Investigating Object Orientation Effects Across 18 Languages

Sau-Chin Chen[1]

[1] Department of Human Development and Psychology
Tzu-Chi University
Hualien
Taiwan

Author Note

Funding statement. Glenn P. Williams was supported by the Leverhulme Trust Research Project Grant (RPG-2016-093). Krystian Barzykowski was supported by the National Science Centre, Poland (2019/35/B/HS6/00528). Zoltan Kekecs was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Science. Patrícia Arriaga was supported by the Portuguese National Foundation for Science and Technology (UID/PSI/03125/2019). Gabriel Baník was supported by Charles University Grant Agency (PRIMUS/20/HUM/009).

Ethical approval statement. Authors who collected data on site and online had the ethical approval/agreement from their local institutions. The latest status of ethical approval for all the participating authors is available at the public OSF folder (https://osf.io/e428p/ "IRB approvals" in Files).

Acknowledgement. We appreciated the major contributions from the contributors as below. Chris Chartier and Jeremy Miller managed and monitored progress. Erin Buchanan provided guidelines to improve the inter-lab progress website management and managed the JATOS server for online data collection. Arti Parganiha, Asil Özdoğru, Attila Szuts, Babita Pande, Danilo Zambrano Ricaurte, Gabriel Baník, Harry Manley, Jonas Kunst, Krystian Barzykowski, Marco Antonio Correa Varella, Marietta Papadatou Pastou, Niv Reggev, Patrícia Arriaga, Stefan Stieger, Vanja Ković and Zahir Vally managed the material translation from English to the other languages. Roles of each collaborator are available in the public table (https://osf.io/mz97h/). We thank the suggestions from the editor and two reviewers on our first and second proposals.

The authors made the following contributions. Sau-Chin Chen: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing.

Correspondence concerning this article should be addressed to Sau-Chin Chen, No. 67, Jei-Ren St., Hualien City, Taiwan. E-mail: csc2009@mail.tcu.edu.tw

Abstract

Mental simulation theories of language comprehension propose that people automatically create mental representations of objects mentioned in sentences. Representation is often measured with the sentence-picture verification task, in which participants first read a sentence implying the shape/size/color/object orientation and, on the following screen, a picture of an object. Participants then verify if the pictured object either matched or mismatched the implied visual information mentioned in the sentence. Previous studies indicated the match advantages of shapes, but findings concerning object orientation were mixed across languages. This registered report describes our investigation of the match advantage of object orientation across 18 languages, which was undertaken by multiple laboratories and organized by the Psychological Science Accelerator. The preregistered analysis revealed that there is no compelling evidence for a global match advantage, although some evidence of a match advantage in one language was found. Additionally, the match advantage was not predicted by mental rotation scores which does not support current embodied cognition theories.

Keywords: language comprehension, mental simulation, object orientation, mental rotation, cross-lingual research

Word count: 5,138 words in total; Introduction: 1,242 words

Investigating Object Orientation Effects Across 18 Languages

Mental simulation of object properties is a major topic in conceptual processing research (Ostarek & Huettig, 2019; Scorolli, 2014). Theoretical frameworks of conceptual processing describe the integration of linguistic representations and situated simulation (e.g., reading about bicycles integrates the situation in which bicycles would be used, Barsalou, 2008; Zwaan, 2014). Proponents of situated cognition assume that perceptual representations are able to be generated during language processing. Recently, neuroimaging studies have explored and attempted to corroborate this hypothesis by examining the cortical activation patterns from seeing visual images and reading text (see the summary of Ostarek & Huettig, 2019).

One empirical index of situated simulation is the mental simulation effects measured in the sentence-picture verification task (see Figure 1). This task requires participants to read a probe sentence displayed on the screen. On the following screen, the participants see a picture of an object and must verify whether the object was mentioned in the probe sentence. Verification response times are operationalized as the mental simulation effect, which occurs when people are faster to verify pictured objects whose properties match those of objects implied in the probe sentences. For example, the eagle was moving through the air would be matched faster if an eagle was depicted flying, rather than stationary.

(Insert Figure 1 about here)

Mental simulation effects have been demonstrated for object shape (Zwaan et al., 2002), color (Connell, 2007), and orientation (Stanfield & Zwaan, 2001). Subsequent replication studies revealed consistent results for the shape but inconsistent findings for the color and orientation effects (De Koning et al., 2017; Rommers et al., 2013; Zwaan & Pecher, 2012), and the theoretical frameworks do not provide researchers much guidance regarding the potential causes for this discrepancy. With the accumulating concerns about the lack of reproducibility, researchers have found it challenging to update the theoretical framework in terms of mental simulation effects being unreplicable (e.g., Kaschak & Madden, 2021). Researchers who intended to improve the theoretical framework necessarily require a reproducible protocol for measuring mental simulation effects.

An additional facet of this research is the linguistic representations of object properties may play a role in the unreliability of the mental simulation effect. Mental simulation effects for object shape have consistently appeared in English (Zwaan et al., 2017; Zwaan & Madden, 2005; Zwaan & Pecher, 2012), Chinese (Li & Shang, 2017), Dutch (De Koning et al., 2017; Engelen et al., 2011; Pecher et al., 2009; Rommers et al., 2013), German (Koster et al., 2018), Croatian (Šetić & Domijan, 2017), and Japanese (Sato et al., 2013). Object orientation, on the other hand, has produced mixed results across languages: see positive evidence in English (Stanfield & Zwaan, 2001; Zwaan & Pecher, 2012) and Chinese (Chen et al., 2020), null evidence in Dutch (De Koning et al., 2017; Rommers et al., 2013), and German as second language (Koster et al., 2018). Among the studies of shape and orientation, the results indicated smaller effect sizes of object orientation than that of object shape (e.g., d = 0.10 vs. 0.17; in Zwaan and Pecher (2012); 0.07 vs. 0.27 in De Koning et al. (2017)). To understand the causes for the discrepancies among object properties and languages, it is imperative to consider the cross-linguistic and experimental factors of the sentence-picture verification task.

