Bilingualism: Language and Cognition

cambridge.org/bil

Research Article

Cite this article: Wei, Y., Yang, W., Oppenheim, G. and Thierry, G. (2025). The abstract concept of perceived power is embodied to a lesser extent in the second language. *Bilingualism: Language and Cognition* 1–11. https://doi.org/10.1017/S1366728925100588

Received: 19 September 2024
Revised: 14 March 2025
Accepted: 26 August 2025

Keywords:
conceptual representation; embodiment; metaphor; bilingualism; perceived power

Corresponding author:
Yufen Wei;
Email: yfw21bgp@bangor.ac.uk

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

The abstract concept of perceived power is embodied to a lesser extent in the second language

Yufen Wei[1], Wenwen Yang[1], Gary Oppenheim[1] and Guillaume Thierry[1,2]

[1]School of Psychology and Sport Science, Bangor University, Bangor, UK and [2]Faculty of English, Adam Mickiewicz University, Poznan, Poland

Abstract

Embodied cognition theory posits that language comprehension is grounded in sensorimotor experience. For instance, abstract concepts such as perceived power are metaphorically associated with spatial information such as physical size. Here, using a size judgement task, we investigated whether perceived power embodiment differs between languages in Chinese–English bilinguals. Asked to make judgements regarding the physical size of words, participants responded faster and made fewer errors to high-power words (e.g., king) presented in bold and large font than in thin and small font, while no such effect was found for low-power words. Furthermore, this congruency effect was stronger in bilinguals' L1 (Chinese) than in their L2 (English). Thus, while embodiment of perceived power is detectable in both languages of bilinguals, it appears weaker in the L2. This study highlights cross-linguistic similarities and differences in the embodiment of abstract concepts and contributes to our understanding of conceptual knowledge grounding in bilinguals.

Highlights

- Abstract concepts such as perceived power are grounded in sensorimotor experiences and metaphorically linked to physical size.
- The embodiment of perceived power is significantly stronger in a bilingual's first language than in the second language.
- Differences in L2 embodiment are likely influenced by factors such as proficiency and cultural variations.

1. Introduction

Theories of embodied cognition posit that conceptual knowledge is grounded in our sensory–motor experience (Barsalou, 1999, 2008; Gallese & Lakoff, 2005; Glenberg & Kaschak, 2002; Pulvermüller, 2005). For example, when a person thinks about an object such as a paper-printed book, memories of earlier physical experiences such as holding, flipping through and reading a book are reactivated. In other words, the brain partially acts as if one is perceiving and interacting with a book when encountering the concept of a book.

Empirical support for embodiment theory largely comes from the stimulus–response compatibility (SRC) effects, where responses are faster and more accurate when the nature of response (perceptual dimension) matches some stimulus features (conceptual dimension). This effect has been observed in studies on concrete concepts (e.g., Pecher et al., 2003; Zwaan & Yaxley, 2003) and action representations (e.g., Glenberg & Kaschak, 2002; Hauk et al., 2004). For instance, Šetić and Domijan (2007, Experiment 2) examined the influence of spatial associations in lexical processing using a binary categorization task. Participants categorized words referring to entities typically associated with either an upper (e.g., butterfly) or lower (e.g., carpet) spatial position, presented above or below the centre of the screen. Reaction times were faster when words appeared in spatially congruent as compared to incongruent positions, supporting the idea that conceptual representations integrate sensory and spatial experiences.

However, proponents of embodiment face a significant challenge when it comes to the representation of concepts that are more abstract, for which we lack direct bodily experiences. Conceptual Metaphor Theory (CMT) postulates that abstract concepts (the 'target domain') are grounded via metaphorical associations to concrete concepts (the 'source domain', e.g., Gallese & Lakoff, 2005; Lakoff & Johnson, 1999). Spatial metaphor has been widely explored as a means of understanding abstract concepts. Research shows that abstract concepts such as time, emotion, morality and perceived power are spatially embodied along the vertical axis (e.g., Meier & Robinson, 2004; Schubert, 2005). For instance, Meier and Robinson (2004) examined the relationship between emotion and vertical spatial orientation by asking participants to indicate

https://doi.org/10.1017/S1366728925100588 Published online by Cambridge University Press

whether words such as ‘hero’ and ‘liar’ had a positive or negative meaning to test the Metaphor Congruency Effect. Positive-valence words were evaluated faster when presented above (congruent) as compared to below (incongruent) a fixation point, while negative words showed the reverse pattern, suggesting a link between valence and vertical position.

An alternative account challenges the assumption that metaphorical mapping drives spatial congruency effects. The polarity correspondence principle posits that faster responses in binary categorization tasks benefit from structural alignment between stimulus and response features, rather than from perceptual mapping (Proctor & Cho, 2006; see also Treccani et al., 2019). Under this framework, stimulus attributes and spatial attributes of responses are coded as positive (+polar, e.g., hero – UP) or negative (−polar, e.g., liar – DOWN), with faster responses occurring when the polarities align (e.g., Lakens, 2012).

The debate between CMT and the polarity correspondence principle raises an important question: how are abstract concepts that do not have an inherent binary distinction, such as *perceived power*, represented? Unlike morality or emotion, perceived power is inherently relative: one does not denote how powerful a person is per se, but rather whether one individual has more or less social power than another.

Perceived power is often metaphorically mapped onto vertical space, as reflected in expressions such as ‘high status’ or ‘low rank’ (e.g., Schubert, 2005; Wei et al., 2024; Wu et al., 2016). Using a Stroop-like semantic judgement paradigm, Schubert (2005) showed that participants respond faster and more accurately when high-power words are presented in relatively higher screen positions, suggesting that processing perceived power automatically activates vertical space information. Challenging this assumption, Lakens et al. (2011) showed that the vertical representation of power is not absolute but relational. When participants placed high- and low-power words along a vertical scale, they positioned high-power words higher only when low-power words were present, whereas no difference was observed for low-power words irrespective of the presence of high-power words. This effect weakened when evaluating high-power words in isolation in a between-subject design, indicating that power representation depends on contextual contrast rather than fixed spatial mappings.

Beyond vertical space, power has also been associated with physical size (e.g., He et al., 2015; Schubert et al., 2009; Yap et al., 2013). Larger entities often convey dominance, an association reflected linguistically. For instance, in western culture, expressions such as *big boss* or small *fry* relate power to size and in Mandarin Chinese, the same associations exist as in *daguan* – senior official (*da* means big) versus *xiaozu* – low-ranking soldier (*xiao* means small). Schubert et al. (2009) tested this link using an interference paradigm and found slower, less accurate responses when font size mismatched the implied power of words (e.g., professor in small font). He et al. (2015) demonstrated bidirectional effects: participants primed with power-related words perceived subsequent stimuli as larger, while larger stimuli biased participants to perceive words as more powerful.

What is unknown however is whether embodied cognition is at work when individuals learn a new language. Arguably, embodiment applies to language learning from birth since the native language is acquired alongside early bodily experiences. However, this may not be the case for a language learnt beyond childhood. The growing evidence supporting embodied cognition theory has predominantly come from research on first language processing, with very little theoretical or empirical consideration of second language acquisition and bilingualism (see Kühne & Gianelli, 2019 for a review). We can consider the following three scenarios for embodied cognition in bilinguals:

1. *Two monolinguals in one brain.* As a straightforward extension of the monolingual literature, one might naively assume fully language-selective semantic access and therefore expect bilinguals’ embodiment effects in each of their languages to simply mirror those of monolingual speakers. For instance, if a bilingual’s L1 represents power only along the vertical axis, and their L2 represents power only in terms of physical size, then one would expect participants to show vertical congruency effects when using their L1 and size congruency effects when using their L2.
2. *Full integration.* The Bilingual Interactive Activation Plus model (BIA+, Dijkstra & van Heuven, 2002) assumes that lexical representations of the two languages of bilinguals are integrated. Such models assume non-selective access to orthographic, phonetic and semantic representations across L1 and L2 (e.g., Thierry & Yan, 2007; van Heuven & Dijkstra, 2010; Wu et al., 2013), which would seem to predict embodiment effects of similar magnitudes in the two languages, though the particular metaphors might differ across languages, a bilingual speaker would always show the influence of both. There is considerable empirical evidence to suggest that L2 and L1 processing involve similar access to sensorimotor information, at least for action related and emotion words (Ahlberg et al., 2018; Bergen et al., 2010; Buccino et al., 2017; De Grauwe et al., 2014; Dudschig et al., 2014; see Monaco et al., 2019 for a review). For instance, Dudschig et al. (2014) used an adapted Stroop paradigm to investigate the activation of sensorimotor information in L2 processing. Stimuli were implicit location words (e.g., star, ground) and emotion words (e.g., happy, sad), which participants categorized by performing upward/downward finger movements based on word colour. Participants responded significantly faster when word meaning was congruent with the response location in both languages (e.g., upward response with the word star or happy). The authors argued that sensorimotor experiences are automatically activated in L2 (English) and are not significantly different from that of L1 (German).
3. *Reduced access to L2 semantics.* Other accounts, such as the Revised Hierarchical model (RHM), assume separate lexical levels of representation and a conceptual level shared by both languages. L2 can have an indirect connection to the conceptual level via L1 mediation. As L2 proficiency increases, the link between L2 and conceptual representation strengthens (Kroll & Stewart, 1994). In the context of embodied cognition, this would mean that sensorimotor activation in a bilingual’s two languages should be different, with weaker connections to sensorimotor representations in L2 (e.g., Bai & He, 2022; Foroni, 2015; Qian, 2016; Vukovic & Shtyrov, 2014). Qian (2016) asked participants to judge the perceived power of stimuli displayed either in the upper or lower part of a screen. Participants responded faster when high-power words were presented in the upper part as compared to the lower part of the screen, the effect was stronger in L1 than L2, and it was stronger in L2 learners with a higher proficiency.

