

Article

Context Matters: How Experimental Language and Language Environment Affect Mental Representations in Multilingualism

Laura Sperl^{1,2}, Marta Sofia Nicanço Tomé^{1,3}, Helene Kühn^{1,4} and Helene Kreysa^{1,*} 

¹ Department of General Psychology and Cognitive Neuroscience, Friedrich Schiller University, D-07743 Jena, Germany; laura.sperl@fernuni-hagen.de (L.S.); marta.tome@web.de (M.S.N.T.); hfm.kuehn@icloud.com (H.K.)

² Department of General Psychology: Judgment, Decision Making, Action, FernUniversität Hagen, D-58097 Hagen, Germany

³ Institute of Psychology, Otto von Guericke University Magdeburg, D-39106 Magdeburg, Germany

⁴ Faculty of Psychology, TUD Dresden University of Technology, D-01069 Dresden, Germany

* Correspondence: helene.kreysa@uni-jena.de

Abstract: The Revised Hierarchical Model (RHM) proposed by Kroll and Stewart has been one of the most influential models of late multilingual language processing. While the model has provided valuable insights into language processing mechanisms, the role of contextual factors for the RHM has not been investigated to date. Such contextual effects could be manifold, including individual speakers' language profiles (such as age of acquisition, proficiency, and immersion experiences), experimental factors (such as different instruction languages), and environmental factors (such as societal language(s)). Additionally, it also appears promising to investigate the applicability of the RHM to non-native multilingual speakers from diverse backgrounds. To investigate whether some of the mentioned contextual factors affect non-native language processing, we designed three online experiments requiring answers in German and English, but tested speakers whose first language was *neither* German *nor* English. They performed a series of translation, picture-naming, and recall tasks based on Kroll and Stewart, as well as providing detailed information on their proficiencies, profiles of language use, and exposure. Experiment 1, conducted with speakers living in Germany, established the paradigm and investigated the role of *individual differences* in linguistic background. While Experiment 2 focused on the short-term effects of the *experimental context* by varying whether instructions were provided in German or in English, Experiment 3 examined the longer-term role of the current *language environment* by comparing individuals living in German-speaking countries with speakers living in societies where neither experimental language is spoken regularly. As in Kroll and Stewart, both the response language and the list type constituted key variables affecting response times and accuracy, known as language asymmetry and category interference. Importantly, the strength of this asymmetry was affected by participants' immersion experiences, suggesting a certain dynamic development in multilingual language processing. In addition, context also seemed to play a role for experimental performance, especially the language environment examined in Experiment 3. Hence, speakers' individual linguistic backgrounds and experience with the experimental languages, as well as additional contextual factors, need to be considered when conducting multilingual experiments and drawing conclusions about multilinguistic processing.

Keywords: multilingualism; foreign language processing; contextual factors; non-native speakers; psycholinguistics; language context; experimental instructions; environmental language



Citation: Sperl, Laura, Marta Sofia Nicanço Tomé, Helene Kühn, and Helene Kreysa. 2024. Context Matters: How Experimental Language and Language Environment Affect Mental Representations in Multilingualism. *Languages* 9: 106. <https://doi.org/10.3390/languages9030106>

Academic Editors: Margreet Vogelzang, Jacopo Torregrossa and Mandy Wigdorowitz

Received: 15 November 2023

Revised: 29 February 2024

Accepted: 6 March 2024

Published: 19 March 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

More than half of the world's population is multilingual (Crystal 2007; Grosjean 2021), meaning that people speak and understand more than one language (Aronin and Singleton 2012; Cenoz 2013; Grosjean 2021; Kemp 2009). Within the large body of research on multilingual language processing (e.g., Aronin and Singleton 2012; Kroll and Stewart

1994; Lin and Lei 2020; Pavlenko 2011), a particularly influential theoretical framework is the Revised Hierarchical Model (RHM) proposed by Kroll and Stewart (1994). This model describes the mental processing of several languages by late multilinguals, i.e., individuals who learnt a foreign language after early childhood (Kroll et al. 2010). It suggests that two lexical items which carry the same meaning in different languages, for example *apple* in English and *Apfel* in German, are stored cognitively in distinct lexical stores (French and Jacquet 2004; Heredia 1997; Kroll et al. 2010), while the shared meaning of both words resides within a common, language-independent conceptual store (Gürel 2004; Kroll and Tokowicz 2001). Importantly, bidirectional connections exist between the separate lexical stores of the native (L1) and subsequently acquired foreign language (L2) (Broersma and Bot 2006; Brown and Gullberg 2008; Degani et al. 2011; Kroll and Stewart 1994; Lagrou et al. 2011). However, the connections differ in strength and are therefore *asymmetrical*: translation from L2 to L1 tends to be faster and more accurate compared to that in the reverse direction (Kroll et al. 2002; Kroll and Sholl 1992; Kroll and Stewart 1994; Palmer et al. 2010; Potter et al. 1984; Sperl et al. 2023; Zheng et al. 2020). The model explains this phenomenon by suggesting that the preferred pathway from the L1 word (e.g., *apple*) to the corresponding word in L2 (*Apfel*) is mediated via its equivalent representation in the conceptual store. In contrast, as depicted in Figure 1, translation from L2 to L1 is possible directly, without requiring the conceptual store.

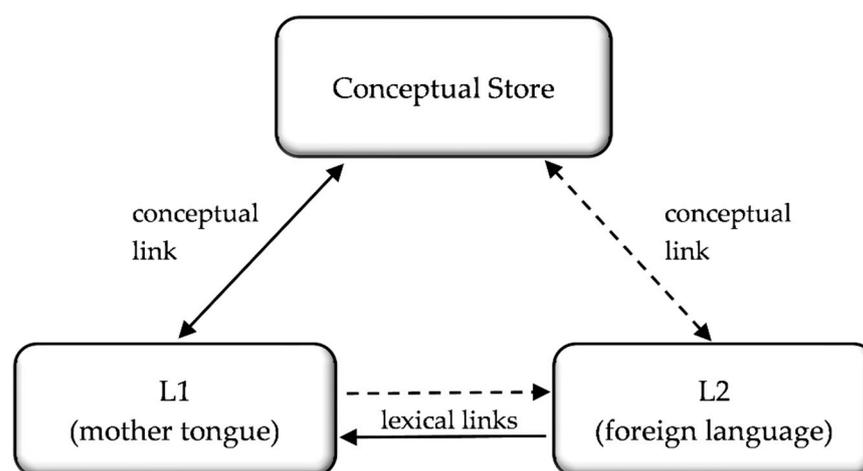


Figure 1. The Revised Hierarchical Model adapted from Kroll and Stewart (1994, p. 158). Continuous lines symbolize strong links, while dashed lines symbolize less intense connections.

In addition to proposing bidirectional asymmetrical routes for translation, Kroll and Stewart (1994) discussed the phenomenon of *category interference*, which means that participants' performance depends on whether stimuli are presented in semantically structured lists (e.g., containing only birds) or mixed lists (e.g., a list containing a mixture of semantic categories: birds, vehicles, furniture, fruit, etc.). Interestingly, participants were slower to respond when translating from categorically organized lists than from mixed lists, but only when the answer was required in L2 (i.e., in the direction where the cognitive route was assumed to be conceptually mediated). This is attributed to the activation of an entire category, resulting in an increased number of competing lexical entries (e.g., *dove* vs. *sparrow*). Consequently, the complexity of the language processing task is heightened, requiring participants to select between similar entries. In turn, this means that they respond more slowly and make more errors than with mixed lists. At the same time, this deeper processing yields a benefit in terms of improved subsequent recall (Kroll and Stewart 1994). In sum, the RHM is based on three key findings, namely, (1) an *asymmetry effect*, whereby translation into L1 is faster and more accurate than into L2; (2) *category interference*, which results in slower response times and reduced accuracy for semantically organized lists compared to mixed lists; and (3) *category facilitation*, with enhanced recall of words from

organized lists compared to mixed lists. These findings shed light on the dynamics of language processing, where the activation of category information can both impede and enhance cognitive performance, offering valuable insights into the mechanisms underlying multilingual cognition.

In the 30 years since its publication, the RHM has been intensively questioned and refined by numerous studies (Heredia 1997; Kroll and Tokowicz 2001; Kroll et al. 2010; Palmer et al. 2010). Recently, Sperl et al. (2023) investigated the applicability of the RHM to non-native multilingual speakers. Their Experiment 2 resembled Kroll and Stewart's (1994) Experiment 3, but instead of comparing participants' native language to a second language, they compared asymmetry and category effects for two different foreign languages (e.g., L2 and L3). Surprisingly, an asymmetry resembling the one seen between L1 and L2 was also present between L2 and L3, instead of being less pronounced or even absent for non-native languages. This unexpected finding prompts a reconsideration of the RHM as proposed by Kroll and Stewart (1994). Originally developed to explain the relationship between L1 and L2 regardless of their identity, the model appears to apply more broadly to relationships among multiple non-native languages. It may be that any of an individual's foreign languages can, under certain conditions, assume the role typically played by the L1. This 'pseudo-native' language could potentially manifest similar effects to a true L1 in its interactions with other languages.

However, understanding such differences between foreign languages requires us to identify which factors affect the relative status of two or more foreign languages. Proficiency and age of acquisition are often reported in this context. However, a recent study by Sperl et al. (2023) showed that a language learned later in life can still be superior in status compared to one learned earlier (in their case German, which was on average learned at an age of 16.5 years, compared to English, with 8.2 years). In addition, participants in this study exhibited a highly similar proficiency level regarding the two experimental languages ($M = 4.5$ on a scale from 1—A2 to 6—C2 level). Nonetheless, asymmetries were clearly evident between the two languages. One plausible factor that could also play a role in determining the relative status of an individual's foreign languages could be the *language context* surrounding the multilingual speaker.

On the one hand, this could be the immediate *experimental context*, i.e., the language of the current investigation (Grosjean 1989; Wu and Thierry 2010). For instance, recruiting for, introducing, and conducting an experiment in German could potentially make German particularly salient (Elston-Güttler et al. 2005; Wu and Thierry 2010) and facilitate its processing, leading to enhanced asymmetry and category effects. Essentially, a German experimental language context could strengthen the activation of German due to its immediate contextual presence, thereby activating its vocabulary. On the other hand, the relative status of two foreign languages might also be influenced by the surrounding language context in everyday life as a more long-term factor reflecting the *language environment* of the individual. This type of context would be expected to shape a person's language habits and influence the frequency of both active language use and passive language exposure (see also Wigdorowitz et al. 2022, 2023). Recent research has indeed demonstrated an influence of the linguistic context on both lexical access in different languages and general cognitive control abilities (Beatty-Martínez et al. 2020).

Here, a person's language competence may also be profoundly affected by *immersion experiences* i.e., the intense act of learning a language implicitly by immersing oneself into the respective culture and context (Dewey 2007; Juan-Garau and Lyster 2018; Wilkinson 1998). Studies have found that acquiring a language implicitly, which occurs by being surrounded constantly by the language in the respective country, may lead to a higher and more efficient, almost mother-tongue-like, language gain (Freed et al. 2004; Morgan-Short et al. 2012). This method of acquisition is similar to learning one's native language, taking place naturalistic and "accidentally", without actively studying grammar or vocabulary (Juan-Garau and Lyster 2018; Kearney 2010). A foreign context facilitates learning opportunities which a classroom is not able to provide. It provides interactions with native

speakers. These improve language skills significantly, especially if conversations go beyond small talk and into a deeper level (Baker-Smemoe et al. 2014). A study by Morgan-Short et al. (2012) even found native-like brain activation patterns in participants who learned the foreign language implicitly, while no differences were found in participants who were taught explicitly (e.g., using a grammar-focused classroom method). This finding supports the idea that the conditions of learning a language play a critical role and that certain circumstances, e.g., an immersion experience compared to studying purely in a classroom, may lead to the acquisition of a ‘pseudo-native’ language.

To summarize, both short-term exposure to an *experimental language context* or long-term exposure to a certain *environmental language context* (potentially associated with a certain degree of *immersion experience*) could activate the corresponding vocabulary and render this particular language salient (Montrul 2015; Treffers-Daller 2019), leading to differential asymmetries and category effects between several foreign languages.

The Current Study

In order to identify how short- vs. long-term contextual factors affect the relative status of two (or more) foreign languages, three online experiments were conducted with foreign language speakers. These experiments were loosely based on Kroll and Stewart’s (1994) theoretical framework. Specifically, Experiment 1 followed up on the unexpected findings of Sperl et al. (2023), replicating and establishing the paradigm in an online format and focusing on individual differences between speakers’ language biographies. Next, two further experiments manipulated and compared two distinct types of language context: Experiment 2 investigated the short-term role of the experimental language by varying the language of the instructions, while Experiment 3 targeted the wider long-term language environment by contrasting participants living in countries where different societal languages are spoken. It seems plausible that either or both types of language context could highlight one foreign language over another, giving it a (temporarily) similar status to the L1 in the classic RHM. As in Kroll and Stewart (1994), this should be visible in the form of asymmetry effects, category interference, and category facilitation, based on the two dependent variables response time (RT) and accuracy.

All three experiments were conducted online and in parallel, with data acquisition taking place between July 2021 and December 2022. Personal networks and social media postings were used to recruit an exclusive expert sample of late-multilingual speakers who were proficient in both experimental languages (German and English), although neither language was their L1. This approach allows for a meaningful comparison with the RHM, while also enabling us to identify conditions which lead foreign languages to achieve ‘pseudo-native’ status. It compares two foreign languages (instead of a native and a foreign language) in order to examine the salience of a language under different contextual conditions. The key manipulation consisted of highlighting the German language in different ways, potentially giving it a ‘pseudo-native’ status via either the short-term experimental context (Experiment 2) or the longer-term environmental context (Experiment 3).

2. Experiment 1: Online Replication and Immersion Experiences

Experiment 1 examined the applicability of the RHM (Kroll and Stewart 1994) to subsequent languages (L2, L3 and so forth: note that, in the interest of readability, we will always refer to the two foreign target languages of this study as L2 and L3 for all participants, irrespective of their order of acquisition). This experiment was conducted online due to pandemic restrictions on lab studies. At the same time, it allowed us to test a form of data acquisition that would make it possible to assess a rather exclusive sample of multilingual participants in different locations. In this experiment, participants were late multilinguals currently living in German-speaking countries (i.e., individuals who acquired their L2 only after entering school).

Hypotheses. As already mentioned, we expected *asymmetry effects*, with faster RT and higher accuracy for answers in German than in English. In other words, although both

languages are assumed to be connected bidirectionally (connections from L2 to L3 and vice versa), participants should respond faster and more precisely in their more salient foreign language (here: German), as they would in their L1, compared to their less salient language (here: English). Such a finding would suggest varying strengths of mental pathways between lexical stores. Furthermore, we expected *category interference* effects, where participants would respond slower and less accurately when completing tasks with semantically organized lists rather than mixed lists. However, analogously to Kroll and Stewart (1994), these effects were only anticipated in the less salient language. Therefore, we predicted that participants would experience category interference, particularly in English. Lastly, we predicted *category facilitation*, whereby more items would be correctly recalled from organized lists due to more in-depth processing during longer response times (Kroll and Stewart 1994).

In addition to an online replication of Sperl et al. (2023), Experiment 1 also looked at the role played by *immersion experiences*. We predicted that the language of immersion, in this case German, would be highlighted and therefore assume the role of a pseudo-native language. Moreover, participants with a deeper immersion were supposed to show a stronger asymmetry effect, with faster RT and higher accuracy for answers in German than in English. In other words, we hypothesized that the asymmetry effect would grow stronger with deeper immersion into the German language due to it becoming more salient in comparison to English. In an exploratory manner, Experiment 1 also investigated whether category interference and recall facilitation were associated with the strength of immersion.