Cross-linguistic, Methodological, and Cognitive Factors

Several factors might contribute to cross-linguistic differences in the match advantage of orientation as a mental simulation effect, and we focused on context, methodological, and cognitive factors. Researchers have argued that languages differ in how they encode motion and placement events in sentences (Newman, 2002; Verkerk, 2014). In addition, the potential role of mental rotation as a confound has been considered (Rommers et al., 2013). We expand on how the context, experimental, and cognitive factors hinder the improvement of theoretical frameworks below.

Context Factors. The probe sentences used in object orientation studies usually contain several motion events (e.g., "The ant walked towards the pot of honey and tried to climb in."). The languages we probed in this study encode motion events in different ways, and grammatical differences between language encodings could explain different match advantage results. According to Verkerk (2014), Germanic languages (e.g., Dutch, English, German) generally encode the manner of motion in the verb (e.g., 'The ant dashed'), while conveying the path information through satellite adjuncts (e.g., 'towards the pot of honey'). In contrast, other languages, such as the Romance family (e.g., Portuguese, Spanish) more often encode path in the verb (e.g., 'crossing,' 'exiting'). Crucially, past research on the match advantage of object orientation is exclusively based on Germanic languages, and yet, there were differences across those languages, with English being the only one that consistently yielded the match advantage. As a minor difference across Germanic languages in this regard, Verkerk (2014) notes that path-only constructions (e.g., 'The ant went to the feast') are more common in English than in other Germanic languages.

Another topic to be considered is the lexical encoding of placement in each language, as the stimuli contains several placement events (e.g., 'Sara situated the expensive plate on its holder on the shelf.'). Chen et al. (2020) and Koster et al. (2018) noted that some Germanic languages, such as German and Dutch, often make the orientation of objects more explicit than English. Whereas in English readers could use the verb "put" in both "She put the book on the table" and "She put the bottle on the table," in both Dutch and German, readers could instead say "She laid the book on the table," and "She stood the bottle on the table." In these literal translations from German and Dutch, the verb "lay" encodes a horizontal orientation, whereas the verb "stand" encodes a vertical orientation. This distinction extends to verbs indicating existence. As Newman (2002) exemplified, an English speaker would be likely to say "There's a lamp in the corner," whereas a Dutch speaker would be more likely to say "There 'stands' a lamp in the corner." Nonetheless, we cannot conclude that these cross-linguistic differences are affecting the match advantage across languages because the current theories (e.g., language and situated simulation, Barsalou, 2008) do not precisely define the complexity of linguistic aspects such as placement events.

Methodological factors. Inconsistent findings on the match advantage of object orientation may be due to reliability in task design. For example, studies failing to detect the match advantage may not have required participants to verify the probe sentences they had read (see Zwaan, 2014). Without such a verification, participants might have paid less attention to the meaning of the probe sentences, in which they would have been less likely to form a mental representation of the objects (e.g., Zwaan & van Oostendorp, 1993). In this regard, it is relevant to acknowledge that variability originating from individual differences and other characteristics of experiments can substantially influence the results (Barsalou, 2019; Kaschak & Madden, 2021).

Cognitive Factors. Since Stanfield and Zwaan (2001) showed a match advantage of object orientation, later studies on this topic have examined the association between the match advantage and alternative cognitive mechanisms rather than the situated simulation. Spatial cognition is one of the potential cognitive mechanisms, which may be measured with mental rotation tasks. Studies have suggested that mental rotation tasks offer valid reflections of previous spatial experience (Frick & Möhring, 2013) and of current spatial cognition (Chu & Kita, 2008; Pouw et al., 2014). De Koning et al. (2017) suggested that effectiveness of mental rotation could increase with the depicted object size. Chen et al. (2020) examined this implication in use of the picture-picture verification task that was designed using the mental rotation paradigm (Cohen & Kubovy, 1993). In each trial of this task, two pictures appear on opposite sides of the screen. Participants had to verify whether the pictures represent identical or different objects. This study not only indicated shorter verification times for the same orientation (i.e., two identical pictures presented in horizontal or vertical orientation) but also showed the larger time difference for the large size object (i.e., pictures of bridges versus pictures of pens). The pattern of results were consistent among their investigated languages: English, Dutch, and Chinese. In comparison with the results of sentence-picture verification and picture-picture verification, Chen et al. (2020) depicted that mental rotation may affect the comprehension in some languages versus others by converting the picture-picture verification times to the mental rotation scores that were the discrepancy of verification times between the identical and different orientation[1]. With this measurement, we explore the relation of mental rotation in spatial cognition and orientation effect in comprehension across the investigated languages.

Purposes of this study

To scrutinize the discrepancies findings across languages and cognitive factors, we examined the reproducibility of the object orientation effect in a multi-lab collaboration. Our preregistered plan aimed at detecting a general match advantage of object orientation across languages and evaluated the magnitude of match advantage in each specific language. Additionally, we examined if match advantages were related to the mental rotation index. Thus, this study followed the original methods from Stanfield and Zwaan (2001) and addressed two primary questions: (1) How much of the match advantage of object orientation can be obtained within different languages and (2) How do differences in the mental rotation index affect the match advantage across languages?

Method

Hypotheses and Design

The study design for the sentence-picture and picture-picture verification task was mixed using between-participant (language) and within-participant (match versus

[1] In the preregistered plan, we used the term "imagery score" but this term was confusing. Therefore, we used "mental rotation scores" instead of "imagery scores" in the final report.

mismatch object orientation) independent variables. In the sentence-picture verification task, the match condition reflects a match between the sentence and the picture, whereas in the picture-picture verification, it reflects a match in orientation between two pictures. The only dependent variable for both tasks was response time. The time difference between conditions in each task are the measurement of mental simulation effects (for the sentence-picture task) and mental rotation scores (for the picture-picture task). We did not select languages systematically, but instead based on our collaboration recruitment with the Psychological Science Accelerator (PSA, Moshontz et al., 2018).

We pre-registered the following hypotheses:

(1) In the sentence-picture verification task, we expected response times to be shorter for matching compared to mismatching orientations within each language. In the picture-picture verification task, we expected shorter response time for identical orientation compared to different orientations. We did not have any specific hypotheses about the relative size of the object orientation match advantage in different languages.

(2) Based on the assumption that the mental rotation is a general cognitive function, we expect equal mental rotation scores across languages but no association with mental simulation effects (see Chen et al., 2020).

Participants

The preregistered power analysis indicated $n = 156$ to 620 participants for 80% power for a directional one-sample t-test for a $d = 0.20$ and 0.10, respectively. A mixed-model simulation suggested that $n = 400$ participants with 100 items (i.e., 24 planned items nested within at least five languages) would produce 90% power to detect the same effect as Zwaan and Pecher (2012). The laboratories were allowed to follow a secondary plan: a team collected at least their preregistered minimum sample size (suggested 100 to 160 participants, most implemented 50), and then determine whether or not to continue data collection via Bayesian sequential analysis (stopping data collection if $BF_{10} = 10$ or $1/10)$[2].

We finally collected data in 18 languages from 50 laboratories. Each laboratory chose a maximal sample size and an incremental n for sequential analysis before their data collection. Because the preregistered power analysis did not match the final analysis plan, we additionally completed a sensitivity analysis to ensure sample size was adequate to detect small effects, and the results indicated that each effect could be detected at a 2.23 millisecond range for the mental simulation effect of object orientation. Appendix 1 summarizes the details of sensitivity analysis.

The original sample sizes are presented in Table 1 for the teams that provided raw data to the project. Data collection proceeded in two broad stages: initially we collected data in the laboratory. However, when the global Covid-19 pandemic made this practice impossible to continue, we moved data collection online. For the in person data collection, demographic data was collected within a bundled, unrelated study that participants completed (Phills et al., in preparation). When the data collection was moved online, the demographic data collection was integrated into the current study $n_{demographic} = 4{,}605$. In both cases, the demographic data was separated from the experimental data. The in-person data required the experimenter to enter the lab ID information into the second study. Data entry errors in this stage resulted in some demographic information being excluded due to the inability to match to a particular lab $n = 39$. Additionally, some participants completed only the bundled study, and therefore, demographic sample sizes may be higher than the data collected for this study.

In total, 4,249 unique participants completed the study with 2,844 completing the in

[2] See details of power analysis in the preregistered plan, p. 13 ~ 15. https://psyarxiv.com/t2pjv/

person version and 1,405 completing the online version. The in person version included 35 research teams and the online version included 19 with 50 total teams across both data collection methods (i.e., some labs completed both in person and online data collection). Based on recommendations from research teams, two sets of data were excluded from all analyses due to participants being non-native speakers. Figure 2 provides a flow chart for participant exclusion and inclusion for analyses. All participating laboratories had either ethical approval or institutional evaluation before data collection. All data and analysis scripts are available on the source files (CODE OCEAN). Appendix 2 summarizes the average characteristics by language and laboratory.

(Insert Figure 2 about here)

General Procedure and Materials

In the beginning of the sentence-picture verification task, participants had to correctly answer all the practice trials. Each trial started with a left-justified and horizontally centered fixation point displayed for 1000 ms, immediately followed by the probe sentence. The sentence was presented until the participant pressed the space key, acknowledging that they understood the sentence. Then, the object picture (from Zwaan & Pecher, 2012) was presented in the center of the screen until the participant responded or it disappeared after 2 s. Participants were instructed to verify that the object on screen was mentioned in the probe sentence as quickly and accurately as they could. Following the original study (Stanfield & Zwaan, 2001), a memory check test was carried out after every three to eight trials to ensure that the participants had read each sentence carefully.

The picture-picture verification task used the same object pictures. In each trial, two objects appeared on either side of the central fixation point until either the participant indicated that the pictures displayed the same object or two different objects or until 2 s elapsed. In the trials where the same object was displayed, the pictures on each side were presented in the same orientation (both were horizontal/vertical) or different orientations (one was horizontal, one was vertical).

The study was executed using OpenSesame software for millisecond timing (Mathôt et al., 2012). After data collection moved online, in order to minimize the differences between on-site and web-based studies, we converted the original Python code to Javascript and collected the data using OpenSesame through a JATOS server (Lange et al., 2015; Mathôt & March, 2022). We proceeded with the online study from February to June 2021 after the changes in the procedure were approved by the journal editor and reviewers. Following the literature, we did not anticipate any theoretically important differences between the two data collection methods (see Anwyl-Irvine et al., 2020; Bridges et al., 2020; de Leeuw & Motz, 2016). The instructions and experimental scripts are available at the public OSF folder (https://osf.io/e428p/ “Materials” in Files).

Analysis Plan

Our first planned analysis[3] employed a random-effects meta-analysis model that estimated the match advantage across laboratories and languages. The meta-analysis summarized the median reaction times by match condition to determine the effect size by laboratory ($d = \frac{Mdn_{Mismatch} - Mdn_{Match}}{\sqrt{MAD^2_{Mismatch} + MAD^2_{Match} - 2 \times r \times MAD_{Mismatch} \times MAD_{Match}}} \times \sqrt{2 \times (1 - r)}$). For the languages for which at least two teams collected data, we computed the meta-analytical effect size of that language.

Next, we ran planned mixed-effect models using each individual response time from the sentence-picture verification task as the dependent variable. In each analysis we first built a simple linear regression model with a fixed intercept only. Then, we systematically added fixed effect and random intercepts arriving at the final model. First, the random intercepts were added to the model one-by-one in the following order: participant ID,

[3] See the analysis plan in the preregistered plan, p. 19 ~ 20. https://psyarxiv.com/t2pjv/ This plan was changed to a random-effects model to ensure that we did not assume the exact same effect size for each language and lab.

target, laboratory ID, and finally language. We compared the model fit measured by the AIC at each of these steps, to determine the best random effect structure for the model. Models with lower AIC were preferred over models with higher AIC, and in a case where the difference in AIC did not reach 2, the model with the fewer parameters was preferred. Then, the fixed effect of matching condition (match vs. mismatch) was added to the model. Language-specific mixed-effect models were conducted in the same way if the meta-analysis showed a significant orientation effect.

According to the preregistration, we planned to first evaluate the equality of mental rotation scores across languages using an ANOVA. However, this plan was updated to use mixed models instead due the nested structure of the data. The same analysis plan was used for model building and selection as described above for the sentence-picture verification task. The last planned analysis was to use mental rotation scores for the prediction of mental stimulation with an interaction between language and mental rotation scores to determine there were differences in prediction of match advantage based on language. This model was updated to a mixed-effects model to control for the random effect of the data collection lab.

Decision criterion. *p*-values were interpreted using the preregistered alpha level of .05. *p*-values for each effect were calculated using the Satterthwaite approximation for degrees of freedom for individual predictor coefficients and meta-analysis (Luke, 2017).

Intra-lab analysis during data collection. Before data collection, each lab decided whether they wanted to apply a sequential analysis (Schönbrodt et al., 2017) or whether they wanted to settle for a fixed sample size. The preregistered protocol for labs applying sequential analysis established that they could stop data collection upon reaching the preregistered criterion ($BF_{10} = 10 \ or \ -10$), or the maximal sample size. Each laboratory chose a fixed sample size and an incremental n for sequential analysis before their data collection. Two laboratories (HUN_001, TWN_001) stopped data collection at the preregistered criterion, $BF_{10} = -10$. Fourteen laboratories did not finish the sequential analysis because (1) twelve laboratories were interrupted by the pandemic outbreak; (2) two laboratories (TUR_007E, TWN_002E) recruited English-speaking participants to comply with institutional policies. Lab-based records were reported on a public website as each laboratory completed data collection (details are available in Appendix 3).

Results

Data Screening

As shown in Figure 2, entire participants were first removed from the sentence-picture and picture-picture tasks if they did not perform at 70% accuracy. Next, the data were screened for outliers. Our preregistered plan excluded outliers based on a linear mixed-model analysis for participants in the third quantile of the grand intercept (i.e., participants with the longest average response times). After examining the data from both online and in-person data collection, it became clear that both a minimum response latency and maximum response latency should be employed, as improbable times existed at both ends of the distribution (Kvålseth, 2021; Proctor & Schneider, 2018). The minimum response time was set to 160 ms. The maximum response latency was calculated as two times the mean absolute deviation plus the median calculated separately for each participant. Exclusions were performed at the trial level for these outlier response times.

In order to ensure equivalence between data collection methods, we evaluated the response times predicted by the fixed effects of the interaction between match (match versus mismatch) and data collection source (in person, online). We included random intercepts of participant, lab, language, and random slopes of source by lab, and source by language. This analysis showed no difference between data sources: $b = 2.41$, $SE = 2.77$, $t(73729.28) = 0.87$, $p = .385$. Therefore, the following analyses did not separate in person and online data. Table 2 provides a summary of the match advantage by language for the sentence-picture verification task.

Meta-Analysis

```
## pdf
##    2
```

The planned meta-analysis examined the effect overall and within languages wherein at least two laboratories had collected data (Arabic, English, German, Norway, Simplified Chinese, Traditional Chinese, Slovak, and Turkey). Figure 3 showed a significant positive orientation effect across German laboratories (b = 16.68, 95% CI [7.75, 25.62]) but did not reveal a significant overall effect (b = 2.05, 95% CI [-2.71, 6.82]). Also, a significant negative orientation effect was found in the Hungarian (b = -20.00, 95% CI [-29.60, -10.40]) and the Serbian laboratory (b = -17.25, 95% CI [-32.26, -2.24]), although in these languages only a single laboratory participated, so no language-specific meta-analysis was conducted.

(Insert Figure 3 about here)

Mixed-Linear Modeling

First, an intercept only model of response times with no random intercepts was computed for comparison purposes 1008828.79. The model with the participant random intercept was an improvement over this model 971783.32. The addition of a target random intercept improved model fit over the participant intercept only model 969506.32. Data collection lab was then added to the model as a random intercept, also showing model improvement 969265.28, and the random intercept of language was added last 969263.66 which did not show model improvement at least 2 points change. Last, the fixed effect of match advantage was added with approximately the same fit as the three random-intercept model, 969265.06. This model did not reveal a significant effect of match advantage: b = -0.17, SE = 1.20, t(69830.14) = -0.14, p = .887.

We conducted an exploratory mixed-effect model on German data as this was the only language indicating a significant match advantage in the meta-analysis. An intercept-only model with random effects for participants, target, and lab was used as a comparison, as the last random effect (language) could not be used in this model, 55828.57. The addition of the fixed effect of match showed a small improvement over this random-intercept model, 55824.52. This model did not reveal a significant effect of match advantage: $b = 4.84$, $SE = 4.12$, $t(4085.71) = 1.17$, $p = .241$. All the details of the above fixed effects and random intercepts are summarized in Appendix 4.

Mental Rotation Scores

Using the same steps as described for the sentence-picture verification mixed model, we first started with an intercept only model with no random effects for comparison 1029639.26. The addition of subject 980138.90, item 977307.03, lab 976991.96, and language 976987.98 random intercepts all subsequently improved model fit. Next, the match effect for object orientation was entered as the fixed effect for mental rotation score, 973324.45, which showed improvement over the random intercepts model. This model showed a significant effect of object orientation, $b = 32.30$, $SE = 0.53$, $t(79605.20) = 61.24$, $p = < .001$, such that identical orientations were processed faster than rotated orientations. The coefficients of all considered mixed-effects models are reported in Appendix 5, along with all effects presented by language.

Prediction of Match Advantage

The last analysis included a mixed effects regression model using the interaction of language and mental rotation to predict match advantage. First, an intercept only model was calculated for comparison, 42678.66, which was improved slightly by adding a random intercept of data collection lab, 42677.80. The addition of the fixed effects interaction of language and imagery improved the overall model, 42633.44. English was used as the comparison group for all language comparisons. No interaction effects or the main effect of mental rotation were significant, and these results are detailed in Appendix 5.

Discussion

Results from the meta-analysis and mixed-effects models on match advantage show similar, but slightly convergent results. The meta-analysis showed a small, but greater than zero, effect size for German, while the mixed-effects German model did not support these findings. Both analyses agree that the match advantage effect for object orientation was not supported. In contrast, mixed-effect models indicated significant mental rotation differences with an advantage for identical rotations. However, this rotation advantage does not predict the match advantage nor interact with language to predict object orientation effects. We summarize the lessons learned on the methodology, analysis, and theoretical issues and attempt to address in which aspect the hypotheses obtained the disconfirmative evidence from the current findings.

Methodology

This study reflected the difficulty of investigating cognition across languages, especially when dealing with effects that require large sample sizes (see Loken & Gelman, 2017; Vadillo et al., 2016). Our data collection deviated from the preregistered plan because due to the COVID-19 pandemic. Due to the lack of participant monitoring online, and an inspection of the data, we *post hoc* used filtering on outliers in terms of participants' response times for both too quick and too slow responses. After these exclusions, the mixed-effect model confirmed no difference of response times between in person and online data. Although we combined the two data sets in the final data analysis, it is worth considering that online participants' attention may be easily distracted given the lack of any environmental control and lack of experimenter assistance.

When using sentence-picture verification task as a comprehension task, researchers have had to insert the comprehension questions or memory checks among the experimental trials (Chen et al., 2020; Stanfield & Zwaan, 2001). Kaschak and Madden (2021) pointed out this setting could trigger the participants to consciously generate mental imagery while reading the probe sentence. If the current results showed significant match advantages, we may have had to evaluate the contribution of participants' strategy. However, we do not find that mental imagery predicted match advantage, which implies that this strategy was not effective or unsupported.

Analysis Issues

The sensitivity analysis indicated that a small effect was potentially detectable, and the limited number of trials could be an influencing factor to why the effect was not detectable. Most studies use approximately 24 items (12 match and 12 mismatch), however, these items vary in length and difficulty, which may not be completely controlled using random effects for item. In a classical cognitive capacity measurement, such as Stroop task and Flanker task, the suggested trial numbers are beyond 100 to decrease the trial-level noise (Rouder et al., 2019).

Theoretical Issues

Mental simulation theories of comprehension have suggested that cognitive processing converts discourse into either abstract symbols or grounded mental representations (Barsalou, 1999, 2009; Zwaan, 2014). This study did not support differences in match advantage (minus German effects in the meta-analysis), and therefore, may not support an embodied view of the priming-based mechanism for the reading task as like sentence-picture verification (Kaschak & Madden, 2021).

The original probe sentences (see Stanfield & Zwaan, 2001; Zwaan & Pecher, 2012) were the researchers' creations which were compatible with the experimental demands but may not capture the theoretical complexity proposed by embodied views. These sentences describe the interaction between one actor and one object. A different study (Chen et al., 2020) that found the orientation effect used lab created sentences as well. In comparison with the simple sentences (e.g., Chen et al. used I saw "something"), the second set of sentences addressed how English participants from the original study may have comprehended the sentences and which language-specific aspects may alter the sentence content in non-English studies. We suggest that further explorations could employ the original object pictures after simple and complex sentences. The results will help establish specific guidelines for exploring sentence content.

A secondary task used sentence-picture verification was designed to encourage participants to understand the probe sentences. However, the verification task could potentially have been answered without realizing sentence content. A secondary task could be designed to explore the probe meaning that would require participants to deeply process sentences. Even with the concern of the secondary task inspiring the use of strategies instead of comprehension (e.g., Rommers et al., 2013), a new set of items could explore the effect of secondary task demands (memory checks; comprehension questions). These studies are necessary to distinguish the effects from the targeted cognitive processing and strategy in many language topics, such as semantic priming (McNamara, 2005).

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Barsalou, L. W. (2019). Establishing generalizable mechanisms. *Psychological Inquiry*, *30*(4), 220–230. https://doi.org/10.1080/1047840X.2019.1693857

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660. https://doi.org/10.1017/S0140525X99002149

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*, 1281–1289. https://doi.org/10.1098/rstb.2008.0319

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Chen, S.-C., de Koning, B. B., & Zwaan, R. A. (2020). Does object size matter with regard to the mental simulation of object orientation? *Experimental Psychology*, *67*(1), 56–72. https://doi.org/10.1027/1618-3169/a000468

Chu, M., & Kita, S. (2008). Spontaneous gestures during mental rotation tasks: Insights into the microdevelopment of the motor strategy. *Journal of Experimental Psychology: General*, *137*(4), 706–723. https://doi.org/10.1037/a0013157

Cohen, D., & Kubovy, M. (1993). Mental rotation, mental representation, and flat slopes. *Cognitive Psychology*, *25*, 351–382. https://doi.org/10.1006/cogp.1993.1009

Connell, L. (2007). Representing object colour in language comprehension.

Cognition, *102*, 476–485. https://doi.org/10.1016/j.cognition.2006.02.009

De Koning, B. B., Wassenburg, S. I., Bos, L. T., & Van der Schoot, M. (2017). Mental simulation of four visual object properties: Similarities and differences as assessed by the sentence-picture verification task. *Journal of Cognitive Psychology*, *29*(4), 420–432. https://doi.org/10.1080/20445911.2017.1281283

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*, *48*(1), 1–12. https://doi.org/10.3758/s13428-015-0567-2

Engelen, J. A. A., Bouwmeester, S., de Bruin, A. B. H., & Zwaan, R. A. (2011). Perceptual simulation in developing language comprehension. *Journal of Experimental Child Psychology*, *110*(4), 659–675. https://doi.org/10.1016/j.jecp.2011.06.009

Frick, A., & Möhring, W. (2013). Mental object rotation and motor development in 8- and 10-month-old infants. *Journal of Experimental Child Psychology*, *115*(4), 708–720. https://doi.org/10.1016/j.jecp.2013.04.001

Kaschak, M. P., & Madden, J. (2021). Embodiment in the Lab: Theory, Measurement, and Reproducibility. In M. D. Robinson & L. E. Thomas (Eds.), *Handbook of Embodied Psychology* (pp. 619–635). Springer International Publishing. https://doi.org/10.1007/978-3-030-78471-3_27

Koster, D., Cadierno, T., & Chiarandini, M. (2018). Mental simulation of object orientation and size: A conceptual replication with second language learners. *Journal of the European Second Language Association*, *2*(1). https://doi.org/10.22599/jesla.39

Kvålseth, T. O. (2021). Hick's law equivalent for reaction time to individual stimuli. *British Journal of Mathematical and Statistical Psychology*, *74*(S1), 275–293. https://doi.org/10.1111/bmsp.12232

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLOS ONE*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Li, Y., & Shang, L. (2017). An ERP study on the mental simulation of implied object color information during Chinese sentence comprehension. *Journal of Psychological Science*, *40*(1), 29–36. https://doi.org/10.16719/j.cnki.1671-6981.20170105

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Mathôt, S., & March, J. (2022). Conducting linguistic experiments online with OpenSesame and OSWeb. *Language Learning*, *72*(4), 1017–1048. https://doi.org/10.1111/lang.12509

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

McNamara, T. P. (2005). *Semantic Priming: Perspectives From Memory and Word Recognition*. Psychology Press.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., … Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515.

https://doi.org/10.1177/2515245918797607

Newman, J. (2002). 1. A cross-linguistic overview of the posture verbs “Sit,” “Stand,” and “Lie.” In J. Newman (Ed.), *Typological Studies in Language* (Vol. 51, pp. 1–24). John Benjamins Publishing Company. https://doi.org/10.1075/tsl.51.02new

Ostarek, M., & Huettig, F. (2019). Six Challenges for Embodiment Research. *Current Directions in Psychological Science*, *28*(6), 593–599. https://doi.org/10.1177/0963721419866441

Pecher, D., van Dantzig, S., Zwaan, R. A., & Zeelenberg, R. (2009). Language comprehenders retain implied shape and orientation of objects. *The Quarterly Journal of Experimental Psychology*, *62*(6), 1108–1114. https://doi.org/10.1080/17470210802633255

Pouw, W. T. J. L., de Nooijer, J. A., van Gog, T., Zwaan, R. A., & Paas, F. (2014). Toward a more embedded/extended perspective on the cognitive function of gestures. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00359

Proctor, R. W., & Schneider, D. W. (2018). Hick’s law for choice reaction time: A review. *Quarterly Journal of Experimental Psychology*, *71*(6), 1281–1299. https://doi.org/10.1080/17470218.2017.1322622

Rommers, J., Meyer, A. S., & Huettig, F. (2013). Object shape and orientation do not routinely influence performance during language processing. *Psychological Science*, *24*(11), 2218–2225. https://doi.org/10.1177/0956797613490746

Rouder, J., Kumar, A., & Haaf, J. M. (2019). *Why Most Studies of Individual Differences With Inhibition Tasks Are Bound To Fail* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/3cjr5

Sato, M., Schafer, A. J., & Bergen, B. K. (2013). One word at a time: Mental representations of object shape change incrementally during sentence processing. *Language and Cognition*, *5*(04), 345–373.

https://doi.org/10.1515/langcog-2013-0022

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. https://doi.org/10.1037/met0000061

Scorolli, C. (2014). Embodiment and language. In L. Shapiro (Ed.), *The Routledge handbook of embodied cognition* (pp. 145–156). Routledge.

Šetić, M., & Domijan, D. (2017). Numerical Congruency Effect in the Sentence-Picture Verification Task. *Experimental Psychology*, *64*(3), 159–169. https://doi.org/10.1027/1618-3169/a000358

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, *12*(2), 153–156. https://doi.org/10.1111/1467-9280.00326

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review*, *23*(1), 87–102. https://doi.org/10.3758/s13423-015-0892-6

Verkerk, A. (2014). *The evolutionary dynamics of motion event encoding.* [PhD thesis]. Radboud Universiteit Nijmegen.

Zwaan, R. A. (2014). Embodiment and language comprehension: Reframing the discussion. *Trends in Cognitive Sciences*, *18*(5), 229–234. https://doi.org/10.1016/j.tics.2014.02.008

Zwaan, R. A., Diane Pecher, Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (2017). Participant Nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin & Review*, 1–5. https://doi.org/10.3758/s13423-017-1348-y

Zwaan, R. A., & Madden, C. J. (2005). Embodied sentence comprehension. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception and* *action in memory, language, and thinking* (pp. 224–245). Cambridge University Press.

Zwaan, R. A., & Pecher, D. (2012). Revisiting mental simulation in language comprehension: Six replication attempts. *PLoS ONE*, *7*, e51382. https://doi.org/10.1371/journal.pone.0051382

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science*, *13*, 168–171. https://doi.org/10.1111/1467-9280.00430

Zwaan, R. A., & van Oostendorp, H. (1993). Do readers construct spatial representations in naturalistic story comprehension? *Discourse Processes*, *16*(1-2), 125–143. https://doi.org/10.1080/01638539309544832

Table 1

Demographic and Sample Size Characteristics

Language	SP Trials	PP Trials	SP N	PP N	Demo N	Female N	Male N	M Age	SD Age
Arabic	2544	2544	106	106	107	42	12	32.26	18.59
Brazilian Portuguese	1200	1200	50	50	50	36	13	30.80	8.73
English	45189	45336	1884	1889	2055	1360	465	21.71	3.85
German	5616	5616	234	234	248	98	26	22.34	3.40
Greek	2376	2376	99	99	109	0	0	33.86	11.30
Hebrew	3576	3571	149	149	181	0	0	24.25	9.29
Hindi	1896	1896	79	79	86	57	27	21.66	3.46
Magyar	3610	3816	151	159	168	3	1	21.50	2.82
Norwegian	3576	3576	149	149	154	13	9	25.22	6.40
Polish	1368	1368	57	57	146	0	0	23.25	7.96
Portuguese	1488	1464	62	61	55	26	23	30.74	9.09
Serbian	3120	3120	130	130	130	108	21	21.38	4.50
Simple Chinese	2040	2016	85	84	96	0	1	21.92	4.68
Slovak	3881	3599	162	150	325	1	0	21.77	2.33
Spanish	3120	3096	130	129	146	0	0	21.73	3.83
Thai	1200	1152	50	48	50	29	9	21.54	3.81
Traditional Chinese	3600	3600	150	150	186	69	46	20.89	2.44
Turkish	6456	6432	269	268	274	36	14	21.38	4.59

Note. SP = Sentence Picture Verification, PP = Picture Picture Verification. Sample sizes for demographics may be higher than the sample size for the this study, as participants could have only completed the bundled experiment. Additionally, not all entries could be unambigously matched by lab ID, and therefore, demographic sample sizes could also be less than data collected.

Table 2

Descriptive Statistics by Language

Language	Accuracy Percent	Mismatching	Matching	Match Advantage
Arabic	90.65	580.25 (167.53)	581.00 (200.89)	-0.75
Brazilian Portuguese	94.87	641.00 (136.40)	654.50 (146.78)	-13.50
English	95.04	576.75 (124.17)	578.75 (127.87)	-2.00
German	96.53	593.00 (106.75)	576.00 (107.12)	17.00
Greek	92.35	753.50 (225.36)	728.50 (230.91)	25.00
Hebrew	96.73	569.50 (98.59)	574.50 (110.45)	-5.00
Hindi	91.32	638.50 (207.19)	662.00 (228.32)	-23.50
Hungarian	96.47	623.00 (111.94)	643.00 (129.73)	-20.00
Norwegian	96.93	592.50 (126.39)	612.00 (136.03)	-19.50
Polish	96.11	601.00 (139.36)	586.00 (108.23)	15.00
Portuguese	95.01	616.50 (144.55)	607.00 (145.29)	9.50
Serbian	94.78	617.75 (158.64)	635.00 (168.28)	-17.25
Simplified Chinese	92.39	655.00 (170.50)	642.50 (158.64)	12.50
Slovak	96.45	610.50 (125.28)	607.25 (117.87)	3.25
Spanish	94.32	663.00 (147.52)	676.00 (154.19)	-13.00
Thai	93.92	652.50 (177.91)	637.75 (130.10)	14.75
Traditional Chinese	94.41	625.00 (139.36)	620.00 (123.06)	5.00
Turkish	95.38	654.50 (146.04)	637.00 (126.02)	17.50

Note. Average accuracy percentage, Median response times and median absolute deviations (in parentheses) per match condition (Mismatching, Matching); Match advantage (difference in response times).

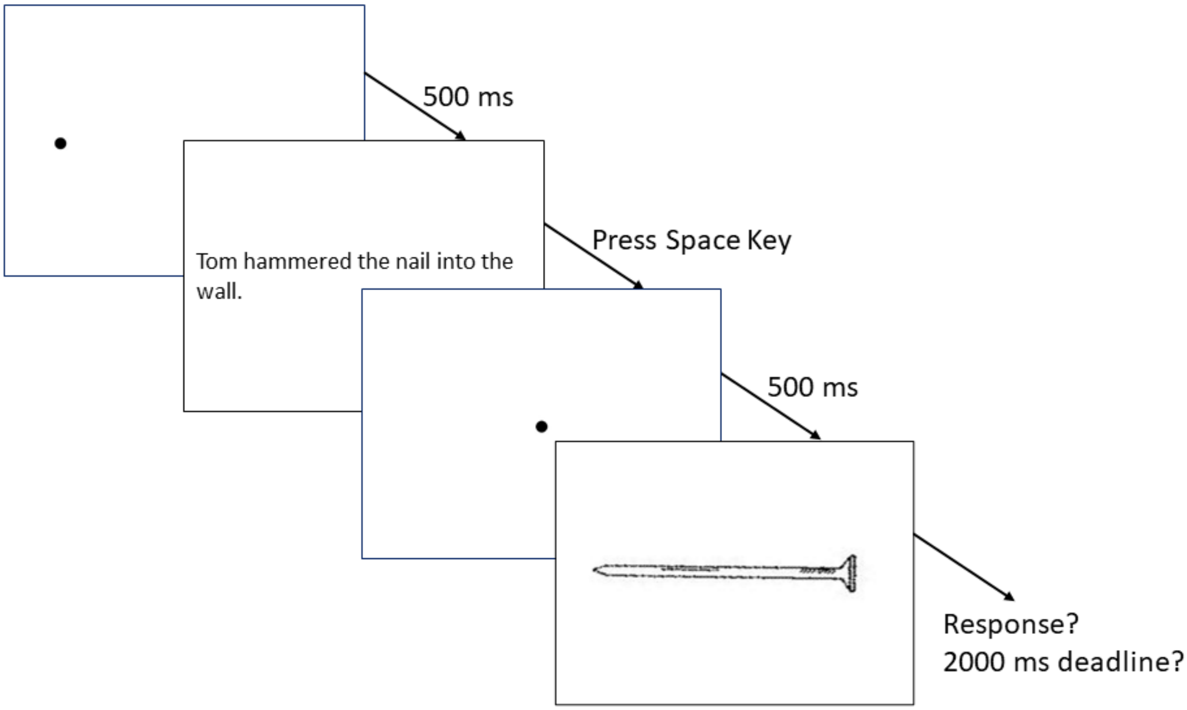

Figure 1

Procedure of sentence-picture verification task.

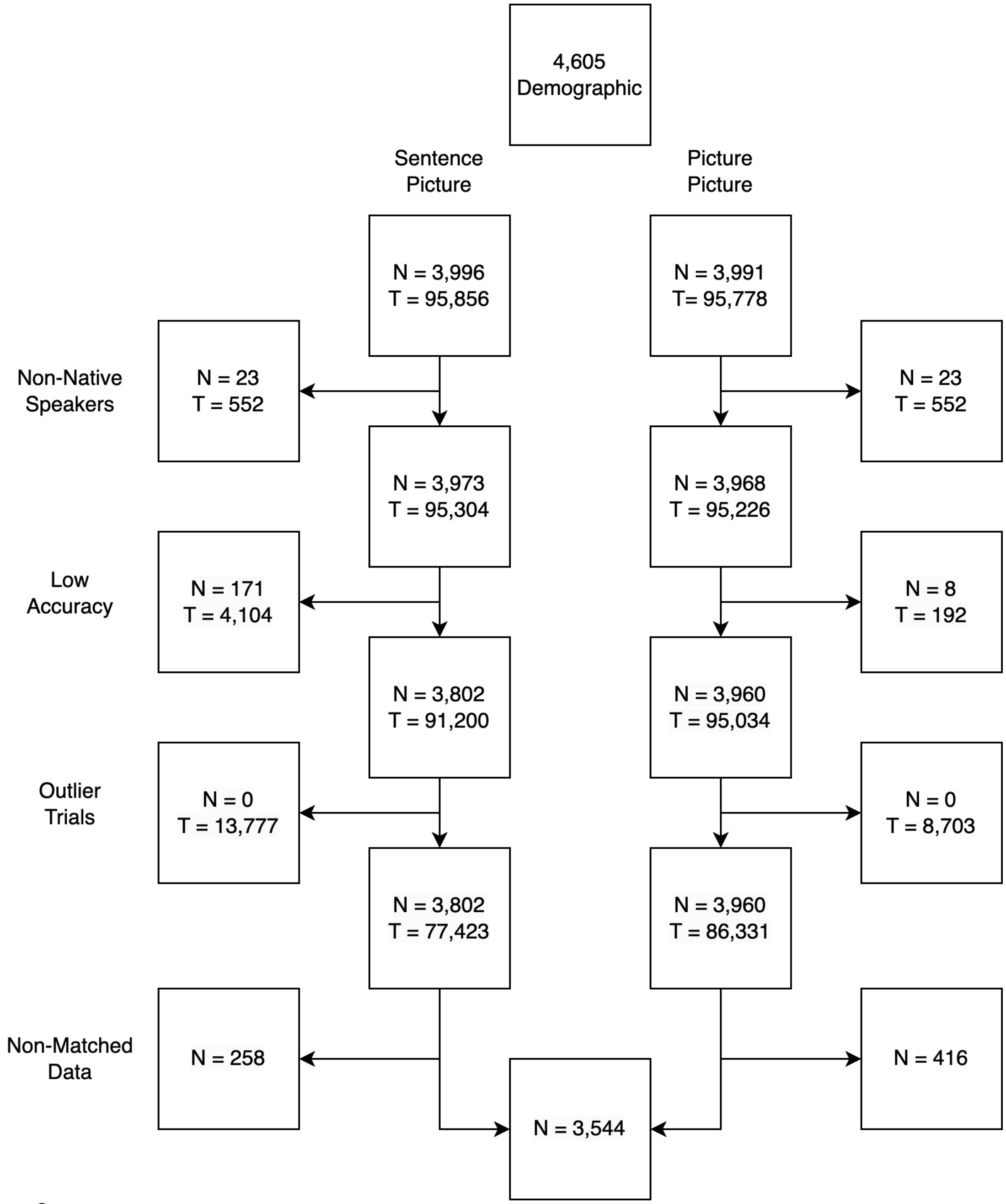

Figure 2

Sample size and exclusions. N = number of unique participants, T = number of trials. The final combined sample was summarized to a median score for each match/mismatch condition, and therefore, includes one summary score per person.

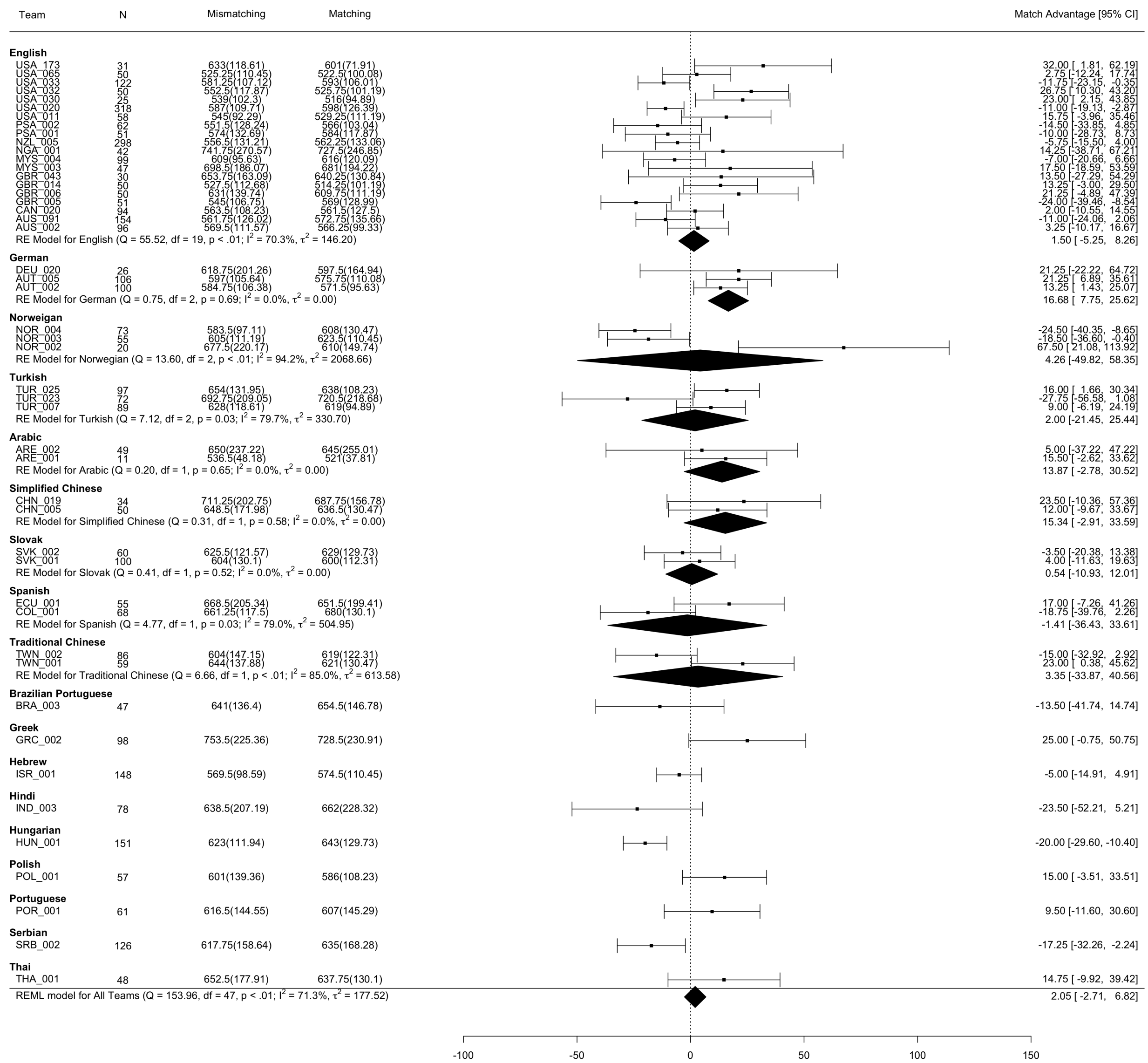

Figure 3

Meta-analysis on match advantage of object orientation for all languages