Wei et al. (2024) previously assessed whether the vertical mapping of perceived power applies equally in the two languages of bilinguals using event-related potentials (ERPs) and auditory stimuli presented above or below the participant’s sitting position. Chinese–English

https://doi.org/10.1017/S1366728925100588 Published online by Cambridge University Press

bilinguals responded faster in congruent (high-power words played from above) than incongruent conditions (high-power words played from below), with no such congruency effect for low-power words. Curiously, the congruency effect was not found in L2 English either behaviourally or in the ERP data. In addition, there was no embodiment effect in L1 speakers of English. Here, we chose to move to the realm of physical size as a more salient mapping for perceived power, in an attempt to detect differences between languages within bilinguals and across groups in English at the behavioural level.

In this study, we asked Chinese–English bilinguals to engage in a size judgement task while ignoring animal names (to ensure semantic processing of every word presented). The task was intended to measure whether the metaphorical association between perceived power and physical size holds true to a similar extent across the two languages. We also tested a control group of L1 English speakers to assess potential embodiment effects for the same concepts in English. Stimuli could refer to either human entities varying in power status (e.g., king/servant – 'power words') or animal names (e.g., dog), serving as fillers. High-power words were considered as congruent when presented in a larger and bolder font and as incongruent when presented in a smaller and thinner font.

In line with accounts of embodied cognition and word processing in bilinguals (Barsalou, 1999, 2008; Gallese & Lakoff, 2005; Kroll & Stewart, 1994), we hypothesized (1) a congruency effect in the processing of perceived power and (2) a weaker congruency effect in L2 (English) than in L1 (Chinese) in late, sequential bilinguals with medium proficiency in L2. If our hypotheses are correct, we would predict faster response times (RTs) and greater accuracy when participants see a word in the congruent as compared to incongruent condition. Critically, if bilinguals have weaker or less direct L2 semantics connections, then this effect should be weaker in their L2 than in their L1.

2. Material and methods

2.1. Participants

A *priori* power analysis was conducted using the *Superpower* package (Lakens & Caldwell, 2021) in R to estimate the required sample size for a 2 (Language: Chinese vs. English) × 2 (Congruency: congruent vs. incongruent) within-subject design. The analysis was based on data from a pilot study involving 13 participants (nine females, M = 22.9, SD = 0.7). Only correct responses were included when calculating the mean response times (RTs) and standard deviations (SD). Mean RTs, SD and the correlation (r) for each combination of factor levels were provided in the power analysis. We simulated 10,000 sets of observations adhering to the distributional properties of the data. The result was that a sample size of 12 participants would be required to achieve a threshold of 90% power ($\alpha = 0.05$) for detecting a large-sized ($\eta_p^2 = 0.59$) main effect of Language, 77 participants would be needed to detect a medium-sized ($\eta_p^2 = 0.13$) main effect of Congruency and 88 participants would be required to detect a medium-sized ($\eta_p^2 = 0.12$) interaction between Language and Congruency.[1]

We recruited 104 unbalanced Chinese–English bilinguals from China and 111 native English speakers from the UK for this online study. All participants reported normal or corrected-to-normal vision and had no learning or language disabilities. Twelve bilingual participants and 20 native English participants were excluded from the analyses due to ineligibility, or withdrawal before experiment completion, or their near- or below-chance accuracy in at least one language context. Specifically, we set a rejection threshold of <60% accuracy, based on a discontinuity in the empirical distribution of mean accuracy in the bilinguals' dataset. Our analyses thus included 92 Chinese–English bilinguals (65 females, M = 21, SD = 1.9) and 91 native English participants (65 females, and one as non-binary, M = 23, SD = 6.9).

Participants' language background was assessed via the Language Experience and Proficiency Questionnaire (LEAP-Q; Kaushanskaya et al., 2020; Table 1) on Qualtrics (https://www.qualtrics.com). Bilingual participants reported an average of 15 ± 3 years of formal education (undergraduate or postgraduate level), high daily exposure to Chinese (M = 7.2, SD = 2.3 on a scale from 0 – never to 10 – always), and moderate daily exposure to English (M = 4.5, SD = 1.6 on a scale from 0 – never to 10 – always). They reported very good proficiency in Mandarin Chinese (M = 8.3, SD = 1.6 on a scale from 0 – none to 10 – perfect) and adequate proficiency in English (M = 5.2, SD = 1.9) (Figure 1). Native English controls reported excellent proficiency in English (M = 9.1, SD = 1 on a scale from 0 – none to 10 – perfect) and did not report having fluent knowledge of another language.

Participants signed consent forms online before taking part in this study and were compensated for their participation in either money transfer (an amount of RMB 30) or course credit (1.5 credits). This study was approved by the ethics committee of the Bangor University (approval no. 2021-17074).

2.2. Materials

The stimuli consisted of a total of 90 Chinese words and their corresponding English-language translation equivalents (see Supplementary Appendix 1). We used the translation from Oxford

Table 1. Chinese–English bilingual participants' self-reported language backgrounds

Measure	Mean	SD
Age of Chinese acquisition	1.8 years	2
Age of English acquisition	7.6 years	3.5
Daily Chinese use	80.5%	16
Daily English use	19.5%	16

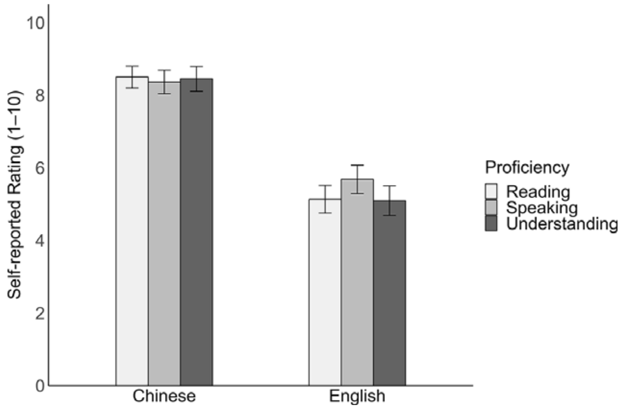

Figure 1. Chinese–English bilinguals' self-reported ratings of Chinese and English proficiency (10-point scale). Error bar represents confidence interval.

[1] The results of pilot data and the script for power analysis can be found on https://osf.io/upwf7/?view_only=2ed718f461674582876e2e543185a52b

https://doi.org/10.1017/S1366728925100588 Published online by Cambridge University Press

English Dictionary and AI translator. Among these stimuli, 60 words were considered as critical items, representing human roles with relatively high/low level of perceived power (e.g., king, servant). There were also 30 filler items consisting of animal names (e.g., dog, rabbit). In the practice session, an additional six Chinese and English words were used.

Thirty Chinese–English bilinguals who did not participate in the experiment rated an initial set of 60 Chinese words for perceived power, familiarity and valence on a five-point Likert scale. Due to low familiarity or the ambiguity of perceived power in the norming, six high-power and three low-power words were replaced prior to testing. Thus, norming results reported here did not include the replaced items. Independent sample t test showed that the perceived power of the high-power words was rated significantly higher (M = 3.69 ± 0.74) than that of low-power words (M = 1.47, SD = 0.18), $t(49) = 15.2$, $p < .001$, Cohen's $d = 4.26$. Valence ratings for high-power words were also significantly higher (M = 3.62 ± 0.44) than for low-power words ($M = 2.51$, $SD = 0.61$), $t(49) = 7.32$, $p < .001$, Cohen's $d = 2.05$, reflecting a positive correlation between the measures ($r = 0.70$, $p < .001$). In addition, neither familiarity ratings (high-power: $M = 4.25$, $SD = 0.42$; low-power: $M = 4.19$, $SD = 0.44$; $t(49) = 0.52$, Cohen's $d = .15$) nor log lexical frequency significantly differed between the high- and low-power conditions (estimated via the SUBTLEX-CH corpus, Cai & Brysbaert, 2010; high-power: $M = 2.45$, $SD = 0.53$; low-power: $M = 2.31$, $SD = 0.55$; $t(47) = 1.3$, $p = .361$, Cohen's $d = .26$).