2.1. Methods

2.1.1. Participants

Participants were non-German and non-English native speakers who had been living at the time of the experiment in Germany for a minimum of three months and had advanced knowledge of both German and English as foreign languages.

Initially, 60 people participated in the experiment. However, a total of 15 participants had to be excluded due to technical problems ($N = 4$), an accuracy below 50% ($N = 5$), for being early multilinguals ($N = 3$), or because their L1 was German ($N = 3$). The final sample of 45 participants was on average 38 years old ($M = 38.02$ years; $SD = 11.29$, range: 20–59, with 35 females and 10 males). The recruitment strategy consisted of using personal and social media contacts and language teachers. Psychology students were rewarded with course credits.

On average, participants spoke four different languages ($M = 4.11$; $SD = 1.17$; range: 3–7), including German, English, and their native language, which could be Spanish ($N = 15$), Serbian ($N = 5$), Portuguese ($N = 5$), French ($N = 3$), Turkish ($N = 2$), Arabic ($N = 1$), Armenian ($N = 1$), Basque ($N = 1$), Bulgarian ($N = 1$), Catalan ($N = 1$), Chinese ($N = 1$), Croatian ($N = 1$), Georgian ($N = 1$), Greek ($N = 1$), Italian/Albanian ($N = 1$), Kurdish ($N = 1$), Norwegian ($N = 1$), Romanian ($N = 1$), Serbo-Croatian ($N = 1$), or Swedish ($N = 1$). For a full overview of the languages participants reported being able to speak, see Table S1 in the Supplemental Materials.

Contextual usage of languages. On average, participants had been living in Germany for about 12 years, with a wide range between 9 months and 32 years ($SD = 8.71$). Consequently, although German was not their native language, most participants (69%) used German in everyday life; 33% of participants used their native language; and 22% used English (total responses exceeded 100% as they could name more than one option). A similar distribution was found for the language used at work (69% German, 40% native language, and 27% English). Nonetheless, most participants reported usually thinking in their native language (78%, with 40% German and 13% English).

Age of acquisition. On average, participants started learning German during early adulthood at the age of 20 years ($M = 19.84$, $SD = 8.53$; range: 7–39) and English at the age of 11 years ($M = 10.6$, $SD = 4.68$, range: 5–30), making them late multilinguals in both

languages. In total, they had been using German for about 18 years ($M = 18.09$, $SD = 10.78$; range: 2–39) and English for about 27 years ($M = 27.48$, $SD = 10.31$, range: 7–49) at the time of the experiment.

Proficiency. On a scale from 1 to 6, with 1 representing an A1 proficiency level and 6 representing a C2 proficiency level according to the definitions of the [Council of Europe \(2001\)](#); described in detail in Table S2 in the Supplemental Materials), participants indicated an overall proficiency of 4.98 in German ($SD = 0.89$, range: 3–6) and 4.44 in English ($SD = 1.01$, range: 2–6). The average proficiency in German therefore corresponded to C1, which is an advanced level, allowing most participants to understand and use a wide range of complex language and express themselves effectively, even in academic or professional contexts. Average proficiency in English was slightly lower, between B2 and C1, where B2 represents an upper-intermediate level, allowing speakers to express their thoughts spontaneously and without much difficulty ([Council of Europe 2001](#)).

Receptive and productive use. When distinguishing between proficiency skills, on a scale from 1 (“never”) to 7 (“very often”), participants reported using German very often both receptively (reading and listening; $M = 6.00$, $SD = 0.79$, range: 4–7) and productively (writing and speaking; $M = 5.91$, $SD = 0.97$, range: 4–7). English was reported as being used slightly less (receptively: $t(76.88) = 4.27$, $p < .001$; productively: $t(76.77) = 5.45$, $p < .001$), though it was still employed frequently by most participants (receptively: $M = 5.09$, $SD = 1.18$, range: 1–7; productively: $M = 4.49$, $SD = 1.46$, range: 1–7).

2.1.2. Materials

Experiment 1 consisted of two questionnaires (language and immersion), two types of experimental task (translation and picture naming), and a recall task. The questionnaires and task instructions were all presented in written form in German.

Language questionnaire. Aside from assessing demographic data (age, gender, country of residence, time living in Germany, native language, and occupation), this questionnaire assessed participants’ language experience regarding German and English, as described in Table S3 in the Supplemental Materials. It comprised the age of acquisition for German and English, as well as self-assessments of proficiency in German and English according to the criteria described in the Common European Framework of Reference for Languages ([Council of Europe 2001](#)), an official framework for describing language abilities (for further details, see Table S2 in the Supplemental Materials). This scale categorizes language abilities as A, B, and C, indicating elementary, intermediate, and proficient language use, respectively. Each category is further subdivided into specific levels: A1, A2, B1, B2, C1, and C2. These correspond to different proficiency levels, ranging from a beginner’s proficiency to approximately that of a native speaker. While self-assessments are necessarily subjective, the scale and the corresponding descriptions provide a standardized reference for evaluating productive and receptive language skills across different languages and contexts. Further questions assessed the native language and other languages spoken by the participant. Also, participants were asked about how frequently they used German, English, and their native language during their occupation, in everyday life, and in their thoughts. Finally, they reported how often they used these three languages for reading and listening (comprehension) vs. writing and speaking (production).

Immersion questionnaire. The degree of immersion was assessed with a specifically constructed immersion questionnaire. This questionnaire was based on the validated Sojourner Adjustment Measure (SAM) designed by [Pedersen et al. \(2011\)](#), which measures adjustment in people living abroad temporarily. Hence, the immersion questionnaire assessed immersion regarding German, the environmental language context participants were currently living in. The original version of the SAM identified six factors, which included items about language and culture, from which we extracted three language-specific aspects of immersion relevant for this study. These were (1) *social interaction with host nationals*, (2) *language development and use*, and (3) *social interaction with co-nationals*. The respective items were adapted and translated by the authors into German. An additional

fourth aspect, *subjective integration*, was added based on literature research (Dewey 2007; Kinginger 2008; Knight and Schmidt-Rinehart 2002) and included questions about the motivation and interest participants felt towards integrating themselves in Germany. The resulting immersion questionnaire, available in Table S4 in the Supplemental Materials, was presented to participants after completing the experimental tasks and consisted of 24 items with 7-point scales ranging from 1 = *strongly disagree* to 7 = *strongly agree*, or from 1 = *never* to 7 = *always*.

Translation and picture-naming task. The experimental lists and items used for the translation and picture-naming tasks were identical to those used by Sperl et al. (2023). These could be words in German or English or pictures of various objects, such as buildings and vehicles, such as those listed in Table S5 in the Supplemental Materials. The items were deemed to be easy or of intermediate difficulty to answer and avoided cognates that look and sound similar in the two languages (e.g., *Tomate* and *tomato*). In total, 100 stimuli were organized into 10 lists—each containing either mixed, semantically organized (according to the categories *vehicles, buildings, fruits, vegetables, furniture, kitchenware, stationery, and clothing*), or abstract words (e.g., *sadness*). Abstract lists (used for translation tasks only) were included in accordance with the work of Sperl et al. (2023), who were interested in whether the effects of the RHM might also hold true for words that lack a clear pictorial representation in the conceptual store. However, abstract lists are not the primary focus of the current study, and their inclusion is solely to ensure methodological consistency within the established design. Most of the pictures used in the semantically organized and mixed picture-naming lists were originally from the BOSS inventory (Brodeur et al. 2010), supplemented by additional internet searches. To minimize the influence of potential differences in stimulus difficulty, two alternate versions of the experiment were created. Each mixed list of stimuli consisted of words drawn from the semantically organized and abstract word lists in the parallel version. The experiment was implemented in E-Prime 3.0 Go (Psychology Software Tools Inc. 2020).

2.1.3. Design

The experimental part of the study consisted of two types of task. First, participants translated words from English into German, and then vice versa (translation tasks). In the second section, they named a depicted object out loud, first in English and then in German (picture-naming tasks). As depicted in Figure 2, items of the translation tasks were embedded in three types of word list: mixed, semantically organized, and abstract. Since abstract words cannot be depicted well, the picture-naming tasks only consisted of mixed and semantically organized lists. The order of the lists within each task was randomized. However, the order of the items within each list remained the same for all participants, following the precedent of Kroll and Stewart (1994). Experiment 1 used a within-participant design, comparing the performance (response times and accuracy) for each participant in relation to response language, list, and task type.

TASK TYPE:	1. Translation Tasks		2. Picture Naming Tasks	
RESPONSE LANGUAGE:	a. Answers in German	b. Answers in English	a. in German	b. in English
LIST TYPE:	Mixed Lists	Mixed Lists	Mixed Lists	Mixed Lists
	Organized Lists	Organized Lists	Organized Lists	Organized Lists
	Abstract Lists	Abstract Lists		

Figure 2. Overview of the experimental tasks and list types. Experimental tasks and response languages were always presented in this order. The order of list types within each task and language was randomized for every participant.

2.1.4. Procedure

Participants received a link to download an E-Prime Go file (Psychology Software Tools Inc. 2020), which enabled the experiment to run on their personal computers without further supervision or interaction. It began with the language questionnaire (cf. Table S3 in the Supplemental Materials). Next, participants performed the two translation tasks, followed by the two picture-naming tasks. Each task began with two trial items, followed by three word lists for each translation task and two picture lists for each picture-naming task. Each list consisted of 10 stimuli, meaning that in total participants translated and named 100 items. First, they translated 30 words from three lists into German and then 30 words into English; then, they named 20 pictures from two lists in German and then 20 pictures in English. Participants could take a break after each list of 10 items. Each item began with a black fixation cross, presented for 500 ms in the middle of the screen. Next, the stimulus was presented for 3500 ms, and participants responded by speaking into their headset or internal microphone as fast as possible. The recording of each answer began with the presentation of the stimulus and was saved as an individual .wav file.

After the experimental tasks, participants completed the immersion questionnaire (Table S4 in the Supplemental Materials). Finally, in the recall task, they were asked to type out as many stimuli as they could remember. They were instructed to write the words in the language in which they remembered the items first. There was no time limit for the recall task. The entire study took approximately 20 min in total, depending on how quickly the participants answered the questionnaires. All data were saved automatically to the E-Prime Go platform (Psychology Software Tools Inc. 2020).

2.1.5. Data Preparation

After downloading the data of each participant, all relevant variables from the questionnaires were collated into a single table, while the .wav files were processed using Praat (Boersma and Weenink 2016) to assess response times and accuracy.

Language questionnaire data. Proficiency levels in German and English were coded on a 6-point scale according to the Common European Framework of Reference for Languages (Council of Europe 2001), with higher values meaning higher proficiency. The scale ranged from A1 (elementary) to C2 (proficient). The productive and receptive use of German and English were assessed on a 7-point scale from *never* to *always*, with higher values indicating more frequent use. The use of German and English in the contexts of occupation, daily life, and thought processes was coded binarily: 1 if the respective language was used in the particular context, and 0 otherwise. Participants' age of acquisition of German and English were assessed in years, while duration of residence was assessed in months.

Response times and accuracy. A Praat script initially detected the response latency of each trial. These speech onsets were manually inspected and adjusted, e.g., when noise or filler words were captured before the answer and not detected automatically. For accuracy,

correct answers were coded with a 1 and incorrect answers were coded as 0. Synonyms and grammatical variations of the correct answers were accepted (e.g., “sad” instead of “sadness”). Only correct answers were considered for RT. Data from participants with total accuracy < 50% were not included in the analyses as, with only 10 items per list, this did not leave sufficient items to reliably determine RT for all conditions.

Recall performance. Each correct response in the recall task was classified based on list type (mixed, organized, or abstract), task type (translation or picture-naming task), and response language (translating into/naming in German vs. English) which it had originally occurred in. Statistical analyses considered the proportion of words in each category (e.g., words that appeared in a translation task) compared to the total amount of words recalled per participant. Practice items and words that had not occurred in the experiment at all were not accepted. In total, participants could have recalled a maximum of 100 items.

2.1.6. Statistical Analysis

All statistical analyses were performed in R version 4.2.1 (2022-06-23 ucrt) with a significance level of $\alpha = .05$. In the interest of readability, we will generally only mention results when they are statistically significant or directly relevant to our hypotheses.

Asymmetry and category effects. We used the *lmer* package for R (Bates et al. 2015) to run separate mixed-effect models (Baayen et al. 2008; Brown 2021) for RT using linear mixed-effect models (LME; Winter 2020), while generalized mixed-effect models (GLMEs; Whalley 2019) were used to analyze the same effects for the binary outcome of accuracy. These analyses were performed separately for translation and picture-naming tasks since the two tasks differed in complexity. In translation tasks, list type contained three levels of mixed, organized, and abstract word lists, while picture-naming tasks only used mixed and organized lists.

Variables were contrast-coded in all mixed-effects models, as recommended by various researchers (e.g., Brehm and Alday (2022)). This procedure was implemented in the present study to elucidate main effects. As such, the dependent variable (RT or accuracy) was modeled with a response language (translating into/naming in English or German) that had been sum-coded (Brehm and Alday 2022) using a contrast coding of (−0.5, 0.5), with German as the first and English as the secondary category. List type was also included in the model; however, in a dummy-coded contrast for translation tasks, lists were arranged into mixed, organized, and finally, abstract lists (0, 1, 0 for organized and 0, 0, 1 for abstract lists), so that mixed lists formed a baseline for the other two list types. For picture-naming tasks, a sum-coded contrast of (0.5, −0.5) was used for list type, with mixed lists serving as the first and organized lists as the secondary category. Both independent variables (response language and list type) were defined as fixed effects with by-participant and by-item random intercepts, allowing for interactions; the response language was additionally specified with a by-participant-varying slope.¹ Maximum likelihood (Laplace approximation for GLMEs) was used to estimate parameters. In case of convergence issues with the initial model, an iterative selection process using the *allFit* function from the *lme4* R package was used to determine the appropriate optimizers (e.g., *bobyqa*) to facilitate model fit.

Post hoc tests were conducted using estimated marginal means with the R package *emmeans* (Lenth 2023). These post hoc comparisons allowed us to further investigate the main effects of response language and list type, as well as their interaction, while considering the dependencies among the different levels of the grouping variables. In addition to pairwise comparisons between conditions, we specifically tested our key hypothesis about category interference in a combined comparison: responding to semantically organized lists in English should result in lower performance than mixed lists in English as well as organized and mixed lists in German.

Recall. Finally, using the *ez* package for R, recall performance was examined with two-way within-subject repeated-measures analyses of variance (ANOVA) to analyze the effect of list type and response language. Thus, a 3 (list type: mixed vs. organized vs.

abstract word lists) \times 2 (response language: translating into German vs. into English) ANOVA was conducted for recall performance from translation tasks, while a 2 (list type: mixed vs. organized lists) \times 2 (response language: German vs. English) ANOVA was calculated for picture-naming tasks. In cases of sphericity violations, Greenhouse–Geisser-corrected values are reported. Post hoc analyses used Bonferroni-corrected p -values for multiple comparisons.

Immersion questionnaire data. Data from the immersion questionnaire were extracted as one total immersion score (average value over all questions) and we averaged the scores of the single facets. The total score was used to measure the degree of immersion, i.e., how deeply a person reported being immersed into the German language. Sub-scores were used to understand which specific aspects of immersion contributed most to processing a foreign language similarly to one's native language. Correlations with language and residence variables were also investigated.

2.2. Results

2.2.1. Language Characteristics

The number of languages spoken by the participants correlated significantly with their proficiency in German ($r = .42$; $p = .005$) and with using German productively ($r = .35$; $p = .02$), suggesting that participants who spoke many languages tended to consider themselves more fluent in German, but not necessarily in English. The age of acquisition (AoA) for German correlated negatively with various variables regarding the German language, such as proficiency ($r = -.61$; $p < .001$), productive use ($r = -.32$; $p = .03$), receptive use ($r = -.34$; $p = .03$), thinking in German ($r = -.33$; $p = .02$), and using German at work ($r = -.52$; $p < .001$). These links indicate that those who learned German earlier (lower AoA) indicated higher proficiency and the use of the language in various contexts. This was not found for AoA in English, which correlated only with using English at work ($r = -.36$; $p = .01$). A full overview of correlations between language-use variables with corresponding descriptive statistics is presented in Table S6 of the Supplemental Materials.

2.2.2. Translation Tasks

Observing translation tasks overall, participants translated 72% of the presented words correctly ($SD = 0.10$) and responded on average after 1585 ms ($SD = 193$). Mean performance is depicted in Figure 3, and further specified in Table 1, depending on task type, response language, and list type.

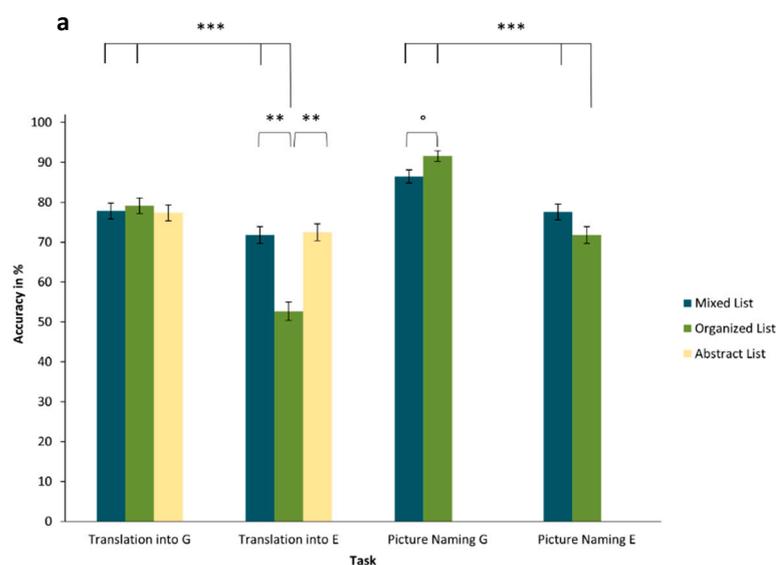


Figure 3. Cont.

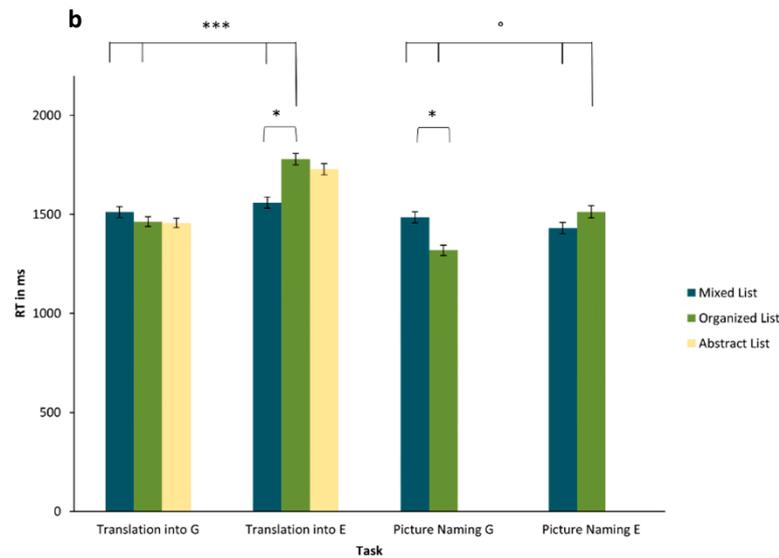


Figure 3. Mean (a) accuracy and (b) response times depending on task type, response language, and list type. The large brackets on top of each subfigure represent the combined comparison of semantically organized lists in English versus mixed lists in English and organized and mixed lists in German (see also Statistical Analysis described in Experiment 1). Error bars depict standard errors. ° $p < .1$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 1. Grand means and SD of the dependent variables (DV) accuracy and response times (RT), listed separately for each task type, response language, and list type.

DV	Task Type	Response Language	List Type							
			Mixed		Organized		Abstract		Overall	
			M	SD	M	SD	M	SD	M	SD
Accuracy	Translation	German	0.78	0.16	0.79	0.16	0.77	0.17	0.78	0.1
		English	0.72	0.14	0.53	0.21	0.72	0.17	0.66	0.15
		Overall A	0.75	0.12	0.66	0.14	0.75	0.13	0.72	0.10
	Picture naming	German	0.86	0.14	0.92	0.1			0.89	0.11
		English	0.78	0.18	0.72	0.21			0.75	0.17
		Overall RT	0.82	0.11	0.82	0.11			0.82	0.09
RT	Translation	German	1546	291	1475	249	1478	257	1495	209
		English	1579	265	1803	340	1746	325	1704	245
		Overall A	1552	217	1602	236	1608	234	1585	193
	Picture naming	German	1505	311	1326	288			1414	278
		English	1450	314	1566	357			1507	294
		Overall RT	1471	258	1414	233			1445	224

Note. Values are not provided for abstract lists in the context of picture-naming tasks, as this particular task does not incorporate such list types. All accuracies ranged from 0 to 1, as correct responses were coded with 1 and wrong answers with 0. RT was measured in milliseconds. $N = 45$.

Accuracy. The mean accuracy for translation into English was somewhat lower ($M = 0.66$, $SD = 0.48$) than for translation into German ($M = 0.78$, $SD = 0.41$). According to the results of the GLME model provided in Table 2, the odds of accurate answers were higher than 50% (significant intercept), but the effect of the response language was not significant. There was also no significant effect of list type, as the estimates of neither organized nor abstract lists were significant. However, the effect of response language differed for organized lists, with greater accuracy seen when responding in German (emmean = 1.93, $SE = 0.33$) than in English (emmean = 0.15, $SE = 0.32$; Table S7 of the Supplemental Mate-

rials presents the estimated marginal means for all six possible combinations of response language, and list type, provided separately for accuracy and response times).

Table 2. Translation tasks: modeling results for accuracy and RT depending on response language and list type.

Fixed Effects	Accuracy			Response Times		
	β	SE(β)	z-Value	β	SE(β)	t-Value
Intercept	1.45	0.24	6.19 ***	1622	53	30.39 ***
Response language	−0.32	0.45	−0.71	62	95	0.66
Organized lists	−0.42	0.31	−1.35	65	65	1.00
Abstract lists	−0.01	0.31	−0.04	28	64	0.44
Response language × organized lists	−1.46	0.61	−2.37 *	277	130	2.13 *
Response language × abstract lists	0.13	0.61	0.20	231	128	1.81 °

Note. Columns β and SE(β) represent estimates and standard errors of the coefficients, respectively. The model intercept (mean of mixed lists for list type and the grand mean for response language) was fixed as the reference level. ° $p < .1$. * $p < .05$. *** $p < .001$.

Finally, post hoc contrast tests (shown in Table 3) were conducted to test the key hypotheses that semantically organized lists would result in lower accuracy when translating into English compared to translating mixed lists, and even more specifically, that translating semantically organized lists into English would prove harder than mixed lists into English, mixed lists into German and organized lists into German. Both of these contrasts were significant, as can be seen in the first and bottom lines of Table 3. In fact, the highest mean probability of success was for translating organized lists into German (emmean = 1.93, SE = 0.33), the lowest was for translating organized lists into English (emmean = 0.15, SE = 0.32).

Table 3. Translation tasks: post hoc tests comparing response language and list type.

Contrasting	Accuracy					Response Times				
	Estimate	SE	df	z-Ratio	p-Value	Estimate	SE	df	t-Ratio	p-Value
Mixed vs. organized lists into English	−1.14	0.43	Inf	−2.68	.01 **	204	96	132	2.13	.04 *
Mixed vs. abstract lists into English	0.05	0.43	Inf	0.12	.91	144	93	122	1.54	.13
Organized vs. abstract lists into English	1.19	0.43	Inf	2.80	.01 **	−60	96	132	−0.63	.53
Mixed vs. organized lists into German	0.31	0.44	Inf	0.71	.48	−73	93	120	−0.79	.43
Mixed vs. abstract lists into German	−0.07	0.43	Inf	−0.17	.87	−88	93	119	−0.95	.35
Organized vs. abstract lists into German	−0.39	0.44	Inf	−0.88	.38	−14	93	119	−0.15	.88
Organized lists into English vs. mixed lists into English, and mixed and organized lists into German	1.46	0.35	Inf	4.12	.001 ***	−270	81	148	−3.32	.001 ***

Note. Accuracy values are provided on the logit scale; no degrees of freedom are available for generalized linear mixed models. * $p < .05$. ** $p < .01$. *** $p < .001$.

Response time. Correspondingly, translations into English ($M = 1681$ ms, $SD = 617$ ms) were somewhat slower than translations into German ($M = 1477$ ms, $SD = 543$ ms). The outcomes of the LME analysis in Table 2 provide greater detail regarding the interplay of response language and list type. Answering in German or English did not have a significant effect, nor did semantically organized or abstract lists differ from mixed lists per se. Echoing the pattern found for accuracies, the post hoc comparisons presented in Table 3

show that participants translated organized lists into English slower than mixed lists, as well as translating organized lists into English slower than both mixed lists into English and mixed or organized lists into German.

2.2.3. Picture-Naming Tasks

On average, participants correctly named 82% of pictures overall ($SD = 0.09$) and were generally somewhat faster to respond in picture-naming ($M = 1445$ ms, $SD = 224$) than in translation tasks ($M = 1585$ ms; $SD = 193$ ms; $t(3126) = 6.72, p < 0.001$). Figure 3 depicts the mean response accuracy and RT based on task type, response language, and list type; these are further described in Table 1.

Accuracy. According to the GLME in Table 4, neither list type nor the interaction between response language and list type were significant. However, response language had a significant effect on its own, with more accurate responses in German than in English. As shown in Table 5, post hoc comparisons revealed that participants were less accurate when naming semantically organized lists in English (emmean = 1.29, $SE = 0.29$) compared to mixed lists in both languages and organized lists in German (full list of emmeans provided in Table S7 in the Supplemental Materials).

Table 4. Picture-naming tasks: modeling results for accuracy and RT depending on response language and list type.

Fixed Effects	Accuracy			Response Times		
	β	$SE(\beta)$	z-Value	β	$SE(\beta)$	t-Value
Intercept	2.12	0.18	11.71 ***	1478	42	35.22 ***
Response language	-1.31	0.37	-3.53 ***	94	70	1.34
List type	-0.18	0.27	-0.66	29	56	0.52
Response language \times list type	1.03	0.53	1.93 °	-278	112	-2.49 *

Note. β and $SE(\beta)$ represent estimates and standard errors of the coefficients, respectively. The first fixed-effect level (semantically organized lists in the target language German) was fixed as the reference level. ° $p < .1$. * $p < .05$. *** $p < .001$.

Table 5. Picture-naming tasks: post hoc tests contrasting response language and list type.

Contrasting	Accuracy					Response Times				
	Estimate	SE	df	z-Ratio	p-Value	Estimate	SE	df	t-Ratio	p-Value
Mixed vs. organized lists in English	-0.34	0.36	Inf	-0.95	0.34	110	81	84	1.35	0.18
Mixed vs. organized lists in German	0.69	0.40	Inf	1.73	0.08 °	-168	80	78	-2.11	0.04 *
Organized lists in English vs. mixed lists in English, and mixed and organized lists in German	-1.10	0.34	Inf	-3.23	0.001 ***	136	72	105	1.88	0.06 °

Note. Accuracy values are provided on the logit scale; no degrees of freedom are available for generalized linear mixed models. ° $p < .1$. * $p < .05$. *** $p < .001$.

Response times. The results from the LME analysis, presented in Table 4, show no significant main effects for the response language and list type, but a significant interaction between the two variables. Overall, participants were slowest to name words from semantically organized lists in English (emmean = 1580, $SE = 70$), but fastest for organized lists in German (emmean = 1347, $SE = 68$). This result is supported by the post hoc test results shown in Table 5, where the difference observed between naming mixed and organized lists in German is significant, with faster RT for mixed than for semantically organized lists (emmean = 1515, $SE = 68$). The main hypothesis, that naming organized lists in English

would differ from naming any other list types in English or in German, was only marginally significant ($p = .06$).

2.2.4. Recall Performance

After the experiment, participants correctly recalled on average 20 words or items from any of the experimental tasks ($M = 19.98$; $SD = 11.73$; range 0–56), with more stimuli recalled from picture-naming ($M = 12.09$; $SD = 6.75$, range = 0–27) than from translation tasks ($M = 7.84$; $SD = 6.09$, range = 0–29). Cases of falsely recalled items were practically non-existent.

Regarding translation tasks, the two-way ANOVA, analyzing the effect of list type (mixed vs. organized vs. abstract lists) and response language (during the translation task: German vs. English) on the mean number of words recalled, revealed a significant main effect for list type ($F [1.95, 85.72] = 4.89$, $p = .01$, Greenhouse–Geisser-corrected). However, subsequent pairwise comparisons using Bonferroni correction yielded only a marginally significant difference ($p = .05$) between mixed ($M = 1.51$, $SD = 1.29$, range = 0–5.5) and abstract ($M = 0.97$, $SD = 1.16$, range = 0–4) word lists. There was no significant main effect for the response language and no interaction between the two variables.

For picture-naming tasks on the other hand, the two-way ANOVA showed a main effect for the response language ($F [1, 44] = 34.05$, $p < .001$), indicating that participants remembered more items from picture lists they had responded to in English ($M = 3.66$, $SD = 1.92$, range = 0–7.5) than in German ($M = 2.39$, $SD = 1.75$, range = 0–6). There was no main effect for list type and no interaction.

2.2.5. Immersion

On a scale from one to seven, participants reported an immersion score of 5.16 on average. This score correlated negatively with overall RT ($r = -.31$; $p = .03$) and positively with overall accuracy ($r = .38$; $p = .009$). The correlations between total immersion score and asymmetries did not reach significance. An overview of the correlations between asymmetries, interference, and immersion sub-scores can be found in Table S8 of the Supplemental Materials.

Accuracy. Two of the four sub-scores of immersion, *language development and use* and *subjective integration*, correlated positively with the asymmetries in translation ($r = .31$; $p = .037$; $r = .38$; $p = .01$, respectively) and picture-naming accuracy ($r = .29$; $p = .049$; $r = .38$; $p = .009$). In all cases, the more participants reported being immersed into the German language and integrated within German culture, the more pronounced their asymmetry between German and English became.

Response times. For RT, only the sub-score *language development and use* was correlated positively with picture-naming asymmetry ($r = .44$; $p = .002$). Again, deeper immersion led to stronger asymmetry effects.

Category interference and recall. Correlations between immersion scores and both category interference and recall performance did not reach significance in any category.

2.3. Discussion

Experiment 1 explored asymmetry effects and category interference in translation and picture-naming tasks. Like Sperl et al. (2023), using an online version of the same paradigm, we found an asymmetry effect for accuracies in both translation and picture naming, with more accurate responses produced in German compared to English. Moreover, translations of organized lists into English were produced more slowly and resulted in more errors compared to translations of organized lists into German and mixed lists into both response languages (category interference), even though participants were proficient but non-native speakers of both languages. Of note, even in online experiments where each participant carried out their assessment on their personal computer outside of our experimental control, and with a very heterogeneous sample of non-native speakers, clear differences in response times were measured with millisecond resolution. The only part of the study where the

online modality seemed to be problematic was the final recall task, where participants seemed to show less motivation to rack their brains for further items before closing their browser window than when sitting in the lab with paper and a pencil (i.e., $M = 20$ items recalled in the online version vs. $M = 36$ ($SD = 10.35$) in [Sperl et al. 2023](#)).

While these findings are important from a practical point of view regarding the feasibility of conducting language production experiments remotely, the finding that experimental performance depended on *immersion experience* is of theoretical importance. Specifically, the degree of immersion affected accuracy and RT. Asymmetries between German and English responses were more pronounced if a person reported a stronger immersion into the German language and culture, and they were also in general more pronounced for picture naming. This result goes hand in hand with the suggestion that people who are deeply immersed into a language process this language similarly to a native speaker. In fact, the results resemble the pattern found in previous studies with native speakers (e.g., [Sperl et al. 2023](#)), although the non-native participants show longer response times overall.

Arguably, this finding could suggest that immersion might constitute a driving force for the dynamic aspect postulated by [Kroll and Stewart \(1994\)](#). However, in their model, the asymmetry becomes smaller with growing language skills, while the asymmetry in our studies increases for highly immersed speakers of two foreign languages. The rationale behind this important difference is apparent. In the original design, a comparison was made between an L1 and an L2, where strong immersion experiences rendered the L2 more native-like, reducing the difference between the two languages. However, in our study, we investigated *two foreign* languages, where intense immersion experiences accentuated the contrast between the target languages by arguably making one of the two relatively more native-like ('pseudo-native' status). It also seems reasonable to assume that people who report stronger immersion in German consequently use less English than people reporting weaker immersion. In fact, participants who reported higher English than German language skills showed the opposite asymmetry effect, which is to say negative, instead of positive, correlations with asymmetries (e.g., picture-naming accuracy asymmetry: $r = -.52$, $p < .001$). However, this trend was not looked at in detail since only immersion into the German language was measured.

In sum, the results suggest that language context, specifically immersion experience, plays an important role in non-native language processing. Hence, it seems that not only the language level, as originally proposed by [Kroll and Stewart \(1994\)](#), but also the method of learning influences the performance in language tasks.

2.4. Hypotheses Relating to Language Context (Experiments 2 and 3)

While this suggestion is based purely on correlations with the depth of immersion in Experiment 1, Experiment 2 deliberately manipulated language context, focusing on the experimental language context, and randomly assigning participants to receive either German or English instructions. The complimentary experiment, Experiment 3, varied the environmental language context by assessing individuals living in various non-German and non-English environments compared to those in German-speaking countries. In both cases, if short-term or long-term language contexts influence the relative status of foreign languages, translating into English should take more time and result in more errors compared to translating into German in samples with a German context (German instructions in Experiment 2 and living in German-speaking countries in Experiment 3). In contrast, the group in the non-German context (English instructions in Experiment 2 and living in non-German-speaking countries in Experiment 3) was expected to demonstrate a reverse pattern, where translating into German would take longer and be more prone to errors than translating into English. In other words, we expected to observe an *asymmetry effect*, as in Experiment 1, and that this effect would be accentuated by the language context. This was also expected for picture-naming tasks, where answers were expected to be faster and more correct in German than in English for participants in German contexts. Additionally, groups in both contexts were expected to translate and name semantically organized lists

slower and less accurately compared to mixed lists. This difference was hypothesized to be particularly pronounced when translating from the more salient language (e.g., German in a German context) into the less salient one (English in a German context). In other words, we again expected *category interference*, as in Experiment 1, and therefore that specifically semantically organized lists would result in longer RT and worse accuracy when translating from German into English in a German context. Importantly, if the experimental context was able to give one of the experimental languages a ‘pseudo-native’ status, then we would expect an analogous, but opposite pattern for English in the English context condition (asymmetry and interference when translating from English into German instead).

3. Experiment 2: Experimental Language

Experiment 2 manipulated the experimental language context, with participants receiving either German or English instructions as well as being shown a short film clip to prime them in the respective language. The actual online experiment was otherwise identical to that in Experiment 1 (which used only German instructions), except that no immersion questionnaire was used.

3.1. Methods

3.1.1. Participants

Sample requirements and recruitment were the same as in Experiment 1. Seventy-one non-German and non-English multilinguals living in Germany initially took part. Of these, 30 participants had to be excluded due to either technical problems ($N = 5$), not reaching the minimum accuracy of 50% in the experiment ($N = 8$), for being early multilinguals or because their mother tongue was German ($N = 14$), or for not adhering to the experimental instructions ($N = 2$). The final sample ($n = 41$) consisted of 21 female and 20 male participants, with an average age of 33 years ($M = 32.78$, $SD = 12.39$, range = 18–75).

Participants spoke on average 4 different languages ($M = 4.15$, $SD = 1.17$, range = 3–7), including German, English and their native language. This could be Portuguese ($N = 8$), Spanish ($N = 5$), Arabic ($N = 3$), Dutch ($N = 3$), French ($N = 3$), Italian ($N = 3$), Russian ($N = 2$), Arabic/French/Amazigh ($N = 1$), Armenian ($N = 1$), Azerbaijani ($N = 1$), Bulgarian ($N = 1$), Chinese, ($N = 1$), Fula ($N = 1$), Georgian ($N = 1$), Hindi ($N = 1$), Hungarian ($N = 1$), Persian ($N = 1$), Russian/Ukrainian ($N = 1$), Upper Silesian ($N = 1$), Turkish ($N = 1$), or Hungarian ($N = 1$). A full overview of languages spoken by participants can be found in Table S9 in the Supplemental Materials.

Contextual use of languages. Participants had lived in Germany ($N = 39$) or in Austria ($N = 2$; please note that German is the official language everywhere in Austria) for an average of 10.31 years ($SD = 9.79$, range = 0.42–50). Despite German not being their native language, the majority (83%) used it in their daily life, while 22% used their native language and 20% used English. At work, 86% used German and 22% used English; only 7% used their native language. In contrast, most participants usually reported thinking in their native language (73%); 32% tended to think in German, while 25% thought in English (multiple responses possible).

Age of acquisition. Participants, on average, started learning German during early adulthood at the age of 18 years ($M = 18.02$, $SD = 8.94$, range = 5–37) and learning English at 12 years ($M = 11.41$, $SD = 4.17$, range = 6–25), making them late multilinguals in both languages. At the time of the experiment, participants had thus been using German for about 15 years ($M = 14.76$, $SD = 10.42$, range = 1–55) and English for 21 years ($M = 21.37$, $SD = 11.17$, range = 6–50).

Proficiency. On a scale of 1 to 6, participants rated their proficiency in German as 4.88 ($SD = 0.84$, range = 3–6) and in English as 4.39 ($SD = 0.97$, range = 2–6); see Experiment 1 for further explanations). Participants who received German ($N = 19$) or English instructions ($N = 22$) were similar with regard to all language and residence variables (for an exhaustive overview, please refer to Table S10 in the Supplemental Materials).

3.1.2. Materials

Experiment 2 was identical to Experiment 1, except that all instructions were given in either German or English, thus establishing two distinct experimental contexts. Furthermore, to strengthen the contextual language, participants were primed by watching a short trailer of the movie “Inside Out” (approx. 5 min), which was shown before the start of the experiment in either English or German depending on the language context.

3.1.3. Design

A mixed design compared the performance of participants in two distinct immediate language contexts (between participants): German and English. Apart from this, Experiment 2 followed the same design as Experiment 1, with identical tasks (translation and picture-naming tasks; within participants) and list types (mixed, semantically organized, and abstract word lists; within participants) used in the two parallel versions. By comparing participant performance across these two contexts, we aimed to assess the impact of the experimental language on task performance (RT and accuracy) depending on task and list type.

3.1.4. Procedure

The procedure was identical to that in Experiment 1, except for the inclusion of the video clip and the removal of the immersion questionnaire. The video clip was presented before the translation and picture-naming tasks. After the video, in order to simulate a small-talk situation, participants were asked to express their opinions about the movie aloud. Unfortunately, due to a programming error, the receptive use (reading and listening) of German and English was not assessed in the German experimental context. Therefore, the variables of receptive and productive use were not analyzed in Experiment 2.

3.1.5. Data Preparation

Data preparation was identical to that performed for Experiment 1. All statistical analyses incorporated the additional between-participant factor, *experimental context* (German or English instructions), in addition to response language (German or English) and list type (mixed, organized, and abstract lists in translation tasks; mixed and organized lists for picture naming). These were defined as fixed effects in the mixed models, with by-participant and by-item random intercepts allowing for interactions. The response language was also specified with a by-participant varying slope, similarly to Experiment 1.² Contrast coding was computed as in the first experiment, with the additional variable language context being sum-coded (−0.5, 0.5), with German and English as the first and second categories, respectively.

3.2. Results

3.2.1. Language Characteristics

The age of acquisition in German only correlated with the usage of the native language during thinking ($r = .36, p = .02$), suggesting that participants who learned German later in life had a higher tendency to think in their mother tongue than, e.g., in German. Furthermore, consistent with Experiment 1, using German in different situations exhibited a negative correlation with the usage of English, e.g., the correlation between using German and English at work ($r = -.61, p < .001$). This suggested a contradictory relationship between the usage of English and German. A complete presentation of the descriptive values and correlations of language variables can be found in Table S11 of the Supplemental Materials.

3.2.2. Translation Tasks

The mean accuracy was 67% across all contexts, response languages, and list types in translation tasks ($SD = 0.11$), and the average RT was 1651 ms ($SD = 193$). Means are depicted in Figure 4, and further specified in Table 6, depending on the instruction language, task type, response language, and list type.

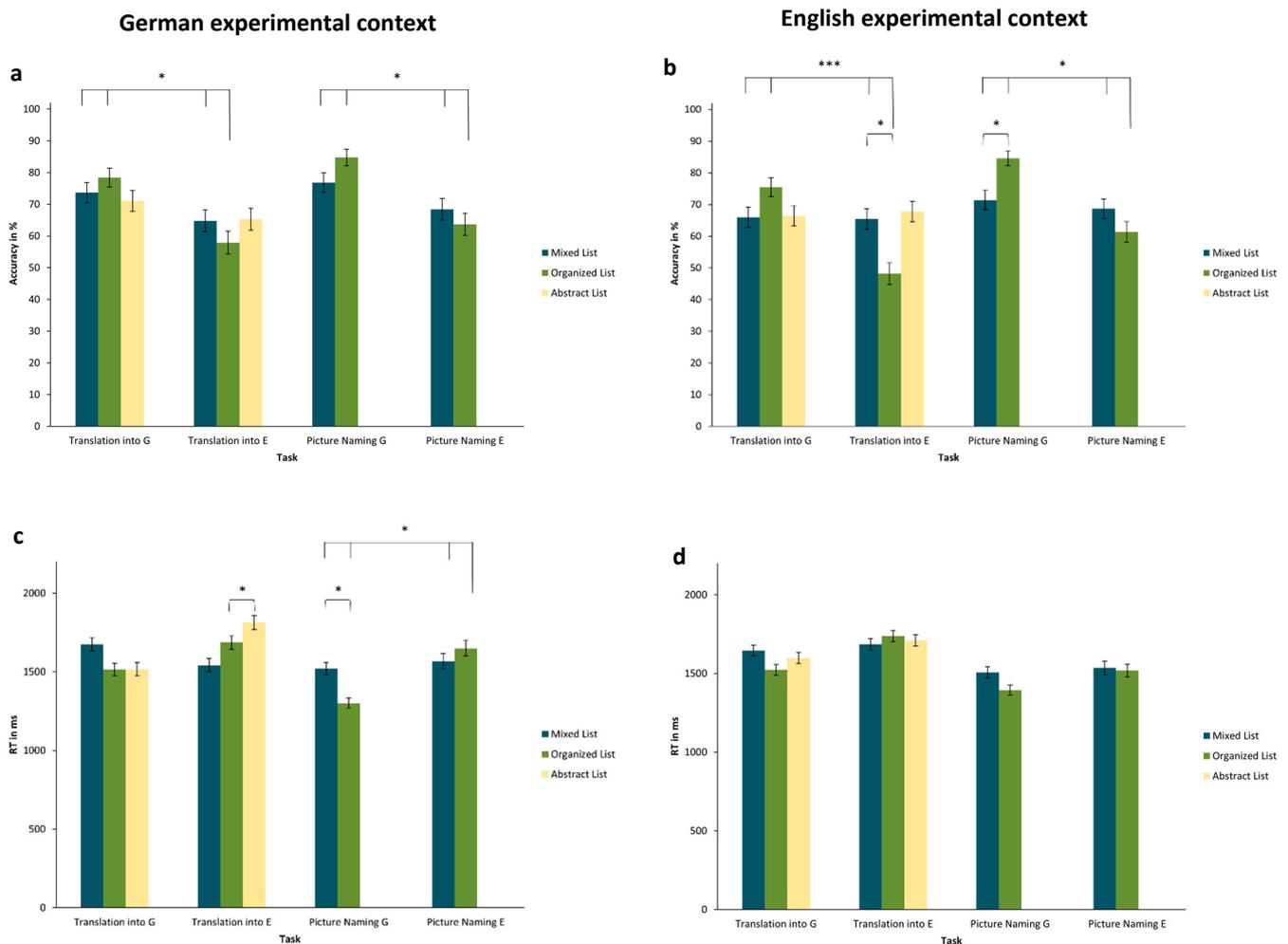


Figure 4. Mean accuracies of participants in the (a) German and (b) English experimental contexts depending on task type, response language, and list type (top). Mean response times of participants in the (c) German and (d) English experimental contexts depending on task type, response language, and list type (bottom). The large brackets on top of each subfigure represent the combined comparison of semantically organized lists in English versus mixed lists in English and organized and mixed lists in German (see also Statistical Analysis described in Experiment 1). Error bars represent standard errors. * $p < .05$. *** $p < .001$.

Accuracy. Table 7 displays the GLME analysis results. Language context, response language, and list type did not significantly affect accuracy separately, nor did context and response language interact (no asymmetry and no asymmetry differences between contexts). However, response language and organized lists interacted significantly, with the highest accuracy found for translating semantically organized lists into German in experiments with German instructions (emmean = 1.80, SE = 0.39) and the lowest for translating organized lists into English with English instructions (emmean = -0.10, SE = 0.37; see Supplemental Table S12 for the full list of estimated marginal means). The post hoc contrast analysis results shown in Table 8 support this finding, demonstrating that translating organized lists into English with English instructions (emmean = -0.10, SE = 0.38) resulted in lower accuracy compared to mixed-word lists (emmean = 0.93, SE = 0.38). Post hoc tests showed further evidence that translating organized lists into English was harder than translating mixed lists in both languages and organized lists in German when participants were exposed to German instructions. Importantly, the same pattern was also observed for English instructions. These results suggest that accuracy is influenced by the level of list type and the response language, demonstrating category interference.

Table 6. Grand means and SD of the dependent variables' (DV) accuracy and response times (RT) separately for each task type, response language, and list type.

DV	Task Type	Response Language	List Type							
			Mixed		Organized		Abstract		Overall	
			M	SD	M	SD	M	SD	M	SD
Accuracy	Translation	German GC	0.74	0.15	0.78	0.15	0.71	0.14	0.74	0.09
		English GC	0.65	0.18	0.58	0.22	0.65	0.2	0.63	0.16
		Overall GC	0.69	0.13	0.68	0.17	0.68	0.14	0.69	0.11
		German EC	0.66	0.2	0.75	0.15	0.66	0.19	0.69	0.11
		English EC	0.65	0.18	0.48	0.22	0.68	0.19	0.6	0.16
		Overall EC	0.66	0.15	0.62	0.14	0.67	0.17	0.65	0.11
		Overall	0.7	0.18	0.77	0.15	0.69	0.17	0.72	0.1
	Picture naming	German GC	0.77	0.15	0.85	0.17			0.81	0.15
		English GC	0.68	0.2	0.64	0.24			0.66	0.18
		Overall GC	0.73	0.11	0.74	0.11			0.73	0.1
		German EC	0.71	0.16	0.85	0.17			0.78	0.15
		English EC	0.69	0.19	0.61	0.24			0.65	0.2
		Overall EC	0.7	0.14	0.73	0.14			0.71	0.13
		Overall	0.74	0.16	0.85	0.17			0.79	0.15
RT	Translation	German GC	1698	264	1542	332	1545	374	1591	289
		English GC	1552	276	1701	292	1859	321	1713	205
		Overall GC	1626	192	1605	275	1690	281	1642	220
		German EC	1685	293	1531	184	1619	195	1593	168
		English EC	1693	273	1734	293	1747	292	1741	221
		Overall EC	1682	260	1616	170	1670	156	1658	171
		Overall	1691	277	1536	260	1585	290	1592	229
	Picture naming	German GC	1628	280	1719	289	1799	307	1728	211
		Overall	1656	230	1611	222	1679	220	1651	193
		German GC	1538	276	1315	201			1424	200
		English GC	1628	343	1683	431			1668	356
		Overall GC	1556	196	1461	214			1509	177
		German EC	1512	222	1400	266			1456	209
		Overall EC	1566	266	1565	309			1559	223
Picture naming	Overall EC	1532	198	1461	185			1496	162	
	German	1524	246	1361	239			1441	203	
	English	1595	302	1620	371			1609	293	
	Overall	1543	195	1461	196			1502	167	

Note. GC and EC represent German and English context, respectively. Values are not provided for abstract lists in the context of picture-naming tasks, as these task do not incorporate such list types. All accuracies range from 0 to 1, as correct responses are coded with 1 and wrong answers with 0. RT is measured in milliseconds. *N* = 41 with *N* = 19 in the German and *N* = 22 in the English context.

Response times. Table 7 also presents the LME analysis results for RT, showing no significant asymmetry effects across or within contexts. The interaction between response

language and organized lists observed was only marginally significant ($p = .08$). However, a significant interaction was observed between the three variables when dealing with abstract lists. The fastest answers occurred when translating organized lists into German under both German ($emmean = 1586, SE = 85$) and English instructions ($emmean = 1575, SE = 82$). Table 8 presents the post hoc contrast analysis results that show a significant difference between translations of mixed and abstract lists into English under German instructions, with the slowest RT observed for abstract lists ($emmean = 1866, SE = 84$).

Table 7. Translation tasks: modeling results for accuracy and RT depending on language context, response language, and list type.

Fixed Effects	Accuracy			Response Times		
	β	$SE(\beta)$	z-Value	β	$SE(\beta)$	t-Value
(Intercept)	1.03	0.25	4.13 ***	1717	53	32.38 ***
Main effect for language context	-0.27	0.26	-1.04	46	72	0.65
Main effect for response language	-0.23	0.47	-0.48	-9	91	0.10
Organized lists	-0.1	0.33	-0.3	-29	63	-0.46
Abstract lists	-0.03	0.32	-0.11	9	62	0.14
Interaction of context with response language	0.57	0.4	1.44	130	92	-1.41
Interaction of context with organized lists	-0.1	0.25	-0.39	-20	59	-0.34
Interaction of context with abstract lists	0.2	0.25	0.82	-66	58	-1.15
Response language \times organized lists	-1.3	0.65	-1.99 *	223	125	-1.78 $^{\circ}$
Response language \times abstract lists	0.03	0.65	0.05	202	124	-1.62
Context \times response language \times organized lists	-0.9	0.5	-1.79 $^{\circ}$	-56	117	0.48
Context \times response language \times abstract lists	-0.12	0.49	-0.25	-266	115	2.31 *

Note. Columns β and $SE(\beta)$ represent estimates and standard errors of the coefficients, respectively. The model intercept was fixed as the reference level (mixed lists for list type and the grand mean for both language context and response language). $^{\circ} p < .1$. * $p < .05$. *** $p < .001$.

Table 8. Translation tasks: post hoc tests comparing context, response language and list type.

Context	Comparisons	Accuracy					Response Times				
		Estimate	SE	df	z-Ratio	p	Estimate	SE	df	t-Ratio	p
German	Mixed vs. organized lists into German	0.38	0.51	Inf	0.74	.46	-144	98	171	-1.47	.14
	Mixed vs. abstract lists into German	-0.18	0.50	Inf	-0.36	.72	-126	99	176	-1.27	.21
	Mixed vs. organized lists into English	-0.48	0.49	Inf	-0.97	.33	106	103	202	1.03	.30
	Mixed vs. abstract lists into English	-0.09	0.49	Inf	-0.18	.86	209	101	193	2.08	.04 *
English	Mixed vs. organized lists into German	0.73	0.50	Inf	1.48	.14	-137	98	164	-1.40	.16
	Mixed vs. abstract lists into German	0.08	0.48	Inf	0.17	.87	-58	98	172	-0.59	.55
	Mixed vs. organized lists into English	-1.02	0.48	Inf	-2.13	.03 *	58	102	194	0.57	.57
	Mixed vs. abstract lists into English	0.05	0.48	Inf	0.11	.91	10	98	170	0.10	.92
German	Organized lists into English vs. mixed lists into English, and mixed and organized lists into German	0.94	0.41	Inf	2.28	.02 *	-105	87	209	-1.20	.23
English	Organized lists into English vs. mixed lists into English, and mixed and organized lists into German	1.23	0.40	Inf	3.05	.001 ***	-141	87	213	-1.62	.11

Note. Accuracy values are provided on the logit scale; no degrees of freedom are available for generalized linear mixed models. * $p < .05$. *** $p < .001$.

3.2.3. Picture-Naming Tasks

Overall, participants were both more accurate ($t(3639) = -4.00, p < .001; M = 0.72, SD = 0.11$) and faster to respond in picture-naming tasks ($t(2515) = 6.55, p < .001; M = 1502 \text{ ms}, SD = 167 \text{ ms}$) than in translation tasks ($M = 1651 \text{ ms}, SD = 193$). Figure 4 and Table 6 provide further details.

Accuracy. The GLMEs presented in Table 9 show asymmetry effects independently of context. Moreover, both the response language and its interaction with list type significantly impacted accuracy, with responses to semantically organized lists in German under German instructions showing the highest accuracy (emmean = 2.34, $SE = 0.42$). Naming organized lists in English with both German (emmean = 0.76, $SE = 0.40$) and English (emmean = 0.65, $SE = 0.38$) instructions proved to be the least accurate method. The post hoc contrast analysis presented in Table 10 clarifies the interaction: while the comparison between organized lists in English and mixed lists in English, as well as organized and mixed lists in German, was again significant, in this case Figure 4 shows that this may have been driven by a surprisingly good performance when naming organized lists in German (see Supplemental Table S12 for the full list of estimated marginal means). For this reason, this cannot be claimed as evidence for category interference in the hypothesized direction.

Table 9. Picture-naming tasks: modeling results for accuracy and RT depending on language context, response language and list type.

Fixed Effects	Accuracy			Response Times		
	β	$SE(\beta)$	z-Value	β	$SE(\beta)$	t-Value
(Intercept)	1.42	0.19	7.61 ***	1546	36	43.02 ***
Main effect for language context	-0.14	0.25	-0.55	-32	57	-0.55
Main effect for response language	-0.97	0.38	-2.55 *	173	68	2.54 *
Main effect for list type	-0.17	0.3	-0.56	89	53	1.69 °
Interaction of context with response language	0.17	0.52	0.31	-130	104	-125
Interaction of context with list type	-0.11	0.26	-0.4	12	60	0.20
Interaction of response language with list type	1.27	0.6	2.1 *	-176	106	-1.67
Interaction of context with response language and list type	0.45	0.53	0.85	210	120	1.75 °

Note. β and $SE(\beta)$ represent estimates and standard errors of the coefficients, respectively. The first fixed-effect level (semantically organized lists in the target language German) was fixed as the reference level. ° $p < .1$. * $p < .05$. *** $p < .001$.

Table 10. Picture-naming tasks: post hoc tests contrasting response language and list type.

Context	Comparisons	Accuracy					Response Times				
		Estimate	SE	df	z-Ratio	p	Estimate	SE	df	t-Ratio	p
German	Mixed vs. organized lists in German	0.64	0.49	Inf	1.31	.19	-224	86	134	-2.61	.01 *
	Mixed vs. organized lists in English	-0.41	0.46	Inf	-0.89	.37	57	92	164	0.63	.53
English	Mixed vs. organized lists in German	0.97	0.47	Inf	2.06	.04 *	-131	84	122	-1.56	.12
	Mixed vs. organized lists in English	-0.52	0.45	Inf	-1.17	.24	-60	89	142	-0.67	.50
German	Organized lists in English vs. mixed lists in English, and mixed and organized lists in German	-0.97	0.44	Inf	-2.22	.03 *	197	86	153	2.28	.02 *
English	Organized lists into English vs. mixed lists into English, and mixed and organized lists into German	-0.94	0.42	Inf	-2.23	.03 *	32	83	148	0.39	.70

Note. Accuracy values are provided on the logit scale; no degrees of freedom are available for generalized linear mixed models. * $p < .05$.

Response times. Table 9 presents the LME analysis results, showing that response language influenced RT significantly (asymmetry effect independent of context), while

list type and its interaction with language context and response language had marginal effects. The post hoc comparisons in Table 10 show that, while no differences were found at all between lists and response languages with English instructions, a similar pattern was found for accuracies under German instructions, with the significant contrast for organized lists in English vs. all other conditions probably driven by the fast RT for organized lists when answering in German ($emmean = 1331, SE = 70$). This suggests that the predicted pattern was found in neither the English nor the German context.

3.2.4. Recall Performance

Participants recalled an average of only 17 words out of a possible 100 items ($M = 16.73, SD = 10.81, range = 0–44$). They recalled more items from picture-naming tasks ($M = 11.15, SD = 6.15, range = 0–22$) than from translation tasks ($M = 5.54, SD = 6.28, range = 0–28$), and more words were named in German ($M = 9.63, SD = 7.17, range = 0–31$) than in English ($M = 7.10, SD = 5.36, range = 0–22$). Furthermore, participants in the English context generally recalled fewer words ($M = 13.18, SD = 7.57, range = 0–29$) than those who received German instructions ($M = 20.84, SD = 12.62, range = 0–44$).

Despite the very low numbers of items recalled, a mixed three-way ANOVA was conducted to examine the influence of context, response language, and list type on the mean count from translation tasks. A main effect was found for the context ($F(1,39) = 8.19, p = .01$), with participants who received German instructions reporting slightly more words ($M = 1.39, SD = 1.26, range = 0–4.67$) than those with English instructions ($M = 0.52, SD = 0.60, range = 0–1.83$). A main effect was also observed for list type ($F(1.59,62.10) = 7.04, p = .003$, Greenhouse–Geisser-corrected). Post hoc analysis using Bonferroni correction showed that this effect was driven by the particularly low rates of abstract words recalled ($M = 0.55, SD = 0.97, range = 0–3$).

For picture-naming tasks, the ANOVA only revealed a significant main effect for the response language ($F(1,39) = 18.19, p < .001$), indicating that participants remembered more words previously named in English ($M = 3.30, SD = 1.86, range = 0–6.5$) than in German ($M = 2.27, SD = 1.58, range = 0–6$), which could be due to a recency effect.

3.3. Summary and Discussion

Experiment 2 manipulated the short-term experimental language context by varying the language of the instructions given to participants and by having them watch a short video in either German or English. Echoing the findings from Experiment 1, accuracy was subject to asymmetry effects (for picture naming) and category interference (for translation tasks). Importantly, the pattern of results was highly comparable between the German and the English instruction conditions, suggesting that the current experiment's contextual language does not render that language more native-like, influencing asymmetry and category effects as reported in the RHM (Kroll and Stewart 1994). At least it seems that the current manipulation, primarily involving the instruction language in an online experiment, may not be sufficient to substantially affect the status of the language in question.

4. Experiment 3: Language Environment

In contrast to the short-term context manipulated in Experiment 2, Experiment 3 focused on a more stable and ongoing form of language context—the language spoken in the country the participants were living in. Even for people in international circles, who may only have limited contact with locals, the surrounding language encountered every day on signposts, in media, in shops, and on public transport (to name only some contexts) (Wigdorowitz et al. 2022, 2023) seems highly likely to affect the relative status of the languages spoken by participants. Thus, comparing asymmetry effects and category interference for non-native participants living in German-speaking countries to those living in a non-German/non-English language community makes it possible to test contextual effects of a much more lasting nature than those examined in Experiment 2.

4.1. Methods

4.1.1. Participants

Experiment 3 required a rather exclusive sample of participants for the non-German context condition, examining people who were currently living in a country where neither German nor English was part of the surrounding language environment, but who nonetheless possessed advanced (but not native) language skills in both German and English. Initially, 39 participants were recruited, but 20 of these had to be excluded for exceeding the error threshold of 50% ($N = 6$), for living in Germany ($N = 5$), for misunderstanding the experiment’s instructions ($N = 1$), or for being early multilinguals in German or English ($N = 6$). For the remaining 19 participants, a yoked control design was applied in order to form a control group, i.e., the German context condition (see Supplemental Material Table S13). Overall, 19 participants were selected from the sample acquired for Experiment 1, which consisted of non-German and non-English native speakers residing in Germany. Participants were matched based on (1) native language (or language family if there was no match) and (2) similar language proficiency in German and English. Table 11 presents descriptive statistics for the total sample, as well as for both contextual conditions.

Table 11. Descriptive values for language and residence variables in the total sample and the two context conditions (German vs. non-German).

Variable	Overall ($N = 38$)			German Context ($N = 19$)			Non-German Context ($N = 19$)		
	M	SD	Range	M	SD	Range	M	SD	Range
Age	32.37	10.84	20–55	37.79	11.75	20–55	26.95	6.42	22–45
Proficiency G	5	0.93	3–6	5	1	3–6	5	0.88	3–6
Proficiency E	4.63	1.08	3–6	4.47	1.02	3–6	4.79	1.13	3–6
AoA G	16.29	7.96	6–39	20.05	9.2	9–39	12.53	3.96	6–23
AoA E	9.42	3.76	4–20	10.58	4	6–20	8.26	3.19	4–16
Duration of learning G	16.08	9.17	3–38	17.74	9.83	5–35	14.42	8.4	3–38
Duration of learning E	22.95	10.02	7–49	27.21	11.81	7–49	18.68	5.33	11–35
Number of languages	4.26	1.46	3–10	4.16	1.21	3–6	4.37	1.71	3–10
Total years of residence	17.35	8.87	1.5–31	12.9	8.07	2–29.08	21.8	7.41	1.5–31
Contextual usage of languages:	Percentage (%)			Percentage (%)			Percentage (%)		
Daily life G	36.59			34.15			2.44		
Daily life E	9.76			9.76			0		
Daily life M	53.66			9.76			43.9		
Work G	53.66			36.59			17.07		
Work E	24.39			12.2			12.2		
Work M	43.9			14.63			29.27		
Thoughts G	24.39			21.95			2.44		
Thoughts E	4.88			2.44			2.44		
Thoughts M	73.17			31.71			41.46		

Note. G, E and M indicate German, English, and mother tongue, respectively. Proficiency was measured on a 6-point scale, with higher values for higher proficiency. Age of acquisition (AoA), duration of language learning, and total years of residence in Germany/Austria were measured in years. Language usage variables were coded with 1 if they applied to the participant, and with 0 if not. Total distribution exceeds 100% as multiple options were possible.

To summarize, the total sample of $N = 38$ participants (29 female, 6 male) had an average age of 32 years ($M = 32.37$, $SD = 10.84$, range = 20–55). As in Experiment 1 and 2, participants were proficient on average in four different languages ($M = 4.26$, $SD = 1.46$, and range = 3–10), including German, English, and their native language. Native languages were Spanish ($N = 8$), Italian ($N = 5$), French ($N = 4$), Dutch ($N = 4$), Portuguese ($N = 4$), Bulgarian ($N = 2$), Hungarian ($N = 2$), Armenian ($N = 1$), Finnish ($N = 1$), Georgian ($N = 1$), Italian, Albanian ($N = 1$), Norwegian ($N = 1$), Romanian ($N = 1$), Swedish ($N = 1$), Serbian ($N = 1$), or Serbo-Croat ($N = 1$). A full overview of participants’ language skills is provided in Table S14 in the Supplemental Materials.

Contextual use of languages. The participants in the non-German countries had been living there for 22 years ($M = 21.80$, $SD = 7.41$, range = 1.5–31); the control group, on the other hand, had been living in Germany or Austria for an average of 13 years ($M = 12.90$, $SD = 8.07$, range = 2–29.08).

Age of acquisition. On average, participants in the non-German context started learning German during early adulthood, at around 16 years of age ($M = 16.29$, $SD = 7.96$, range = 6–39), and English at a younger age of about 9 years ($M = 9.42$, $SD = 3.76$, range = 4–20), thus being late multilinguals in both languages. In total, they had been learning German for approximately 16 years ($M = 16.08$, $SD = 9.17$, range = 3–38) and English for about 22 years ($M = 22.95$, $SD = 10.02$, range = 7–49).

Proficiency. Participants in the non-German context self-reported their proficiency in German as $M = 5.00$ ($SD = 0.93$, range = 3–6) and in English as $M = 4.63$ ($SD = 1.08$, range = 3–6) on a scale of 1 to 6, with 1 representing an A1 proficiency level and 6 representing a C2 proficiency level (Council of Europe 2001). Both means corresponded to an advanced level (C1), which allowed most participants to understand and use a wide range of complex language and express themselves effectively.

Receptive and productive use. Unsurprisingly, participants living in a German context used the German language more frequently for receptive ($M = 6.10$, $SD = 0.57$, range: 5–7) and productive purposes ($M = 6.05$, $SD = 0.78$, range: 4–7) than the group living in the non-German countries (receptive: $M = 5.37$, $SD = 0.76$, range: 4–6; $t(33.28) = -3.38$, $p < 0.01$; productive: $M = 4.68$, $SD = 1.16$, range: 2–6, on a 7-point scale; $t(31.55) = -4.27$, $p < 0.001$). Regarding English, there were no significant differences between participants in the German and non-German contexts for either receptive (German: $M = 5.16$, $SD = 1.07$, range: 2–7; non-German: $M = 5.57$, $SD = 1.43$, range: 2–7; $t(33.35) = 1.02$, $p = 0.31$) or productive use (German: $M = 4.79$, $SD = 1.36$, range: 2–7; non-German: $M = 4.95$, $SD = 1.35$, range: 2–7; $t(36) = 0.36$, $p = 0.72$).

4.1.2. Materials

The materials were identical to those used in Experiment 1, except for the immersion questionnaire, which was not presented to participants living outside Germany.

4.1.3. Design

Experiment 3 followed the same design as Experiment 2, but in this case the between-participant factor was the country of residence: one group was living in German-speaking countries (hereafter: *German context*), the other group in non-German- and non-English-speaking countries (*non-German context*; please note that the simplified labelling as “non-German” reflects the pivotal comparison to the group living in a German context; strictly speaking, “non-German” means both “non-German” and “non-English”). Thus, the non-German context incorporated a variety of environmental languages, none of which were featured in the experiment itself. In every other respect, Experiment 3 shared the same design as Experiment 2, including identical response languages (German and English), task types (translation and picture naming), list types (mixed, organized, and in translation tasks also abstract lists), and parallel versions of the lists.

4.1.4. Procedure

The procedure was identical to Experiment 1.

4.1.5. Data Preparation

Data preparation and statistical analyses were conducted as in Experiment 2. The mixed models of Experiment 3 included the context (German or non-German), response language (English or German), and list types (mixed, organized, and in translation tasks also abstract lists) as fixed effects, with by-participant and by-item random intercepts allowing for interactions. Response language was also specified with a by-participant-varying slope, and all variables were contrast-coded as in Experiment 2.³

4.2. Results

4.2.1. Language Characteristics

Descriptive and correlative values for the two different samples of Experiment 3 are presented in Table S15 of the Supplemental Materials. Whereas Experiment 2 found a negative association between the usage of German and English (such that greater use of one co-occurred with lesser use of the other), Experiment 3 found such a negative association between the usage of German and of participants' native language, particularly for those living in non-German contexts, suggesting that the usage of the native language decreased when using German frequently. For example, using German in daily life correlated negatively with the use of one's native language at work ($r = -.33, p = .04$), in daily life ($r = -.73, p < .001$), and in thoughts ($r = -.58, p < .001$).

4.2.2. Translation Tasks

The overall accuracy rate for translation tasks was 68% ($SD = 0.12$), with an average response time of 1579 ms ($SD = 233$). More detailed information can be found in Figure 5, and in Table 12.

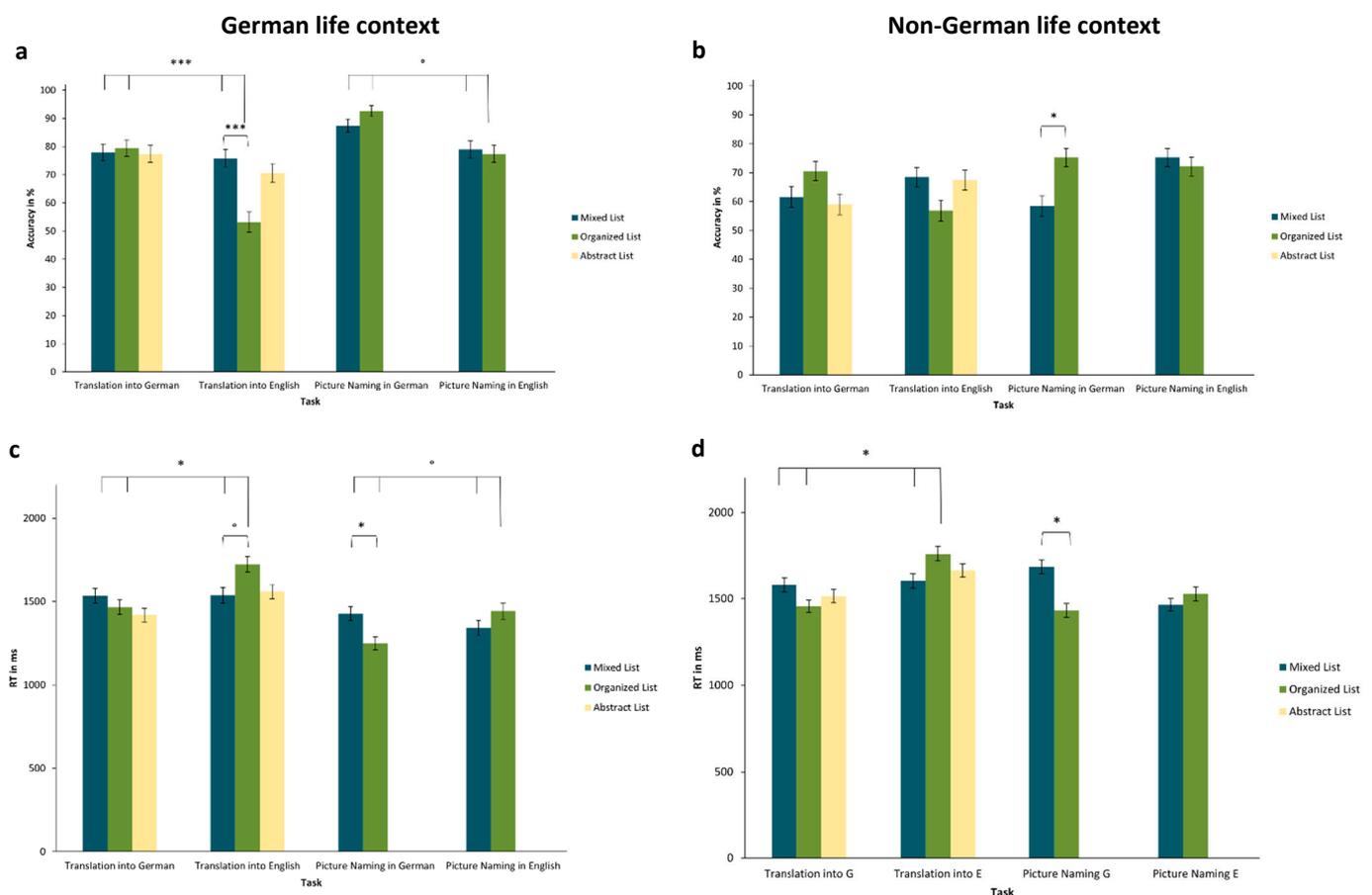


Figure 5. Mean accuracies of participants in the (a) German and (b) non-German life context depending on task type, response language, and list type (top). Mean response times of participants in the (c) German and (d) non-German life context depending on task type, response language, and list type (bottom). The large brackets on top of each subfigure represent the combined comparison of semantically organized lists in English versus mixed lists in English and organized and mixed lists in German (see also Statistical Analysis described in Experiment 1). Error bars represent standard errors. $^{\circ} p < .1$. $* p < .05$. $*** p < .001$.

Table 12. Mean accuracy and RT for each task, response language, and list type.

DV	Task Type	Response Language	List Type							
			Mixed		Organized		Abstract		Overall	
			<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Accuracy	Translation	German GC	0.78	0.14	0.79	0.14	0.77	0.14	0.78	0.09
		English GC	0.76	0.15	0.53	0.2	0.71	0.19	0.66	0.16
		Overall GC	0.77	0.12	0.66	0.13	0.74	0.14	0.72	0.11
		German non-GC	0.62	0.23	0.71	0.16	0.59	0.16	0.64	0.13
		English non-GC	0.68	0.13	0.57	0.18	0.67	0.16	0.64	0.12
		Overall non-GC	0.65	0.15	0.64	0.14	0.63	0.14	0.64	0.11
		Overall	0.7	0.2	0.75	0.16	0.68	0.18	0.71	0.13
	Picture Naming	German GC	0.87	0.14	0.93	0.07			0.9	0.09
		English GC	0.79	0.19	0.77	0.2			0.78	0.19
		Overall GC	0.83	0.12	0.85	0.11			0.84	0.1
		German non-GC	0.58	0.17	0.75	0.15			0.67	0.13
		English non-GC	0.75	0.14	0.72	0.21			0.74	0.16
		Overall non-GC	0.67	0.09	0.74	0.13			0.7	0.09
		Overall	0.73	0.21	0.84	0.15			0.78	0.16
RT	Translation	German GC	1563	297	1495	342	1459	315	1499	273
		English GC	1549	288	1786	365	1591	321	1625	272
		Overall GC	1554	247	1598	322	1521	258	1553	246
		German non-GC	1578	331	1473	272	1535	266	1533	227
		English non-GC	1583	305	1757	283	1675	329	1673	266
		Overall non-GC	1579	274	1608	232	1617	252	1605	222
		Overall	1571	310	1484	305	1497	290	1516	248
	Picture Naming	German GC	1445	335	1256	271			1347	276
		English GC	1395	374	1529	419			1462	359
		Overall GC	1411	321	1355	220			1384	251
		German non-GC	1678	242	1463	341			1547	251
		English non-GC	1478	226	1549	249			1516	217
		Overall non-GC	1559	186	1493	221			1526	180
		Overall	1561	311	1360	321			1447	279

Note. GC and non-GC denote German and non-German contexts, respectively. For accuracies, correct responses were coded with 1 and wrong answers with 0. RT was measured in milliseconds. Total $N = 38$, with $N = 19$ in German and non-German contexts, respectively.

Accuracy. Due to convergence issues with the initial GLME model, the bobyqa optimizer was employed to facilitate model fitting. The results are presented in Table 13. No asymmetry effects were found. The long-term language context showed a significant main effect on accuracy in terms of lower accuracies for the non-German context. In addition, context significantly affected the processing of organized lists. There was also a significant

interaction between response language and semantically organized lists, with translations of organized lists into English exhibiting lower accuracy than translations of organized lists into German (the full set of estimated marginal means is presented in Table S16 of the Supplemental Materials). The post hoc test results shown in Table 14 reveal an interesting pattern. In the German context only, translations into English were less accurate for organized lists than for mixed lists. Similarly, the combined comparison of the translation of organized lists into English vs. mixed lists into English and organized and mixed lists into German differed significantly in the German context. However, neither of these differences were present in the non-German context. Hence, despite showing a similar descriptive pattern, no significant results emerged in the non-German context.

Table 13. Translation tasks: modeling results for accuracy and RT depending on language context, response language, and list type.

Fixed Effects	Accuracy			Response Times		
	β	$SE(\beta)$	z-Value	β	$SE(\beta)$	t-Value
(Intercept)	1.28	0.25	5.16 ***	1595	70	22.9 ***
Main effect for language context	−0.79	0.28	−2.84 **	79	82	0.96
Main effect for response language	0.07	0.45	0.15	9	103	0.09
Organized lists	−0.4	0.32	−1.26	56	70	0.79
Abstract lists	−0.19	0.31	−0.61	−34	69	−0.5
Context × response language	0.58	0.38	1.53	23	97	0.23
Context × organized lists	0.55	0.26	2.09 *	−55	61	−0.91
Context × abstract lists	0.12	0.26	0.46	29	60	0.49
Response language × organized lists	−1.3	0.63	−2.07 *	287	140	2.05 *
Response language × abstract lists	−0.03	0.63	−0.05	141	137	1.03
Context × response language × organized lists	0.34	0.53	0.64	−34	122	−0.28
Context × response language × abstract lists	0.34	0.52	0.66	−63	119	−0.53

Note. β and $SE(\beta)$ represent estimates and standard errors of the coefficients, respectively. The model intercept was fixed as the reference level (mixed lists for list type and the grand mean for both language context and response language). * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 14. Translation tasks: post hoc tests comparing context, response language, and list type.

Context	Comparisons	Accuracy					Response Times				
		Estimate	SE	df	z-Ratio	p	Estimate	SE	df	t-Ratio	p
German	Mixed vs. organized lists into German	0.06	0.50	Inf	0.13	.90	−88	99	163	−0.89	.38
	Mixed vs. abstract lists into German	−0.15	0.49	Inf	−0.31	.76	−105	99	163	−1.06	.29
	Mixed vs. organized lists into English	−1.41	0.48	Inf	−2.94	.001 ***	199	104	191	1.92	.06 °
	Mixed vs. abstract lists into English	−0.36	0.48	Inf	−0.74	.46	36	100	171	0.36	.72
Non-German	Mixed vs. organized lists into German	0.44	0.48	Inf	0.93	.35	−126	103	189	−1.22	.22
	Mixed vs. abstract lists into German	−0.20	0.47	Inf	−0.43	.67	−44	104	201	−0.42	.67
	Mixed vs. organized lists into English	−0.69	0.47	Inf	−1.46	.14	127	104	193	1.22	.22
	Mixed vs. abstract lists into English	−0.06	0.47	Inf	−0.13	.89	34	101	182	0.33	.74
German	Organized lists into English vs. mixed lists into English, and mixed and organized lists into German	1.58	0.39	Inf	4.04	.001 ***	−235	90	212	−2.61	.01 *
Non-German	Organized lists into English vs. mixed lists into English, and mixed and organized lists into German	0.60	0.38	Inf	1.56	.12	−190	89	208	−2.13	.03 *

Note. Accuracy values are provided on the logit scale; no degrees of freedom are available for generalized linear mixed models. ° $p < 0.1$. * $p < 0.05$. *** $p < 0.001$.

Response times. Regarding the LME results provided in Table 13, the only significant finding was an interaction between response language and semantically organized lists, with slower responses for translating organized lists into English than into German. Post hoc tests revealed a marginal difference between translating mixed and organized lists

into German, but this was only the case in the German context. However, the combined comparison was significant for both contexts, implying a very similar pattern of response time results in both contexts.

4.2.3. Picture-Naming Tasks

Overall, the mean accuracy rate was higher ($t(3476) = -6.20, p < .001; M = 0.77, SD = 0.12$) and mean RT was slightly faster ($t(2515) = 5.58, p < .001; M = 1455 \text{ ms}, SD = 227$) in picture-naming compared to translation tasks (accuracy: $M = 0.68, SD = 0.12$; RT: $M = 1579 \text{ ms}, SD = 233$). Further descriptive information is supplied in Figure 5 and in Table 12.

Accuracy. Table 15 presents the results of the GLMEs. The response language had no main effect: for picture naming in English, the mean accuracy was 0.76 ($SD = 0.43$), which was almost identical to that of German ($M = 0.78, SD = 0.41$). In contrast to Experiment 2, context showed a main effect on accuracy, with more correct responses in the German ($M = 0.84, SD = 0.10$) than in the non-German context ($M = 0.70, SD = 0.09$). There was also a significant interaction of context with response language, with a higher accuracy found for answers in German for participants living in a German context compared to non-German contexts. Furthermore, there was an interaction between response language and list type, with substantially more accurate answers in German when naming items from semantically organized lists than from mixed lists, while naming from mixed lists in English was slightly more accurate than naming on the basis of organized lists (see also Figure 5 and the full set of estimated marginal means presented in Table S16 of the Supplemental Materials).

Table 15. Picture-naming tasks: modeling results for accuracy and RT depending on language context, response language, and list type.

Fixed Effects	Accuracy			Response Times		
	β	$SE(\beta)$	z -Value	β	$SE(\beta)$	t -Value
(Intercept)	1.7	0.17	10.02 ***	1492	44	33.71 ***
Main effect for language context	-1.1	0.23	-4.7 ***	157	74	2.13 *
Main effect for response language	-0.34	0.35	-0.99	25	70	0.36
Main effect for list type	-0.36	0.27	-1.33	58	57	1.02
Context \times response language	1.48	0.51	2.90 **	-169	101	-1.67
Context \times list type	-0.05	0.3	-0.17	56	58	0.97
Response language \times list type	1.09	0.55	2.00 *	-310	114	-2.73 **
Context \times response language \times list type	0.27	0.60	0.44	-7	116	-0.06

Note. β and $SE(\beta)$ represent estimates and standard errors of the coefficients, respectively. The model intercept was fixed as the reference level (mixed lists for list type and the grand mean for both language context and response language). * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

Post hoc tests revealed that the combined comparison for organized lists into English compared to mixed lists into English and organized and mixed lists into German was marginally significant in the German context, but not in the non-German context. Moreover, in the non-German context, there was an advantage for organized lists over mixed lists in German (see Table 16).

Response time. For picture naming in English, the mean response time was 1441 ms ($SD = 601.85 \text{ ms}$), very similar to the result for German ($M = 1423 \text{ ms}, SD = 573.60$). LME analysis results are displayed in Table 15, emphasizing the main effect of context on RT, with longer RTs seen for participants in the non-German context ($M = 1384 \text{ ms}, SD = 251$) compared to the German context ($M = 1526 \text{ ms}, SD = 180$), and a significant interaction between response language and list type. Similar to the pattern found for accuracies, participants were substantially faster when naming items in German from organized lists than from mixed lists, while naming mixed lists in English was slightly faster than naming organized lists (again, the full set of estimated marginal means are presented in Table S16 of the Supplemental Materials).

Table 16. Picture-naming tasks: post hoc tests comparing language context, response language, and list type.

Context	Comparisons	Accuracy					Response Times				
		Estimate	SE	df	z-Ratio	p	Estimate	SE	df	t-Ratio	p
German	Mixed vs. organized lists in German	0.82	0.51	Inf	1.60	.11	−184	88	112	−2.07	.04 *
	Mixed vs. organized lists in English	−0.14	0.43	Inf	−.33	.74	123	91	125	1.36	.18
Non-German	Mixed vs. organized lists in German	1.00	0.41	Inf	2.45	.01 *	−243	95	147	−2.55	.01 *
	Mixed vs. organized lists in English	−0.22	0.41	Inf	−.54	.59	71	92	131	0.77	.44
German	Organized lists in English vs. mixed lists in English, and mixed and organized lists in German	−0.81	0.41	Inf	1.98	.05 ^o	155	84	132	1.84	.07 ^o
Non-German	Organized lists in English vs. mixed lists in English, and mixed and organized lists in German	0.12	0.38	Inf	0.30	.76	8	86	142	0.09	.93

Note. Accuracy values are provided on the logit scale; no degrees of freedom are available for generalized linear mixed models. ^o $p < .1$. * $p < .05$.

While the post hoc tests in Table 16 once more show the advantage of using organized lists for the response language German (here in both contexts), the more important observation is again that the combined comparison of the organized lists in English compared to the other three lists shows a marginally significant effect. However, as before, this is only for the German context.

4.2.4. Recall Performance

On average, participants remembered around 23 words ($M = 22.79$, $SD = 12.26$, range = 0–56). As in Experiment 1 and 2, the number of words recalled was higher in picture-naming tasks ($M = 14.16$, $SD = 7.00$, range = 0–27) compared to translation tasks ($M = 8.89$; $SD = 7.38$, range = 0–29). Additionally, participants living in German contexts reported slightly fewer words ($M = 21.11$, $SD = 12.15$, range = 6–56) than those in non-German countries ($M = 24.47$, $SD = 12.46$, range = 0–50). Despite relatively low counts per condition, a mixed three-way ANOVA examined the impact of language context, list type, and response language on recall performance for translation tasks. There were no significant main effects but there was a significant three-way interaction ($F [1.65, 59.41] = 4.34$, $p = .02$, Greenhouse–Geisser-corrected). Two separate repeated-measures ANOVAs for each level of language context revealed no significant effects at all in the non-German context group. However, there was a significant main effect for response language in the German context group ($F [1, 18] = 4.64$, $p = .045$), as well as an interaction between list type and response language ($F [1.62, 29.13] = 3.96$, $p = .04$, Greenhouse–Geisser-corrected). Pairwise comparisons with Bonferroni correction in the German context revealed a significant difference between German and English for semantically organized lists ($p = .04$), but not for mixed ($p = .47$) or abstract ($p = .18$) lists. Participants from German contexts recalled more words from semantically organized lists that they had translated into English ($M = 2.26$, $SD = 2.08$, range = 0–6) than into German ($M = 0.95$, $SD = 1.61$, range = 0–6), while for mixed and abstract lists, more words were reported when the translations had been required in German (German: $M = 1.79$, $SD = 2.51$, range = 0–7; English: $M = 1.37$, $SD = 1.71$, range = 0–5). While these results should not be overinterpreted, they suggest categorical facilitation for participants for whom the German language was particularly salient because they were living in German-speaking countries.

For picture-naming tasks, the mixed three-way ANOVA revealed that, as in Experiment 1 and 2, response language had a significant effect on recall performance ($F [1, 36] = 26.27, p < .001$), as participants remembered more words in English ($M = 4.18, SD = 2.42, \text{range} = 0\text{--}10$) compared to German ($M = 2.89, SD = 2.24, \text{range} = 0\text{--}8$). There were no significant main effects for list type or language context, but there was a significant interaction between list type and response language ($F [1, 36] = 8.84, p = .01$). Specifically, participants remembered more items that they had named in English ($M = 4.55, SD = 2.34, \text{range} = 0\text{--}10$) compared to German ($M = 2.39, SD = 1.53, \text{range} = 0\text{--}6$), though this finding should not be overinterpreted as it may be caused primarily by a recency effect.

4.3. Summary and Discussion

Experiment 3 compared two distinct samples: 19 participants living in countries where neither German nor English was regularly used versus 19 participants from Experiment 1 chosen based on matching criteria (including comparable proficiency in English and German, see *Participants*). While pure asymmetry effects were less pronounced in the findings of this experiment (only one interaction of the response language with the context, impacting accuracies in the picture-naming task), effects of category interference could be observed, especially in the translation tasks. Notably, while both societal language contexts showed a similar pattern of effects in translation regarding RT, for accuracy, critical differences were only observed in the German context (a parallel observation holds true for picture naming, though with only marginal significance). Despite showing a numerically similar descriptive pattern, no significant results emerged in the non-German context. This suggests that the effects might have been (at least partially) attenuated by the non-German living context. Similarly, the finding that differential effects for organized lists in the recall task were only present for the German context, but not for the non-German context, aligns with this assumption. Overall, the results largely replicate the category interference effect found in all experiments, but to a lesser extent in the non-German than in the German language context.

5. General Discussion

Three experiments tested predictions derived from the RHM by Kroll and Stewart (1994) and replicated the model in an online setting and for two non-native languages, which here were German and English (all participants in all three experiments had other native languages). More specifically, they investigated the role that the current linguistic context may play in rendering one foreign language more salient than the other—in a sense, giving it a ‘pseudo-native’ status relative to other foreign languages, comparable with the L1 in Kroll and Stewart (1994). First, the degree of immersion was shown to be significantly associated with the strength of the asymmetry between the two experimental languages (Experiment 1). Hence, profound immersion experiences seem to be able to render one language more native-like. Second, two different forms of current linguistic context were compared regarding asymmetry effects, category interference, and category facilitation: the language highlighted by the instructions of the current experiment (Experiment 2) vs. the language typically spoken in the country the participant is currently living in (Experiment 3). While simple asymmetries between the two languages were not always clear across the three experiments, participants largely showed the predicted category interference, with the processing of organized versus mixed lists depending on the response language. Consistently, the pattern of results for German (as a foreign language) resembled that of the L1 in the RHM. Regarding the linguistic context, the short-term experimental language context in Experiment 2 showed barely any effect, whereas the longer-term context of living in a German (vs. non-German) environment in Experiment 3 affected the size of the results. We will discuss this and other results below.

List effects occurred across all three experiments. Notably, semantically organized lists often interacted with the response language, such that these particular lists led to lower accuracy and slower responses in English, which is the non-prominent foreign

language (category interference). Following the original logic of the RHM, this is due to the conceptual mediation that is required for translation into L2 (in this case, the non-prominent language), which means that the activation of a semantic category (through a semantically organized list) interferes with rapid lexical access. Complementing category interference in production, [Kroll and Stewart \(1994\)](#) postulated that the associated in-depth processing of the semantically organized lists could result in category facilitation in post-experimental recall. While the recall rates in our online studies were generally low, and while we cannot discount possible recency effects for differences between languages, the recall results of Experiment 3 fit well with this original suggestion, and also with the proposed impact of environmental language context. While, more words were recalled from translations of semantically organized lists into English than into German in the German context, this was not the case in the non-German context (here, more words were recalled from translations of organized lists into German).

In contrast to this relatively frequent interaction of list types and response language, differential effects of linguistic context are less clear. We originally hypothesized a pattern of results favoring German over English within a German context (German instructions or residence in German-speaking countries), and vice versa in the English context of Experiment 2. At least for the short-term experimental context investigated in Experiment 2, context did not significantly impact the effect of list type on RT and accuracy, where both experimental conditions largely showed similar results. In contrast, the non-German context of Experiment 3 was not expected to flip the pattern of results, but rather to reduce it, as it did not highlight either of the two target languages. Hence, in this context, German and English should maintain a similar status as proficient foreign languages. In fact, differences in accuracy were specifically smaller for organized lists in non-German contexts. This is exactly what would be expected if the non-German context reduced the salience differences between German and English. In future research endeavors, it would be an interesting alternative approach to also contrast the German environmental language context with a specific English environmental language context (instead of non-German context here). In such a study, living in an English-speaking environment should actually flip the pattern of results found for the German context, making English the more salient, “pseudo-native” language.

In this context, one notable finding was that German responses were faster and more accurate across most of the tasks and contexts of all three experiments. Thus, we suggest that, even if two foreign languages do not inherently differ in status, when an individual resides within a specific linguistic environment (e.g., Germany or Austria for all participants in Experiments 1 and 2 and for half the sample in Experiment 3), the foreign language spoken in that society may become amplified, receiving a ‘pseudo-native’ status. This may explain why the effects of experimental language context were minimal in Experiment 2, where the daily-life context was German for all participants, and similar asymmetries and categorical interference effects between the response languages were found in both experimental context conditions. In fact, even in the non-German context of Experiment 3, many of the participants are likely to have lived in German-speaking countries previously (they were all proficient in the language and not all places of residence were assessed) and may have invested more personal resources and effort into learning a less frequent foreign language such as German (in contrast to English, which is commonly learnt in school in most countries).

This would lead to a common variable for participants in all three experiments, which was explicitly associated with experimental performance in Experiment 1, *immersion*, or in other words, the process of learning and speaking a new language (such as German) while residing in the respective country and its culture (e.g., Germany). Embedded within this environmental context, learners acquire the language implicitly during interactions with native speakers, analogous to learning a native language ([Morgan-Short et al. 2012](#); [Wigdorowitz et al. 2022, 2023](#)). Consequently, immersion learners often amass vocabulary comparable to that of their native language ([DeKeyser 1986](#); [Dewey 2007](#)) and are able to

articulate the foreign language more naturally (Collentine and Freed 2004) and fluently than their other foreign languages (DeKeyser 1986; Freed et al. 2004). Considering these connections, it appears that the lexicon of the respective language can be accessed more rapidly and easily (Heredia 1997). This heightened fluency can alter asymmetry and category effects, favoring the corresponding language, similarly to the dynamic aspect suggested by Kroll and Stewart (1994), which is mainly based on proficiency. Indeed, the performance seen in the non-German context of Experiment 3 suggests that such changes in language status caused by immersion may well persist even when the individual has left the respective country and no longer encounters the language of immersion on a daily basis.

It is interesting to speculate about the processes causing these long-term effects of acquiring language through immersion. For instance, learning by using different senses (e.g., visually, kinesthetic) or even emotions may enhance learning effects, even in long-term memory (Li and Deng 2023; Tyng et al. 2017; Volpe and Gori 2019). Such situations might occur cumulatively in personal interactions during immersion, linked to cultural experiences that provide an emotional and motivational scaffold for learning processes (Goldoni 2013). Concluding, we propose that future research should investigate the impact of immersion experiences on translation asymmetries and category effects directly and longitudinally, rather than relying on the purely correlative approach adopted in this study. This would help the field to identify the factors of immersion that may impact short- and long-term language memory and processing.

Methodological Considerations

Before concluding, it is important to mention a number of methodological considerations relevant to our study. Critically, all data were acquired online and were drawn from a very diverse but specifically selected group of participants. Although we believe that both of these aspects actually constitute strengths of the design, they led to a number of practical challenges, including difficulties with recruitment, technical problems due to E-Prime Go not running on all computers, and a lack of communicative interaction between the experimenters and the participants. First, participants only included individuals who were proficient in both German and English, though neither of these were allowed to be their L1. These constraints excluded multilinguals who did not have (sufficient) knowledge of (one of) these specific languages or who were early bilinguals, thereby probably favoring individuals with a certain socio-economic status, academic background, and technical competence. This type of selection also meant that participants had a wide range of different L1s in all experiments, as well as countries of residence in Experiment 3. This deliberate diversity serves as a strength, as it allows for the generalization of our findings across a spectrum of linguistic backgrounds rather than confining their applicability to a single language group. Moreover, the heterogeneity of our sample not only enhances the external validity of our results but also underscores the robustness of our findings.

Nonetheless, a more detailed assessment of participants' language biographies and their current language context (e.g., using tools as the Contextual and Individual Linguistic Diversity Questionnaire recently developed by Wigdorowitz et al. 2023) and a more objective measurement of proficiency and immersion would be helpful in clarifying the nature of some of the response patterns. It would also be beneficial to control for participant's familiarity with vocabulary tests, perhaps by providing standard training before the actual tests. At the same time, we would like to point out that any standardized assessment of proficiency would have substantially increased the time that participants spent on the study, in addition to the challenges of implementing some of the language skill dimensions in an online setting. As a compromise, the Council of Europe (2001) criteria constitute clear guidelines for the self-assessment of both productive and receptive language skills. They are also widely familiar in the student populations of European countries which most of our participants were from.

Two potential confounding factors that we cannot discount are possible stimulus effects and effects of the fixed order of experimental tasks. In order to control for stimulus

effects, we performed linear mixed model analysis and included item as random effect in the analyses. Moreover, we created two versions of each experiment, assigning items from semantically ordered lists in one version to the mixed lists of the other version. The necessity of avoiding cognates in designing these lists limited the number of available stimuli—there is only a certain number of non-cognate vegetables that non-native speakers can readily name. Nonetheless, in future studies, it would be ideal to match items better in terms of frequency and difficulty. The fixed order of tasks was originally designed to facilitate the coding of correct answers in the lab by Sperl et al. (2023). This was actually not primarily relevant for the online version of the tasks used here, but was maintained for replication purposes. One task that was fairly clearly influenced by the fixed order was the recall task, where participants consistently recalled more items from English than German (which was always the first language to be tested), and more items from picture naming than from translation. The latter finding could also be due in part to the pictorial representation of the items in the picture-naming tasks. However, on the basis of the current experiments, it is not possible to tease these two factors apart.

Incidentally, we believe that the overall problematically low word counts per condition in the recall tasks resulted from the online nature of the experiment. This task was not announced prior to starting the study, and so when it appeared right at the end, many participants probably closed their browser window as soon as they needed to think harder. Similarly, a strong manipulation of the experimental language context, as attempted in Experiment 2, would be easier in a lab experiment than online, e.g., by having participants interact with a native speaker of one of the two target languages. At the same time, the successful acquisition of language production data from an online experiment—replicating existing lab studies and uncovering RT differences in millisecond resolution—is a merit in itself, as it makes it more feasible to conduct multilingual research with participants living anywhere on earth, thus massively increasing the range of potential languages and language contexts for investigation.

6. Conclusions

To our knowledge, these are the first results to suggest that the RHM may be broadened and applied to various foreign languages instead of only a native and a foreign language, because individual foreign languages may attain a ‘pseudo-native’ status under certain daily-life contexts. However, the complex pattern of results also suggests that long-term residence is probably not the only factor involved in determining the relative status of a language, and further research examining language differences in context is needed. By addressing the role of contextual factors, we hope to enhance understanding of how language processing operates in non-native multilingual speakers and to clarify the importance of contextual factors for multilingual research all around the world.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/languages9030106/s1>, Supplemental Materials Tables S1–S16.

Author Contributions: Conceptualization, L.S., H.K. (Helene Kreysa), and H.K. (Helene Kühn); methodology, all authors; software, L.S. and H.K. (Helene Kreysa); formal analysis, H.K. (Helene Kreysa), M.S.N.T. and H.K. (Helene Kühn); investigation, M.S.N.T. and H.K. (Helene Kühn); resources, H.K. (Helene Kreysa); data curation, M.S.N.T.; writing—original draft preparation, M.S.N.T.; writing—review and editing, L.S. and H.K. (Helene Kreysa); visualization, M.S.N.T.; supervision, H.K. (Helene Kreysa) and L.S.; project administration, L.S. and H.K. (Helene Kreysa). All authors have read and agreed to the published version of the manuscript.

Funding: The research reported here received no external funding. For publication, we acknowledge support by the German Research Foundation Projekt-Nr. 512648189 and the Open Access Publication Fund of the Thuringer Universitaets- und Landesbibliothek Jena.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of the Faculty of

Social and Behavioral Sciences at Friedrich Schiller University of Jena (approval no. FSV 19/08, 28 February 2019).

Informed Consent Statement: Informed consent was obtained from all participants involved in the study.

Data Availability Statement: Data will be made available on OSF at doi:10.17605/OSF.IO/4MRY6. Until the upload is completed, it is available on request from the authors.

Acknowledgments: We would like to thank Anna-Sophie Stegmann and Iris Bergmann for their parts in programming and testing Experiments 2 and 3, respectively. Anna Schröger provided valuable assistance with the analysis. We are also grateful to Sarah Strüber, Luca Diehlmann, and Verena Steinhof for their support in participant acquisition and data preparation, to Clara M. Breier for proof-reading and formatting assistance, and to Stefan R. Schweinberger and Markus Ullsperger for support and supervision.

Conflicts of Interest: The authors declare no conflicts of interest.

Notes

- ¹ $RT \sim \text{Response Language} * \text{List Type} + (1 + \text{Response Language}|\text{Subject}) + (1|\text{Item})$ Accuracy $\sim \text{Response Language} * \text{List Type} + (1 + \text{Response Language}|\text{Subject}) + (1|\text{Item})$.
- ² $RT \sim \text{Context} * \text{Response Language} * \text{List Type} + (1 + \text{Response Language}|\text{Subject}) + (1|\text{Item})$ Accuracy $\sim \text{Context} * \text{Response Language} * \text{List Type} + (1 + \text{Response Language}|\text{Subject}) + (1|\text{Item})$.
- ³ $RT \sim \text{Context} * \text{Response Language} * \text{List Type} + (1 + \text{Response Language}|\text{Subject}) + (1|\text{Item})$. Accuracy $\sim \text{Context} * \text{Response Language} * \text{List Type} + (1 + \text{Response Language}|\text{Subject}) + (1|\text{Item})$.

References

- Aronin, Larissa, and David M. Singleton. 2012. *Multilingualism. Impact: Studies in Language and Society*. Amsterdam: John Benjamins Pub. Co., vol. 30.
- Baayen, R. Harald, Douglas J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59: 390–412. [CrossRef]
- Baker-Smemoe, Wendy, Dan P. Dewey, Jennifer Bown, and Rob A. Martinsen. 2014. Variables Affecting L2 Gains during Study Abroad. *Foreign Language Annals* 47: 464–86. [CrossRef]
- Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67: 1–48. [CrossRef]
- Beatty-Martínez, Anne L., Christian A. Navarro-Torres, Paola E. Dussias, María Teresa Bajo, Rosa E. Guzzardo Tamargo, and Judith F. Kroll. 2020. Interactional context mediates the consequences of bilingualism for language and cognition. *Journal of Experimental Psychology, Learning, Memory, and Cognition* 46: 1022–47. [CrossRef]
- Boersma, Paul, and David Weenink. 2016. Praat (Version 6.1.51) [Computer Software]. Available online: <http://www.praat.org/> (accessed on 24 August 2023).
- Brehm, Laurel, and Phillip M. Alday. 2022. Contrast coding choices in a decade of mixed models. *Journal of Memory and Language* 125: 104334. [CrossRef]
- Brodeur, Mathieu B., Emmanuelle Dionne-Dostie, Tina Montreuil, and Martin Lepage. 2010. The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS ONE* 5: e10773. [CrossRef]
- Broersma, Mirjam, and Kees de Bot. 2006. Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative. *Bilingualism: Language and Cognition* 9: 1–13. [CrossRef]
- Brown, Amanda, and Marianne Gullberg. 2008. Bidirectional Crosslinguistic Influence in L1-L2 Encoding of Manner in Speech and Gesture: A Study of Japanese Speakers of English. *Studies in Second Language Acquisition* 30: 225–51. [CrossRef]
- Brown, Violet A. 2021. An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science* 4: 251524592096035. [CrossRef]
- Cenoz, Jasone. 2013. Defining Multilingualism. *Annual Review of Applied Linguistics* 33: 3–18. [CrossRef]
- Collentine, Joseph, and Barbara F. Freed. 2004. Learning Context and Its Effects on Second Language Acquisition: Introduction. Available online: <https://www.cambridge.org/core/journals/studies-in-second-language-acquisition/article/learning-context-and-its-effects-on-second-language-acquisition-introduction/ae07e84dc413a402ac91e34952439e0e> (accessed on 24 August 2023).
- Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (8. pr)*, Klett ed. Cambridge: Cambridge University Press.
- Crystal, David. 2007. *English as a Global Language*, 2nd ed. Cambridge: Cambridge University Press.
- Degani, Tamar, Anat Prior, and Natasha Tokowicz. 2011. Bidirectional transfer: The effect of sharing a translation. *Journal of Cognitive Psychology* 23: 18–28. [CrossRef]

- DeKeyser, Robert Michel. 1986. *From Learning to Acquisition? Foreign Language Development in a U.S. Classroom and during a Semester Abroad*. Stanford: Stanford University.
- Dewey, Dan P. 2007. Language learning during study abroad: What we know and what we have yet to learn. *Japanese Language and Literature* 41: 245–269. [CrossRef]
- Elston-Güttler, Kerrie E., Thomas C. Gunter, and Sonja A. Kotz. 2005. Zooming into L2: Global language context and adjustment affect processing of interlingual homographs in sentences. *Brain Research. Cognitive Brain Research* 25: 57–70. [CrossRef]
- Freed, Barbara F., Norman Segalowitz, and Dan P. Dewey. 2004. Context of Learning and Second Language Fluency in French: Comparing Regular Classroom, Study Abroad, and Intensive Domestic Immersion Programs. *Studies in Second Language Acquisition* 26: 275–301. [CrossRef]
- French, Robert M., and Maud Jacquet. 2004. Understanding bilingual memory: Models and data. *Trends in Cognitive Sciences* 8: 87–93. [CrossRef] [PubMed]
- Goldoni, Federica. 2013. Students' Immersion Experiences in Study Abroad. *Foreign Language Annals* 46: 359–76. [CrossRef]
- Grosjean, François. 1989. Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language* 36: 3–15. [CrossRef]
- Grosjean, François, ed. 2021. *Life as a Bilingual: Knowing and Using Two or More Languages*. Cambridge: Cambridge University Press. [CrossRef]
- Gürel, Ayşe. 2004. Selectivity in L2-induced L1 attrition: A psycholinguistic account. *Journal of Neurolinguistics* 17: 53–78. [CrossRef]
- Heredia, Roberto R. 1997. Bilingual memory and hierarchical models: A case for language dominance. *Current Directions in Psychological Science* 6: 34–39. [CrossRef]
- Juan-Garau, Maria, and Roy Lyster. 2018. Becoming Bilingual through Additive Immersive Programs. In *The Cambridge Handbook of Bilingualism*. Edited by Annick de Houwer and Lourdes Ortega. Cambridge: Cambridge University Press, pp. 213–32.
- Kearney, Erin. 2010. Cultural Immersion in the Foreign Language Classroom: Some Narrative Possibilities. *The Modern Language Journal* 94: 332–36. [CrossRef]
- Kemp, Charlotte. 2009. Defining Multilingualism. In *The Exploration of Multilingualism: Development of Research on L3, Multilingualism and Multiple Language Acquisition*. AILA Applied Linguistics Series, vol. 6. Edited by Larissa Aronin and Britta Hufeisen. Amsterdam: John Benjamins, pp. 11–26.
- Kinginger, Celeste. 2008. Language Learning in Study Abroad: Case Studies of Americans in France. *The Modern Language Journal* 92: 1–124. [CrossRef]
- Knight, Susan M., and Barbara C. Schmidt-Rinehart. 2002. Enhancing the Homestay: Study Abroad from the Host Family's Perspective. *Foreign Language Annals* 35: 190–201. [CrossRef]
- Kroll, Judith F., and Alexandra Sholl. 1992. Lexical and Conceptual Memory in Fluent and Nonfluent Bilinguals. In *Cognitive Processing in Bilinguals*. Amsterdam: Elsevier, pp. 191–204. [CrossRef]
- Kroll, Judith F., and Erika Stewart. 1994. Category Interference in Translation and Picture Naming: Evidence for Asymmetric Connections between Bilingual Memory Representations. *Journal of Memory and Language* 33: 149–74. [CrossRef]
- Kroll, Judith F., and Natasha Tokowicz. 2001. The development of conceptual representation for words in a second language. In *One Mind, Two Languages: Bilingual Language Processing*. Edited by J. L. Nicol. Hoboken: Blackwell Publishers, pp. 49–71.
- Kroll, Judith F., Erica Michael, Natasha Tokowicz, and Robert Dufour. 2002. The development of lexical fluency in a second language. *Second Language Research* 18: 137–71. [CrossRef]
- Kroll, Judith F., Janet G. Van Hell, Natasha Tokowicz, and David W. Green. 2010. The Revised Hierarchical Model: A critical review and assessment. *Bilingualism: Language and Cognition* 13: 373–81. [CrossRef]
- Lagrou, Evelyne, Robert J. Hartsuiker, and Wouter Duyck. 2011. Knowledge of a second language influences auditory word recognition in the native language. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 37: 952–65. [CrossRef]
- Lenth, Russell V. 2023. emmeans: Estimated Marginal Means, aka Least-Squares Means [R Package Version 1.8.5]. Available online: <https://CRAN.R-project.org/package=emmeans> (accessed on 28 August 2023).
- Li, Jianhua, and Sophia W. Deng. 2023. Facilitation and interference effects of the multisensory context on learning: A systematic review and meta-analysis. *Psychological Research* 87: 1334–52. [CrossRef]
- Lin, Zhong, and Lei Lei. 2020. The Research Trends of Multilingualism in Applied Linguistics and Education (2000–2019): A Bibliometric Analysis. *Sustainability* 12: 6058. [CrossRef]
- Montrul, Silvina. 2015. Dominance and proficiency in early and late bilingualism. In *Language Dominance in Bilinguals: Issues of Measurement and Operationalization*. Edited by Jeanine Treffers-Daller and Carmen Silva-Corvalan. Cambridge: Cambridge University Press, pp. 15–35.
- Morgan-Short, Kara, Karsten Steinhauer, Cristina Sanz, and Michael T. Ullman. 2012. Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *Journal of Cognitive Neuroscience* 24: 933–47. [CrossRef]
- Palmer, Shekeila D., Johanna C. van Hooff, and Jelena Havelka. 2010. Language representation and processing in fluent bilinguals: Electrophysiological evidence for asymmetric mapping in bilingual memory. *Neuropsychologia* 48: 1426–37. [CrossRef]
- Pavlenko, Aneta. 2011. *Emotions and Multilingualism [Reprinted]*. Cambridge: Cambridge University Press. [CrossRef]
- Pedersen, Eric R., Clayton Neighbors, Mary E. Larimer, and Christine M. Lee. 2011. Measuring Sojourner Adjustment among American students studying abroad. *International Journal of Intercultural Relations* 35: 881–89. [CrossRef]

- Potter, Mary C., Kwok-Fai So, Barbara Von Eckardt, and Laurie B. Feldman. 1984. Lexical and conceptual representation in beginning and proficient bilinguals. *Journal of Verbal Learning and Verbal Behavior* 23: 23–38. [CrossRef]
- Psychology Software Tools Inc. 2020. *E-Prime Go* [Computer Software]. Available online: <http://www.pstnet.com> (accessed on 28 August 2023).
- Sperl, Laura, Anna Schroeger, Juergen M. Kaufmann, and Helene Kreysa. 2023. Mental Representation of Words and Concepts in Late Multilingualism: A Replication and Extension of the Revised Hierarchical Model. Available online: <https://osf.io/preprints/psyarxiv/yfktw> (accessed on 15 November 2023).
- Treffers-Daller, Jeanine. 2019. What Defines Language Dominance in Bilinguals? *Annual Review of Linguistics* 5: 375–93. [CrossRef]
- Tyng, Chai M., Hafeez U. Amin, Mohamad N. M. Saad, and Aamir S. Malik. 2017. The Influences of Emotion on Learning and Memory. *Frontiers in Psychology* 8: 1454. [CrossRef] [PubMed]
- Volpe, Gualtiero, and Monica Gori. 2019. