.22 [-0.43, 0.88]  |
| Kim 2016                            | 1.19 [ 0.25, 2.13]  |
| Rosenthal-von der Putten et al 2016 | -0.19 [-0.60, 0.22] |
| Taguchi et al 2017                  | 1.63 [ 1.06, 2.20]  |

**Multilevel random-effects model**

0.61 [ 0.38, 0.83]

### 15s-Summary

Dialogue-based CALL includes all chatbots, conversational agents, voice assistants, robots and talking NPCs for language learning.

We conducted a **multilevel meta-analysis** on all the effectiveness studies ever conducted on such systems (250 articles initially), collecting **96 effect sizes**.

**Innovative statistical formulas and models** were implemented to integrate the results.

The **general effect of dialogue-based CALL** practice on L2 proficiency development is **medium-high**, at  $d_{raw} = .61$ . It is comparable, although logically inferior, to the effect of human-human interaction as measured by other meta-analyses (Mackey & Goo, 2007:  $d_{raw} = .75$ ).

Insights from the moderator analysis include a differentiated effect **across proficiency levels** (beginners benefit more than advanced learners), and stronger effects on **production tasks**, particularly on **vocabulary and fluency** measures.

### Moderator analysis

**Differentiated effects across levels:** beginners tend to benefit more

**Tentative modelization of effects of treatment duration:** Time on task + #Session - Time between sessions

**Goal/task-oriented interaction** seems to provide more learning opportunities than **open-ended** (e.g., small talk) or **system-guided** interactions

Spoken and written practice seem to have very similar effects

...but effects could be slightly stronger or more visible on speaking

Learning effects are **much stronger on production outcomes**, and could be close to zero regarding an improvement in comprehension

All 4 CALF dimensions seem to benefit from DBCALL, but the effects seem **stronger on vocabulary & fluency** (and possibly complexity)

Effects are higher when tested through free or constrained production tasks than in other types of instruments

| Type                               | Variable        | df | F     | p     | Values               | k     | d     | SE   | CI                |
|------------------------------------|-----------------|----|-------|-------|----------------------|-------|-------|------|-------------------|
| Population                         | L2 proficiency* | 4  | 9.55  | .049  | intercept            | 4     | 0.69  | 0.38 | [-0.049, 1.436]   |
|                                    |                 |    |       |       | A1                   | 38    | 0.36  | 0.21 | [-0.056, 0.775]   |
|                                    |                 |    |       |       | A2                   | 89    | 0.18  | 0.30 | [-0.416, 0.769]   |
|                                    |                 |    |       |       | B1                   | 77    | -0.42 | 0.25 | [-0.910, 0.066]   |
|                                    |                 |    |       |       | B2                   | 28    | -0.41 | 0.28 | [-0.962, 0.150]   |
| Context                            |                 | 2  | 1.03  | .599  | school               | 18    | 0.68  | 0.23 | [0.235, 1.121]**  |
|                                    |                 |    |       |       | university           | 35    | 0.54  | 0.15 | [0.259, 0.830]*** |
|                                    |                 |    |       |       | military             | 3     | 1.08  | 0.55 | [0.002, 2.160]*   |
|                                    |                 |    |       |       | intercept            | 0.09  | 0.20  |      | [-0.300, 0.484]   |
|                                    |                 |    |       |       | +1 hour on task      | 0.15  | 0.05  |      | [0.049, 0.256]**  |
| Treatment                          | Duration*       | 4  | 10.29 | .036  | +1 session           | 0.31  | 0.11  |      | [0.094, 0.523]**  |
|                                    |                 |    |       |       | +1 week              | -0.19 | 0.08  |      | [-0.338, -0.037]* |
|                                    |                 |    |       |       | narrative            | 4     | 0.31  | 0.49 | [-0.643, 1.261]   |
|                                    |                 |    |       |       | form-focused         | 15    | 0.86  | 0.27 | [0.336, 1.392]**  |
|                                    |                 |    |       |       | goal-oriented        | 71    | 0.56  | 0.16 | [0.244, 0.877]*** |
| Type of system                     |                 | 3  | 1.38  | .710  | reactive             | 6     | 0.57  | 0.37 | [-0.156, 1.287]   |
|                                    |                 |    |       |       | goal-oriented        | 66    | 0.64  | 0.14 | [0.373, 0.907]*** |
|                                    |                 |    |       |       | open-ended           | 6     | 0.56  | 0.36 | [-0.146, 1.276]   |
|                                    |                 |    |       |       | system-guided        | 4     | 0.31  | 0.48 | [-0.627, 1.245]   |
|                                    |                 |    |       |       | spoken               | 25    | 0.59  | 0.17 | [0.256, 0.920]*** |
| System modality                    |                 | 1  | 0.03  | .873  | written              | 61    | 0.63  | 0.17 | [0.293, 0.960]*** |
|                                    |                 |    |       |       | explicit             | 36    | 0.75  | 0.16 | [0.447, 1.059]*** |
|                                    |                 |    |       |       | implicit             | 37    | 0.71  | 0.15 | [0.415, 1.005]*** |
|                                    |                 |    |       |       | none                 | 23    | 0.37  | 0.18 | [0.013, 0.732]*   |
|                                    |                 |    |       |       | spoken               | 35    | 0.74  | 0.16 | [0.427, 1.054]*** |
| Corrective feedback                |                 | 2  | 2.53  | .283  | written              | 61    | 0.52  | 0.14 | [0.249, 0.799]*** |
|                                    |                 |    |       |       | true                 | 72    | 0.68  | 0.13 | [0.428, 0.923]*** |
|                                    |                 |    |       |       | false                | 24    | 0.40  | 0.17 | [0.063, 0.745]*   |
|                                    |                 |    |       |       | comprehension        | 4     | -0.45 | 0.33 | [-1.095, 0.201]   |
|                                    |                 |    |       |       | production           | 80    | 0.76  | 0.16 | [0.453, 1.069]*** |
| Outcome                            | Test modality   | 1  | 1.72  | .190  | vocabulary           | 12    | 0.41  | 0.25 | [-0.083, 0.899]   |
|                                    |                 |    |       |       | holistic proficiency | 11    | 0.76  | 0.26 | [0.263, 1.265]**  |
|                                    |                 |    |       |       | complexity           | 1     | 0.68  | 0.48 | [-0.262, 1.614]   |
|                                    |                 |    |       |       | accuracy             | 49    | 0.52  | 0.18 | [0.176, 0.871]**  |
|                                    |                 |    |       |       | lexicon              | 17    | 0.83  | 0.23 | [0.375, 1.292]*** |
| Matching modality (treatment=test) |                 | 1  | 2.52  | .113  | fluency              | 14    | 0.65  | 0.23 | [0.207, 1.097]**  |
|                                    |                 |    |       |       | metaling. judgment   | 20    | 0.58  | 0.20 | [0.184, 0.969]**  |
|                                    |                 |    |       |       | selected response    | 9     | 0.17  | 0.23 | [-0.280, 0.621]   |
|                                    |                 |    |       |       | constrained resp.    | 32    | 0.71  | 0.18 | [0.355, 1.064]*** |
|                                    |                 |    |       |       | free response        | 35    | 0.76  | 0.18 | [0.412, 1.109]*** |
| Outcome type***                    |                 | 2  | 16.32 | <.001 | short-term           | 73    | 0.62  | 0.12 | [0.388, 0.860]*** |
|                                    |                 |    |       |       | long-term            | 23    | 0.52  | 0.16 | [0.202, 0.838]**  |
| Outcome dimension**                |                 | 6  | 18.68 | .005  |                      |       |       |      |                   |
|                                    |                 |    |       |       |                      |       |       |      |                   |
|                                    |                 |    |       |       |                      |       |       |      |                   |
|                                    |                 |    |       |       |                      |       |       |      |                   |
|                                    |                 |    |       |       |                      |       |       |      |                   |
| Type of test                       |                 | 3  | 7.75  | .051  |                      |       |       |      |                   |
|                                    |                 |    |       |       |                      |       |       |      |                   |
|                                    |                 |    |       |       |                      |       |       |      |                   |
|                                    |                 |    |       |       |                      |       |       |      |                   |
|                                    |                 |    |       |       |                      |       |       |      |                   |
| Temporality of effects             |                 | 1  | 0.60  | .439  |                      |       |       |      |                   |
|                                    |                 |    |       |       |                      |       |       |      |                   |



Download this poster and obtain more info

serge.bibauw@kuleuven.be

KU LEUVEN

UCL  
Université catholique de Louvain

UNIVERSIDAD CENTRAL DEL ECUADOR