The same 30 Chinese–English bilinguals also rated the English translation equivalents for perceived power, familiarity and valence. The perceived power of the high-power English words was rated significantly higher ($M = 3.72$, $SD = 0.59$) than that of low-power English words ($M = 1.45$, $SD = 0.17$), $t(49) = 18.57$, $p < .001$, Cohen's $d = 5.21$. Valence ratings for high-power words were significantly higher ($M = 3.43$, $SD = 0.25$) than for low-power words ($M = 2.67$, $SD = 0.49$), $t(49) = 6.84$, $p < .001$, Cohen's $d = 1.92$. Familiarity ratings did not significantly differ between the high- and low-power conditions (high-power words: $M = 4.32$, $SD = 0.45$; low-power words: $M = 4.25$, $SD = 0.46$, $t(49) = 0.54$, $p = .59$, Cohen's $d = .15$), though log lexical frequency did (estimated via the LexOPS database, Taylor et al., 2020; high-power: $M = 4.52$, $SD = 0.49$; low-power: $M = 3.88$, $SD = 0.76$; $t(49) = 3.81$, $p < .001$, Cohen's $d = 1.00$)[2].

2.3. Procedure

After reading the information sheet and signing the consent form, participants were first asked to fill the language background questionnaire (LEAP-Q; Kaushanskaya et al., 2020) on Qualtrics and completed the online experiment on Pavlovia.

The information sheet provided an overview of the study, focusing on the purpose of examining different language processing patterns between Chinese–English bilinguals and native English speakers. Participants were instructed to determine whether a word was presented with a larger or smaller font size by pressing 'y/n' or 'u/b' keys on the computer keyboard, and not to respond to animal name. No information was provided regarding perceived power or physical size. Response keys and handedness were counterbalanced between participants. Larger stimuli were presented in 28-point bold Microsoft Yahei (Chinese) or Times New Roman (English), while smaller stimuli were presented in the same font in a 17-point non-bold analogue.

Bilingual participants completed the study in both Chinese and English languages, with the order counterbalanced between the two. The stimuli were presented over four blocks (two in Chinese and two in English) preceded by six practice trials. Each word was presented only once per block, either in a larger/bold font or a smaller/thin font, representing congruent or incongruent conditions. High-power words (e.g., king) presented in a larger/bold font were in the congruent condition, while those presented in a smaller/thin font were in the incongruent condition, and vice versa for the low-power words. Each block consisted of 60 experimental trials and 30 filler trials, adding up to 360 trials in total. Native English participants completed the study only in English. The stimuli were presented once in either the congruent or incongruent condition in two blocks (180 trials).

The experiment was programmed in PsychoPy (V2022.1.3, https://www.psychopy.org/) and run online via Pavlovia (https://pavlovia.org/). Bilingual participants read instructions either in Chinese or English, depending on the order of language blocks randomly selected. They then completed a practice session in the same language and were given feedback after each practice trial. Each trial began with a 500 ms centred fixation cross, after which a stimulus was presented in the centre of the screen for up to 2000 ms, or until a response was detected. There was a self-paced break after each 90-trial block. The whole experiment lasted approximately 30 min for Chinese–English bilinguals and around 15 min for native English speakers.

At the end of the experiment, participants were asked why they thought we used different features (font size and boldness) for the stimuli and were debriefed. Two bilinguals and nine native English participants reported being aware of the association between font size and perceived power. While we did not exclude their datasets from our main analyses (because our intention was not to conceal the manipulation of perceived power, considering the simplicity of the task), doing so would not qualitatively change our claimed results (see Appendices 4 and 5).[3]

2.4. Data analyses

As planned in our pre-registration (https://aspredicted.org/Q4Q_RBL), RTs were analysed using linear mixed effects regression (*lmer* function) with the package *lme4* (Bates et al., 2015b) in R (R Core Team, 2022). Accuracy data were analysed using logistic mixed effects regression with the *glmer* function of *lme4* employing a binomial link function. All models were limited to human power words only. Only RTs for correct responses within the range of 200–1500 ms were included in the analyses based on the RT density distribution.

For each model, Congruency (congruent, incongruent) and Language (Chinese, English) or Group (bilingual, monolingual)

[2] We also recruited 30 native English participants to rate English stimulus for perceived power, familiarity and valence. The pattern of ratings and log lexical frequency for perceived power remained the same. There was no significant difference regarding valence ratings (high-power: $M = 3.47$, $SD = 0.41$; low-power: $M = 3.31$, $SD = 0.62$; $t(49) = 1.19$, $p = .24$, Cohen's $d = 0.31$). Familiarity ratings for high-power words were significantly higher ($M = 4.88$, $SD = 0.13$) than for low-power words ($M = 4.80$, $SD = 0.14$), $t(49) = 2.20$, $p = .032$, Cohen's $d = 0.57$.

[3] This suggests that the congruency detected in both groups of participants was not solely driven by explicit strategies but reflects a deeper, automatic mechanism. Furthermore, bilinguals who reported lower awareness still exhibited the expected effects, reinforcing the idea that the power–size association operates at an implicit level.

https://doi.org/10.1017/S1366728925100588 Published online by Cambridge University Press

and their interactions were considered as centred fixed effects. All models initially included the maximal random effects structures allowed by the design, omitting correlations among random effects to facilitate convergence (Barr et al., 2013). Thus, Language and its interactions were modelled as within-subjects but between-items effects, while Group and its interactions were modelled as within-items but between-subjects effects and Congruency was modelled as both a within-subjects and within-items effect. Following the recommendations of Bates et al. (2015a), we employed a parsimonious approach: If a model failed to converge, we used the *lme4*::rePCA function to remove small variance parameters until the model adequately fit the data (for the final analysis scripts, see https://osf.io/upwf7/?view_only=2ed718f461674582876e2e543185a52b).

The *p* values were calculated with the *lmerTest* package (Kuznetsova et al., 2017), using the Satterthwaite approximation to estimate effective degrees of freedom. The significance threshold was set at .05. Pairwise comparisons, if reported, were conducted using the *emmeans* package (Lenth, 2020) in R, with the Bonferroni correction applied to account for multiple comparisons. All models are reported in full in the Supplementary Materials; non-significant main effects and interactions, and effects that were expectedly reproduced across nested models are reported there but not discussed in the main text.

3. Results

3.1. Results based on pre-registration

3.1.1. Accuracy

Only responses to the critical items (human power words) were included in the analysis of accuracy data. The data were also cleaned by removing trials with RTs shorter than 200 ms or longer than 1500 ms as a *priori* improbable values for valid measurements based on the RT density distribution. Of the trials, 6% were removed from the bilingual group and 13% of the trials were removed from the native group. Those removed trials were not counted in the errors, which resulted in an observation of 20,694 data points for the bilingual group, and 9417 data points for the native group in the regression model. Within those data points, there were 817 errors (4% of the trials) for the bilingual group and 539 errors (6% of the trials) for the native group by our counts.

Bilingual participants responded more accurately to Chinese (L1) than English (L2) stimuli overall (β = −0.62, *SE* = .15, *z* = −4.20, 95% CI [−0.91, −0.33], *p* < .001), and were more accurate when responding to congruent than incongruent trials (β = −0.65, *SE* = .10, *z* = −6.32, 95% CI [−0.86, −0.45], *p* < .001). The interaction between Language and Congruency was not significant (β = .31, *SE* = .19, *z* = 1.59, 95% CI [−0.07, 0.69], *p* = .111, Figure 2A).

Comparing the Chinese–English bilinguals' performance in English (L2) to that of native English speakers showed a significant interaction between Congruency and Group (β = −.29, *SE* = .14, *z* = −2.13, 95% CI [−0.56, −0.02], *p* = 0.034, Figure 2), such that both groups responded more accurately to congruent than incongruent trials, but the congruency effect was larger in native English controls (β = 0.77, *SE* = .12, *z* = 6.47, 95% CI [0.54, 1.00], *p* < .001) than in bilingual participants (β = .48, *SE* = .12, *z* = 3.91, 95% CI [0.24, 0.72], *p* < .001). There were also main effects of Congruency (β = −0.63, *SE* = .10, *z* = −6.25, 95% CI [−0.82, −0.43], *p* < .001) and Group, such that L2 bilinguals were more accurate than native English speakers overall (β = −.19, *SE* = .07, *z* = −2.58, 95% CI [−0.33, −0.05], *p* = .001, Figure 2B).

3.1.2 Response *times*

Only correct responses to the critical items (human power words) were included in the analysis of mean response times (RT), which resulted in removal of 4% of the trials from the bilingual group, and 6% of the trials from the native. Then RTs were transformed to produce a normal distribution residual according to a Box–Cox test. Trials with log-transformed RTs more than ±2.5 standard deviations from each participant's condition mean were excluded as outliers (0.4% for bilingual group and 0.6% for native group). A total number of 19,802 data points for the bilingual group and 8817 data points for the native group was inspected in the regression model.

The bilingual participants responded faster to Chinese (L1) than English (L2) stimuli overall (β = .21, *SE* = .02, *t* = 12.69, 95% CI [0.18, 0.24], *p* < .001, Figure 3A), and were faster when responding

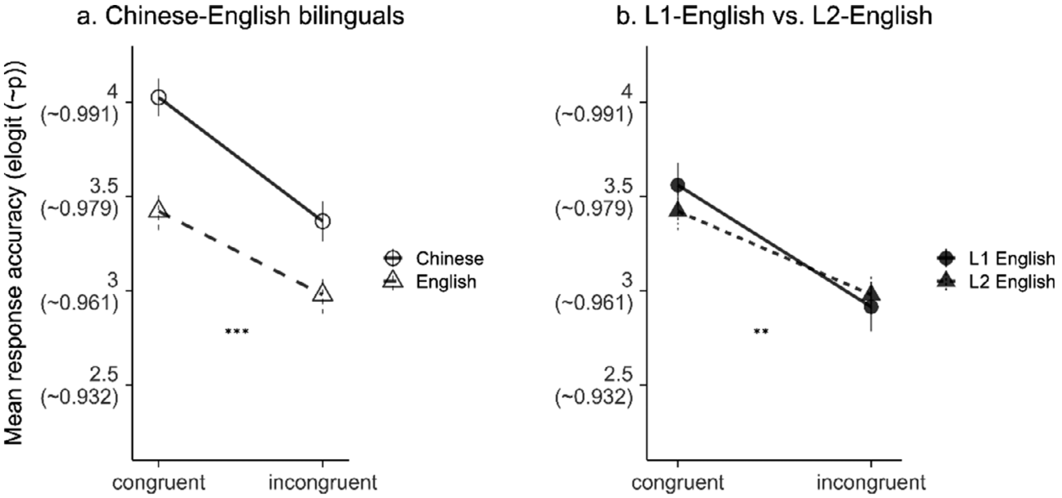

Figure 2. Mean accuracy by Congruency condition for (A) Chinese–English bilinguals in each language and for (B) the group comparison of bilinguals in English (L2) and native English controls in English (L1, right panel). To match the logistic regression analyses, accuracy is calculated as an empirical logit, with additional labels on the *y*-axis providing approximate back-transformed proportion values.

https://doi.org/10.1017/S1366728925100588 Published online by Cambridge University Press

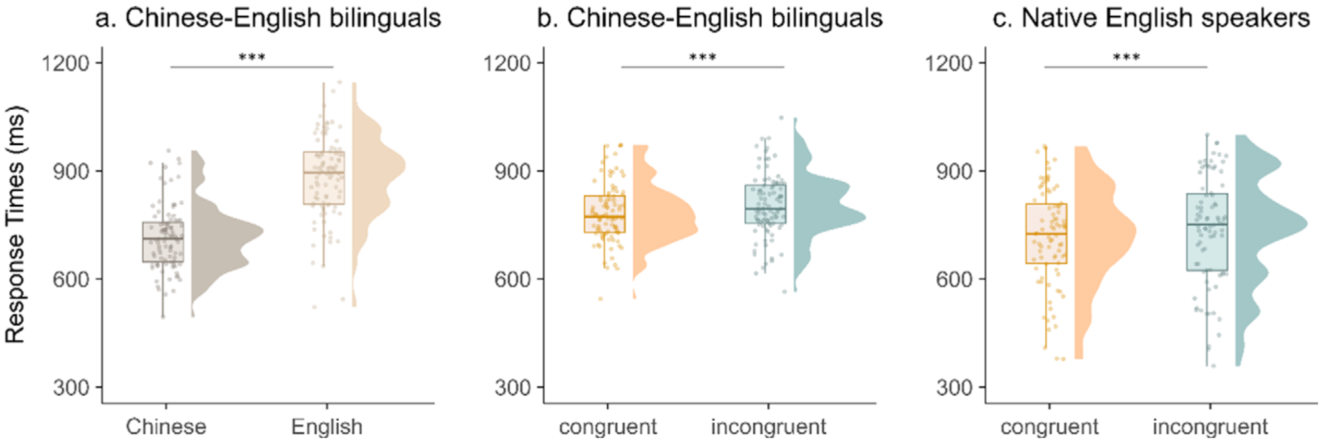

Figure 3. Box and density plots of response times of (A) main effect of Language; (B) main effect of Congruency in Chinese–English bilinguals and (C) main effect of Congruency in native English controls.

to congruent than incongruent trials (*β* = .03, *SE* = .01, *t* = 3.56, 95% CI [0.01, 0.04], *p* < .001, Figure 3B). Again, the interaction between Language and Congruency was not significant (*β* = −.01, *SE* = .01, *t* = −0.53, 95% CI [−0.04, 0.02], *p* = .6).

The cross-group analysis showed significant difference when both groups of participants responded to congruent and incongruent trials (*β* = .02, *SE* = .01, *t* = 2.67, 95% CI [0.01, 0.04], *p* = .009, Figure 3C), and bilingual participants were slower overall (*β* = −.21, *SE* = .00, *t* = −49.21, 95% CI [−0.22, −0.20], *p* < .001, Figure 3B,C). The interaction between Congruency and Group was not significant (*β* = .00, *SE* = .01, *t* = .02, 95% CI [−0.01, 0.02], *p* = 0.99).

3.1.3 Intermediate discussion

Thus, our pre-registered analyses provided little evidence to support the idea that embodiment effects differ across languages in bilingual speakers. However, when comparing bilinguals' performance in English (L2) to that of native English controls in English (L1), we found a stronger congruency effect in native English controls. As mentioned in the Introduction, Wei et al. (2024) recently suggested that power-presentation congruency effects may be stronger for words associated with high power than words associated with low power. To assess the possibility of such a pattern in our behavioural data, we extended the models in our previous analyses to include Power and its interactions as between-items fixed effects.

3.2. Follow-up analyses (not pre-registered)

Logistic mixed effects regressions modelled accuracy rate and linear mixed effects regressions modelled RTs as a function of three predictors: Language (Chinese, English), Congruency (congruent, incongruent) and a new predictor, Power (high power, low power). As before, all models included the maximal random effects structures allowed by the design, omitting correlations among random effects to facilitate convergence. Thus, Language, Power and their interactions were modelled as within-subjects but between-items effects, while Congruency was modelled as both a within-subjects and within-items effect.

3.2.1. Accuracy

For bilingual participants, the logistic regression of the accuracy data revealed a significant Power × Language × Congruency interaction (*β* = −1.27, *SE* = 0.52, *z* = −2.45, 95% CI [−2.29, −0.25], *p* = .014, Figure 4A), such that high-power Chinese words elicited the strongest congruency effect (*β* = 2.05, *SE* = .30, *z* = 6.94, 95% CI [1.47, 2.63], *p* < .001) as compared to low-power Chinese words (*β* = .01, *SE* = .23, *z* = .07, 95% CI [−0.43, 0.46], *p* = .948) and high-power English words (*β* = 0.87, *SE* = .21, *z* = 4.27, 95% CI [0.47, 1.28], *p* < .001; see Supplementary Appendix 3 for the complete set of comparisons). The main effects of Language and Congruency that we detected in the pre-registered analysis also emerged here. There was a significant two-way interaction between Language and Congruency (*β* = 0.54, *SE* = .20, *z* = 2.64, 95% CI [0.14, 0.94], *p* = .008), such that participants made fewer errors on congruent than incongruent trials overall, but this difference was more significant when they were tested in Chinese (*β* = 1.03, *SE* = .17, *z* = 6.12, 95% CI [0.70, 1.36], *p* < .001) than in English (*β* = .49, *SE* = .12, *z* = 4.02, 95% CI [0.25, 0.74], *p* < .001). We also found a main effect of Power (*β* = −.21, *SE* = .10, *z* = −2.08, 95% CI [−0.41, −0.01], *p* = .038). The interaction between Power and Congruency was also significant (*β* = 1.40, *SE* = .26, *z* = 5.35, 95% CI [0.88, 1.91], *p* < .001), such that bilingual participants made fewer errors on congruent than incongruent trials in response to high-power words (*β* = 1.46, *SE* = .18, *z* = 8.04, 95% CI [1.21, 2.02], *p* < .001), while no such difference emerged for low-power words (*β* = .06, *SE* = .15, *z* = .42, 95% CI [−0.24, 0.37], *p* = .675). Moreover, there was a significant interaction between Language and Power (*β* = .49, *SE* = .21, *z* = 2.39, 95% CI [−0.00, 0.87], *p* = .017). Pairwise comparisons showed that participants made fewer errors for high-power than low-power words when tested in Chinese (*β* = .46, *SE* = .17, *z* = 2.76, 95% CI [0.13, 0.79], *p* = .006), but not when tested in English (*β* = −.03, *SE* = .12, *z* = −.28, 95% CI [−0.27, 0.20], *p* = .782).

Concerning accuracy in English across groups, we found a significant three-way interaction between Power, Congruency and Group (*β* = .97, *SE* = .29, *z* = 3.32, 95% CI [0.40, 1.54], *p* < .001, Figure 4B). Although the main effect of Power was not significant (*p* = .553), participants were generally more accurate on congruent than incongruent trials overall (*β* = −0.65, *SE* = .09, *z* = −7.16, 95% CI [−0.82, −0.47], *p* < .001). The main effect of Group was not significant either (*β* = −.13, *SE* = .08, *z* = −1.79, 95% CI [−0.28, 0.01], *p* = .074). A significant Power × Congruency interaction (*β* = 1.08, *SE* = .22, *z* = 4.89, 95% CI [0.65, 1.51], *p* < .001) reflected that a congruency effect found for high-power words (*β* = 1.19, SE = .15, *z* = 8.14, 95% CI [0.90, 1.47], *p* < .001) failed to reach significance in the case of low-power words (*p* = .430), and this pattern was stronger for the L1 group (high-power words:

https://doi.org/10.1017/S1366728925100588 Published online by Cambridge University Press

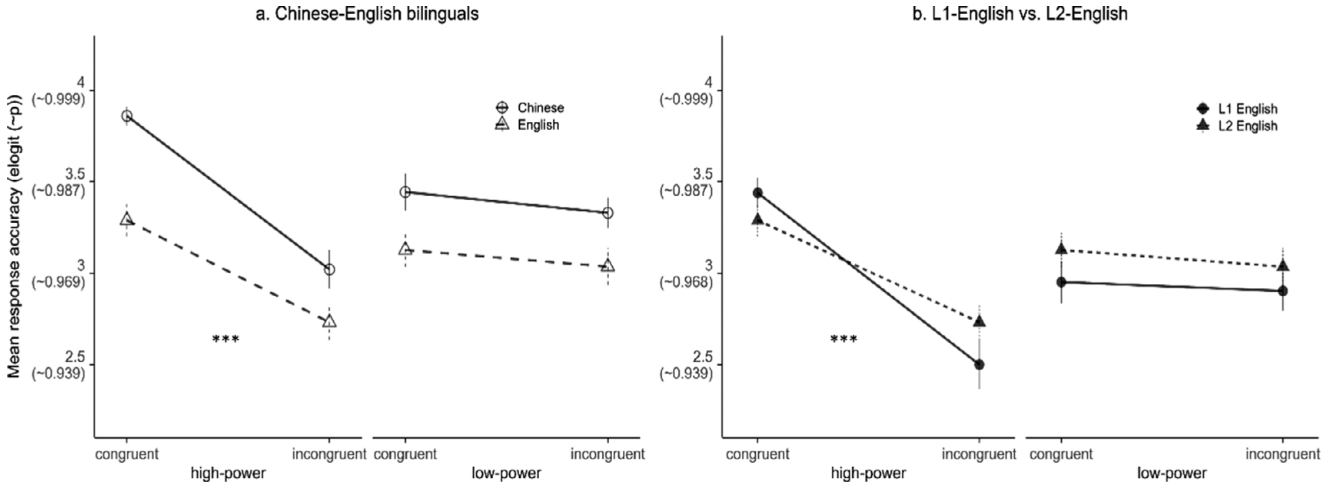

Figure 4. Mean accuracy by Congruency condition (A) for Chinese–English bilinguals in each language and (B) for the group comparison of bilinguals in English (L2) and native English controls in English (L1), for high-power (left panel) and low-power words (right panel).

β = 1.60, *SE* = .19, *z* = 8.64, 95% CI [1.24, 1.97], *p* < .001; low-power words: *β* = .04, *SE* = .16, *z* = 0.28, 95% CI [−0.28, 0.37], *p* = .783) than for the L2 group (high-power words: *β* = 0.77, *SE* = .18, *z* = 4.40, 95% CI [0.43, 1.11], *p* < .001; low-power words: *β* = .17, *SE* = .17, *z* = 1.00, 95% CI [−0.17, 0.51], *p* = .316).

3.2.2. Response times

The expanded linear mixed effects regression model of RTs from bilingual participants revealed a significant interaction between Power and Congruency (*β* = −.07, *SE* = .02, *t* = −4.44, 95% CI [−0.10, −0.04], *p* < .001, Figure 5A), such that the congruency effect was stronger for high-power words (*β* = −.06, *SE* = .01, *z* = −5.87, 95% CI [−0.37, −0.18], *p* < .001) than low-power words (*β* = .01, *SE* = .01, *z* = 0.76, 95% CI [−0.06, 0.13], *p* = .449). There was a main effect of Power, such that high-power elicited faster RTs than low-power words (*β* = .03, *SE* = .01, *t* = 4.10, 95% CI [0.01, 0.04], *p* < .001). The main effects of Language and Congruency found in the pre-registered analysis were also found here. No other interactions were significant (*ps* > .1).

Comparing L2 English to L1 English RTs revealed a significant two-way interaction between Power and Congruency (*β* = −.06, *SE* = .02, *t* = −3.75, 95% CI [−0.09, −0.03], *p* < .001, Figure 5B), underpinned by a strong congruency effect for high-power words (*β* = −.05, *SE* = .01, *t* = −4.75, 95% CI [−0.28, −0.11], *p* < .001) that was not significant for low-power words (*β* = .01, *SE* = .01, *t* = 0.56, 95% CI [−0.06, 0.11], *p* = .578). The analysis also showed that participants responded significantly faster to high- than low-power words (*β* = .03, *SE* = .01, *t* = 4.10, 95% CI [0.02, 0.05], *p* < .001), in addition to the previously reported main effects of Congruency and Group. The three-way interaction between Power, Congruency and Group that we detected in the accuracy data was not significant (*β* = −.01, *SE* = .02, *t* = 0.51, 95% CI [−0.04, 0.02], *p* = .611).

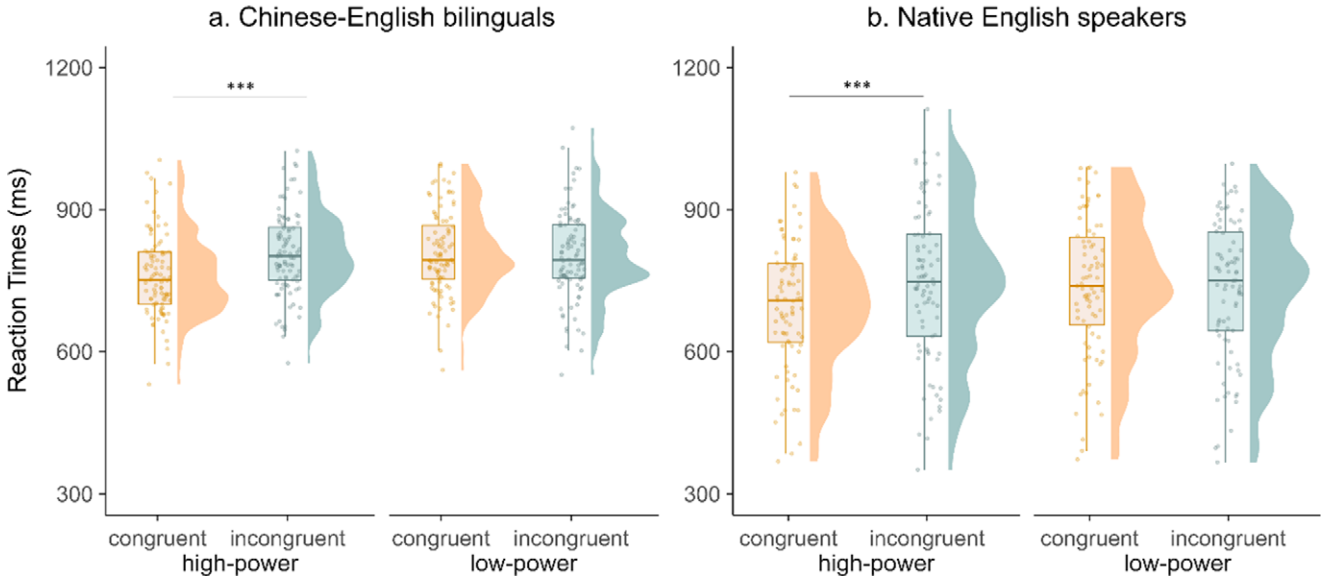

Figure 5. Box and density plots of mean response times by Congruency conditions for high- and low-power words for (A) Chinese–English bilinguals and (B) English native controls.

https://doi.org/10.1017/S1366728925100588 Published online by Cambridge University Press

3.3. Exploratory analyses (adding valence as a predictor)

To exclude the possibility that differences in item valence might have driven the congruency effect detected here, we further extended our accuracy and RT models by adding mean valence ratings for each word (categorical factor) from the norming process described in our Methods section, as a centred fixed effect with by-participant random slopes. Variance Inflation Factor (VIF) values for all predictors in our models were also computed to address possible multicollinearity, ensuring more reliable coefficient estimates and interpretable results. Finally, following the reviewer's suggestion, we conducted a supplementary analysis in which we inverted the roles of power and valence, using valence as a factor and power as a covariate for both accuracy and RT models (see Supplementary Appendix 6 for the report and the complete set of comparisons).

3.3.1. Accuracy

After adding Valence rating as another predictor in our logistic model, neither the main effect of Valence (β = −.06, SE = .11, z = −.48, 95% CI [−0.28, 0.17], p = .628) nor any interaction involving it reached significance (all ps > .3), and all shared parameters remained reasonably close to their estimates from the previous model. We found significant two-way interactions between Language and Congruency and Power and Congruency as reported earlier. Most importantly, the three-way interaction between Language, Power and Congruency remained significant (β = −1.88, SE = 0.69, z = −2.74, 95% CI [−3.22, −0.53], p = .006). Pairwise comparison showed that high-power words elicited the strongest congruency effect when participants were tested in Chinese (β = 2.19, SE = .38, z = 5.71, 95% CI [0.99, 2.03], p < .001) as compared to when they were tested in English (β = 0.56, SE = .26, z = 2.14, 95% CI [0.05, 1.06], p = .032). No such difference was detected for low-power words (ps > .2, see Supplementary Appendix 3 for the complete set of comparisons).

The VIF values for all predictors in this model ranged from 1.11 to 2.33, indicating a low to moderate degree of multicollinearity among the predictors (see Supplementary Appendix 3 for the value corresponding to each predictor). Importantly, the addition of Valence as a predictor (VIF = 2.08) did not substantially increase VIF values for the other factors. This suggests that the inclusion of Valence in the model did not significantly contribute to multicollinearity, and thus, the Language × Power × Congruency interaction detected here was not driven by Valence.

3.3.2. Response times

When adding item Valence, as in the corresponding accuracy analysis, the two-way interaction between Power and Congruency remained significant (β = −.08, SE = .02, t = −3.77, 95% CI [−0.12, −0.04], p < .001) in line with the previous results (high-power words: β = −.07, SE = .01, t = −5.03, 95% CI [−0.40, −0.18], p < .001; low-power words: β = .01, SE = .01, t = 1.13, 95% CI [−0.05, 0.17], p = .259). We also detected a significant interaction between Language and Valence (β = −.04, SE = .02, t = −2.75, 95% CI [−0.07, −0.01], p = .007), reminiscent of the correlation between power and valence ratings for Chinese items (see Methods). The main effects of Language and Congruency reported earlier were also found here, but the main effect of Power was not significant (β = .02, SE = .01, t = 1.75, 95% CI [−0.00, 0.04], p = .082). In addition, we found a main effect of Valence (β = −.02, SE = .01, t = −2.99, 95% CI [−0.04, −0.01], p = .003) with faster RTs for more positive (high-power) words than more negative (low-power) words. No other interactions involving Valence as a predictor was significant (ps > .5).

Again, we calculated the VIF values for the predictors in our linear mixed effect model to avoid the potential multicollinearity issues. The VIF values range from 1.01 to 2.19, indicating a low to moderate degree of multicollinearity among the predictors (see Supplementary Appendix 3 for the corresponding value for each predictor). Importantly, the addition of the Valence predictor did not substantially increase the VIF values, with the VIF for Valence being 2.19. This suggests that the Power × Congruency interaction we found from RTs was not driven by the valence of items. The inclusion of Valence in the model did not qualitatively change any claimed result.

4. Discussion

In this study, we investigated how language of operation affects behavioural correlates of perceived power embodiment in Chinese–English bilinguals and further compared their performance in L2 English to the performance of control participants in L1 English. Our pre-registered analyses failed to detect our hypothesized two-way interaction between language and congruency. Bilingual participants responded more quickly and accurately in their L1 (Chinese) than in their L2 (English) and on congruent trials (high-power words presented in a bold and larger font) than incongruent trials (high-power words presented in a thin and smaller font). The lack of a significant interaction would seem to suggest that the congruency effects were not different across the two languages.

However, recent developments in the field suggest that our pre-registered analysis plan may have been too simplistic. For example, Wei et al. (2024) reported an electrophysiological effect of congruency that was significantly stronger for high-power than low-power words. Although they did not detect any behavioural analogue of that effect, it stands to reason that our analyses may benefit from a similar distinction. We therefore extended our statistical models to include Power as a between-items predictor that could interact with the other fixed effects.

The results of the extended models revealed clear power–size congruency effects in relation to both speed and accuracy. These effects were stronger for high-power than low-power words, with *post hoc* tests indicating significant congruency effects only for the former. Interestingly, a three-way interaction between language, power and congruency showed that the congruency effect for accuracy was stronger when bilinguals used their L1 (Chinese) compared to their L2 (English). Although this interaction only emerged in the accuracy data, it provides the first behavioural analogue to the electrophysiological effect reported by Wei et al. (2024). It should be noted that the lack of an RT effect cannot be interpreted as evidence for the absence of an effect, but rather that the manipulation we have chosen to implement in this experiment failed to affect RTs. Moreover, the Power × Congruency interaction on accuracy was numerically stronger for native English speakers tested in their L1 than bilinguals tested in their L2. Thus, these interactions seem most consistent with the idea that embodiment effects are attenuated in a speaker's second language.

The observed congruency effect on power words in both Chinese–English bilinguals and native English participants supports our initial prediction about the embodiment of perceived power and its metaphorical mapping to physical size. This relationship reflects how physically larger individuals are often perceived as more dominant, an association that extends to abstract representations

https://doi.org/10.1017/S1366728925100588 Published online by Cambridge University Press

of power (Lakoff & Johnson, 1999). Our findings align with previous research (He et al., 2015; Schubert et al., 2009), which is noteworthy given that conceptual attributes of the stimuli were not highlighted by our task: participants did not make overt judgements about perceived power and received no direct instruction about it, but only reported perceptual attributes of the stimuli.

Unlike Wei et al. (2024), who used auditory stimuli to convey vertical position as the embodied reference and only found a congruency effect in ERPs, here we detected such effect at a behavioural level, suggesting a more salient association between perceived power and physical size. This is consistent with research indicating that using physical size to mentally represent social dominance is an innate cognitive ability emerging before language acquisition (Thomsen et al., 2011). Thomsen et al. got infants to observe animated scenarios where two agents differing in physical size attempted to pass each other on a narrow path. Infants exhibited longer gaze durations when a larger agent unexpectedly bowed and yielded to a smaller one as compared to the reverse. Align with this, De Koning et al. (2017) found that physical size is more likely to be activated than spatial information when participants read sentences implying shape, size, colour or orientation of objects. Their sentence–picture verification task revealed varying match advantages across properties, with colour exhibiting the strongest effect, followed by shape, then size, while orientation showed no advantage. This may explain why behavioural effects were found in this study (based on size), whereas Wei et al. failed to detect any (based on orientation).

Another difference from Wei et al. (2024) lies in our observation of embodiment effects in native English participants, which were not detected either behaviourally or in the ERP data. This lack of a significant effect, which they attributed to potential cultural differences between Chinese and English, may have been due to the cross-modal nature of the physical dimension chosen to embody perceived power: auditory stimuli presented from above or below the participant's seating position. This required cross-modal integration and remapping because the origin of sound is not directly linked to visuospatial experience. In addition, note that locating sound origin on the vertical axis is more challenging and less natural than locating sound on a horizontal axis (up – down vs. left – right, e.g., Middlebrooks & Green, 1991). These factors may have contributed to the absence of embodiment effects, while this study associated power and weakness with the more intuitive property of physical size.

We note that our first analysis, which did not distinguish between high- and low-power words, may have blurred the underlying congruency patterns. When we incorporated this distinction in our follow-up analyses, a significant congruency effect emerged, particularly for high-power words. This asymmetrical pattern aligns with previous EEG findings showing increased brain activity when participants respond to congruent than incongruent trials for high-power but not for low-power words (Wei et al., 2024; Wu et al., 2016). While this asymmetry might initially appear to be compatible with a polarity-based explanation (Lakens, 2012; Proctor & Cho, 2006), several aspects of our experimental design cast doubt on such interpretation.

First, our paradigm fundamentally differs from the binary classification tasks typically used in polarity correspondence studies. While those studies have participants explicitly categorize stimuli based on conceptual dimensions, creating direct alignment between stimulus categories and response options (e.g., Lakens, 2012), our participants performed a more perceptually driven task where they judged font size only for non-animal words. This task directed attention primarily to perceptual features, with conceptual processing limited to the animal/non-animal distinction (only implemented to ensure semantic processing of stimuli). Any activation of power-related concepts would thus have been more incidental rather than directly prompted by the task. While one bilingual participant and nine native participants reported awareness of the power-font size association, excluding them from the main analysis did not qualitatively change our results (see Appendices 4 and 5).

Second, the polarity correspondence principle relies on a structured alignment between stimulus and response options, often involving dichotomous categories (e.g., positive/negative, high/low) and explicit mappings (e.g., 'positive' to 'UP' and 'negative' to 'DOWN' response keys). In our study, however, the stimuli were not dichotomous in nature but instead referred to a variety of concrete exemplars on a perceived power continuum (e.g., 'professor', 'employee'). These concepts do not necessarily refer to physically large or small entities but are metaphorically associated with dominance or subservience. Moreover, our task design deliberately avoided explicit structured mappings: response keys were associated with vertical positions ('y'/'n' or 'u'/'b' key pressing) rather than physical size (e.g., we did not use large and small response buttons). This lack of explicit alignment between stimulus and response dimensions makes a polarity-based account of our results unlikely.

The asymmetrical pattern between high-power and low-power words in our results remains compatible with Lakens et al.'s (2011) findings regarding the contextual nature of power representation. Lakens et al. showed that power–space congruency effects are robust in within-subjects designs where both power levels are presented (Exp. 1A, 2A), but weaken or even disappear when high- or low-power words are presented alone (Exp. 1B, 2B). The idea is that power concepts are inherently relational, with high-power words requiring a low-power reference point for optimal processing. Notably, Lakens et al. observed that manipulating the coexistence of high- and low-power words primarily affected the congruency effect for high-power words, while the pattern for low-power words remained relatively stable. They proposed that while 'above' serves as the default endpoint on the vertical axis, with powerful groups typically represented above powerless reference groups, this representational structure is context-dependent. The stronger congruency effects we observed for high-power words likely reflect both this relational nature of power representation and specific characteristics of our participant population. While words like 'president', 'boss' and 'director' consistently evoke clear associations with superior status and dominance, low-power words such as 'student', 'clerk' and 'employee' may not trigger low power associations for our participants, many of whom can identify with such roles. This baseline consideration may explain why congruency effects were less pronounced for low-power words.

Another argument why the effects reported here can be related to embodiment effects concerns cross-linguistic differences. Consistent with previous studies (e.g., Qian, 2016; Wei et al., 2024), we found that perceived power representation is more strongly grounded in L1 than L2. Such cross-linguistic differences are not predicted by the polarity-based account. The weaker embodiment of perceived power in L2 might reflect bilinguals' tendency to process L2 words more literally than L1 words (Kroll & Tokowicz, 2005). Thus, in L2, the instructions might have semantically permeated to the domain of physical size evaluation (Is a king physically bigger than a servant?), whereas the interpretation is more likely to concern the metaphorical level in L1 (Is a King bigger, as in more powerful, than a servant?).

https://doi.org/10.1017/S1366728925100588 Published online by Cambridge University Press

Another consideration is related to L2 acquisition and uses. Bilinguals who live in an L1-dominant environment, primarily acquire and use their L2 in formal and restricted contexts such as school, emphasizing vocabulary and grammar. Consequently, metaphorical associations between perceived power and physical size may be less prevalent in the L2 due to limited real-life exposure. This perspective aligns with the Words as Social Tools (WAT) hypothesis (Borghi et al., 2019), which posits that abstract concepts are acquired and understood through linguistic and social interactions that provide contextual cues and pragmatic information. Late bilinguals encounter abstract words less frequently and in less diverse situations in their L2 than their L1, relying both on sensorimotor simulations and symbolic information. Alternatively, it could be argued that the embodiment effect is modulated by the degree of conceptual overlap between L1 and L2, as proposed by the RHM (Kroll & Stewart, 1994). Given that our participants reported moderate English proficiency, the link between L2 words and their conceptual representation may not have been sufficiently robust to afford full access to embodied representations.

The significant interaction between group, power and congruency, with stronger congruency effects in native English speakers than in bilinguals remains to be explained. While both English and Chinese languages employ metaphors linking perceived power with physical size (e.g., 'big boss' in English; '*yishou zhetian*' [covering the sky with one hand] in Chinese), the salience and frequency of these metaphors likely differ across languages. In English-speaking contexts, such metaphors are deeply embedded in everyday language, whereas in Chinese, they tend to appear more frequently in formal or written contexts. This could explain why the congruency effect was twice as large in native English speakers compared to Chinese–English bilinguals tested in their L2.

Our study has several limitations that could be addressed in the future. First, we manipulated both font size and boldness of the stimuli simultaneously. Further studies could manipulate these two factors orthogonally to disentangle the individual contribution of font size and boldness to the embodiment of power. Second, while including Valence as a predictor showed that valence alone cannot account for the observed congruency effect, the relationship between power and valence requires further investigation. Unlike binary valence distinctions (i.e., positive vs. negative), valence ratings in our stimuli varied continuously, with low-power words being overall mildly negative (average rating 2.51 out of 5). Words like *employee*, *student* and *clerk* had lower power ratings than *president*, *king* or *professor* without necessarily carrying negative connotations. Future studies could explore how valence and/or arousal interact with power embodiment, particularly examining whether congruency effects differ between 'negative' low-power words (e.g., slave, servant) and 'positive' ones (e.g., helper).

5. Conclusion

In order to test how embodied cognition applies to abstract concepts and its potential involvement in second language learning, we investigated behavioural evidence for the embodiment of perceived power in Chinese–English bilinguals and native English speakers. First, we showed that a mildly abstract concept such as perceived power shows effect consistent with predictions from embodiment theory in both languages of bilinguals, even when their two languages strongly differ from a typological viewpoint. Furthermore, we found that embodiment effects linking perceived power to physical size are stronger in a bilingual's L1 than their L2. We contend that differences in embodiment-related congruency effects in bilinguals relate to (i) differences in preferred access to knowledge (literal vs. metaphorical) in L1 and L2, and (ii) differences in acquisition and exposure between languages. In addition, there may be contributions of (iii) socio-cultural and linguistic idiosyncrasies of the languages involved. This study sheds light on the complex interplay between language, cognition and possibly culture, highlighting the need for a comprehensive approach to studying embodied cognition in diverse linguistic and cultural contexts.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S1366728925100588.

Acknowledgements. We thank anonymous reviewers for their constructive comments on our manuscript. Y.W. is supported by the China Scholarship Council (2021[41], 202106070030). G.T. was supported by the Polish National Agency for Academic Exchange (NAWA) under the NAWA Chair Programme (PPN/PRO/2020/1/00006).

Data availability statement. The data that support the findings of this study are openly available in Open Science Framework at https://osf.io/upwf7/?view_only=2ed718f461674582876e2e543185a52b

Competing interests. The authors declare none.

References

Ahlberg, D. K., **Bischoff, H.**, **Kaup, B.**, **Bryant, D.**, & **Strozyk, J. V.** (2018). Grounded cognition: Comparing language × space interactions in first language and second language. *Applied PsychoLinguistics*, **39**(2), 437–459. https://doi.org/10.1017/S014271641700042X.

Bai, Y., & **He, W.** (2022). Involvement of the sensorimotor system in less advanced L2 processing: Evidence from a semantic category decision task. *Frontiers in Psychology*, **13**. https://doi.org/10.3389/fpsyg.2022.980967.

Barr, D. J., **Levy, R.**, **Scheepers, C.**, & **Tily, H. J.** (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, **68**(3), 255–278. https://doi.org/10.1016/J.JML.2012.11.001.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, **22**(4), 577–609. 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.

Bates, D., **Kliegl, R.**, **Vasishth, S.**, & **Baayen, H.** (2015a). *Parsimonious mixed models*. http://arxiv.org/abs/1506.04967

Bates, D., **Mächler, M.**, **Bolker, B. M.**, & **Walker, S. C.** (2015b). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1). https://doi.org/10.18637/jss.v067.i01.

Bergen, B., **Lau, T. T. C.**, **Narayan, S.**, **Sto Janovic, D.**, & **Wheeler, K.** (2010). Body part representations in verbal semantics. *Memory and Cognition*, **38**(7), 969–981. https://doi.org/10.3758/MC.38.7.969.

Borghi, A. M., **Barca, L.**, **Binkofski, F.**, **Castelfranchi, C.**, **Pezzulo, G.**, & **Tummolini, L.** (2019). Words as social tools: Language, sociality and inner grounding in abstract concepts. *Physics of Life Reviews*, **29**, 120–153. https://doi.org/10.1016/j.plrev.2018.12.001.

Buccino, G., **Marino, B. F.**, **Bulgarelli, C.**, & **Mezzadri, M.** (2017). Fluent speakers of a second language process graspable nouns expressed in L2 like in their native language. *Frontiers in Psychology*, **8**. https://doi.org/10.3389/fpsyg.2017.01306.

Cai, Q., & **Brysbaert, M.** (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One*, **5**(6), e10729. https://doi.org/10.1371/journal.pone.0010729.

De Grauwe, S., **Willems, R. M.**, **Rueschemeyer, S. A.**, **Lemhöfer, K.**, & **Schriefers, H.** (2014). Embodied language in first- and second-language speakers: Neural correlates of processing motor verbs. *Neuropsychologia*, **56**(1), 334–349. https://doi.org/10.1016/j.neuropsychologia.2014.02.003.

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.

Dijkstra, T., & **van Heuven, W. J. B.** (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, **5**(3), 175–197. https://doi.org/10.1017/s1366728902003012.

Dudschig, C., **de la Vega, I.**, & **Kaup, B.** (2014). Embodiment and second-language: Automatic activation of motor responses during processing spatially associated L2 words and emotion L2 words in a vertical Stroop paradigm. *Brain and Language*, **132**, 14–21. https://doi.org/10.1016/j.bandl.2014.02.002.

Foroni, F. (2015). Do we embody second language? Evidence for "partial" simulation during processing of a second language. *Brain and Cognition*, **99**, 8–16. https://doi.org/10.1016/j.bandc.2015.06.006.

Gallese, V., & **Lakoff, G.** (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, **22**(3–4), 455–479. https://doi.org/10.1080/02643290442000310.

Glenberg, A. M., & **Kaschak, M. P.** (2002). Grounding language in action. *Psychonomic Bulletin & Review*, **9**(3), 558–565. https://doi.org/10.3758/BF03196313.

Hauk, O., **Johnsrude, I.**, & **Pulvermüller, F.** (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, **41**(2), 301–307. https://doi.org/10.1016/S0896-6273(03)00838-9

He, X., **Chen, J.**, **Zhang, E.**, & **Li, J.** (2015). Bidirectional associations of power and size in a priming task. *Journal of Cognitive Psychology*, **27**(3), 290–300. https://doi.org/10.1080/20445911.2014.996155.

Kaushanskaya, M., **Blumenfeld, H. K.**, & **Marian, V.** (2020). The language experience and proficiency questionnaire (LEAP-Q): Ten years later. *Bilingualism*, **23**(5), 945–950. https://doi.org/10.1017/S1366728919000038.

Kroll, J. F., & **Stewart, E.** (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, **33**(2), 149–174. https://doi.org/10.1006/jmla.1994.1008.

Kroll, J. F., & **Tokowicz, N.** (2005). Models of bilingual representation and processing: Looking back and to the future. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 531–553). New York: Oxford University Press.

Kühne, K., & **Gianelli, C.** (2019). Is embodied cognition bilingual? Current evidence and perspectives of the embodied cognition approach to bilingual language processing. *Frontiers in Psychology*, **10**, 108. https://doi.org/10.3389/fpsyg.2019.00108.

Kuznetsova, A., **Brockhoff, P. B.**, & **Christensen, R. H. B.** (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, **82**(13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Lakens, D. (2012). Polarity correspondence in metaphor congruency effects: Structural overlap predicts categorization times for bipolar concepts presented in vertical space. *Journal of Experimental Psychology: Learning Memory and Cognition*, **38**(3), 726–736. https://doi.org/10.1037/a0024955.

Lakens, D., & **Caldwell, A. R.** (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, **4**(1), 1–16. https://doi.org/10.1177/2515245920951503.

Lakens, D., **Semin, G. R.**, & **Foroni, F.** (2011). Why your highness needs the people: Comparing the absolute and relative representation of power in vertical space. *Social Psychology*, **42**(3), 205–213. https://doi.org/10.1027/1864-9335/a000064.

Lakoff, G., & **Johnson, M.** (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. Basic Books.

Lenth, R.V. (2020) Emmeans: Estimated marginal means. Aka least-squares means R package version 1.5.3.

Meier, B. P., & **Robinson, M. D.** (2004). *Why the sunny side is up associations between affect and vertical position.*

Middlebrooks, J. C., & **Green, D. M.** (1991). Sound localization by human listeners. *Annual Review of Psychology*, **42**, 135–159. https://doi.org/10.1146/annurev.ps.42.020191.001031.

Monaco, E., **Jost, L. B.**, **Gygax, P. M.**, & **Annoni, J. M.** (2019). Embodied semantics in a second language: Critical review and clinical implications. *Frontiers in Human Neuroscience*, **13**. https://doi.org/10.3389/fnhum.2019.00110.

Pecher, D., **Zeelenberg, R.**, & **Barsalou, L. W.** (2003). *Verifying different-modality properties for concepts produces switching costs*. Psychological Science. https://doi.org/10.1111/1467-9280.t01-1-01429.

Proctor, R. W., & **Cho, Y. S.** (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin*, **132**(3), 416–442. https://doi.org/10.1037/0033-2909.132.3.416.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews. Neuroscience*, **6**(7), 576–582. https://doi.org/10.1038/nrn1706.

Qian, W. (2016). Embodied cognition processing and representation of power words by second language learners with different proficiency levels. *Chinese Journal of Applied Linguistics*, **39**(4). https://doi.org/10.1515/cjal-2016-0030.

R Core Team (2022) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Schubert, T. W. (2005). Your highness: Vertical positions as perceptual symbols of power. *Journal of Personality and Social Psychology*, **89**(1), 1–21. https://doi.org/10.1037/0022-3514.89.1.1.

Schubert, T. W., **Waldzus, S.**, & **Giessner, S. R.** (2009). Control over the association of power and size, **27**(1), 1–19. https://doi.org/10.1521/soco.2009.27.1.1.

Šetić, M., & **Domijan, D.** (2007). The influence of vertical spatial orientation on property verification. *Psihologija*, **40**(4), 415–428.

Taylor, J. E., **Beith, A.**, & **Sereno, S. C.** (2020). LexOPS: An R package and user interface for the controlled generation of word stimuli. *Behavior Research Methods*, **52**(6), 2372–2382. https://doi.org/10.3758/s13428-020-01389-1.

Thierry, G., & **Yan, J. W.** (2007). Brain potentials reveal unconscious translation during foreign-language comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(30), 12530–12535. https://doi.org/10.1073/pnas.0609927104.

Thomsen, L., **Frankenhuis, W. E.**, **Ingold-Smith, M.**, & **Carey, S.** (2011). Big and mighty: Preverbal infants mentally represent social dominance. *Science*, **331**(6016), 477–480. https://doi.org/10.1126/science.1199198.

Treccani, B., **Umiltà, C.**, **Tagliabue, M.**, & **Bouharab, A.** (2019). Does perceptual simulation explain spatial effects in word categorization? *Cognition*, **186**, 50–64. https://doi.org/10.1016/j.cognition.2019.02.011.

van Heuven, W. J. B., & **Dijkstra, T.** (2010). Language comprehension in the bilingual brain: fMRI and ERP support for psycholinguistic models. *Brain Research Reviews*, **64**(1), 104–122. https://doi.org/10.1016/J.BRAINRESREV.2010.03.002.

Vukovic, N., & **Shtyrov, Y.** (2014). Cortical motor systems are involved in second-language comprehension: Evidence from rapid mu-rhythm desynchronisation. *NeuroImage*, **102**(P2), 695–703. https://doi.org/10.1016/j.neuroimage.2014.08.039.

Wei, Y. F., **Yang, W. W.**, **Oppenheim, G.**, **Hu, J. H.**, & **Thierry, G.** (2024). Embodiment for spatial metaphors of abstract concepts differs across languages in Chinese-English bilinguals. *Language Learning*, **74**(S1), 224–257. https://doi.org/10.1111/lang.12632.

Wu, X., **Jia, H.**, **Wang, E.**, **Du, C.**, **Wu, X.**, & **Dang, C.** (2016). Vertical position of Chinese power words influences power judgments: Evidence from spatial compatibility task and event-related potentials. *International Journal of Psychophysiology*, **102**, 55–61. https://doi.org/10.1016/j.ijpsycho.2016.03.005.

Wu, Y. J., **Cristino, F.**, **Leek, C.**, & **Thierry, G.** (2013). Non-selective lexical access in bilinguals is spontaneous and independent of input monitoring: Evidence from eye tracking. *Cognition*, **129**(2), 418–425. https://doi.org/10.1016/J.COGNITION.2013.08.005.

Yap, A. J., **Mason, M. F.**, & **Ames, D. R.** (2013). The powerful size others down: The link between power and estimates of others' size. *Journal of Experimental Social Psychology*, **49**(3), 591–594. https://doi.org/10.1016/j.jesp.2012.10.003.

Zwaan, R. A., & **Yaxley, R. H.** (2003). Spatial iconicity affects semantic relatedness judgments. *Psychonomic Bulletin & Review*, **10**(4), 954–958. https://doi.org/10.3758/BF03196557.

https://doi.org/10.1017/S1366728925100588 Published online by Cambridge University Press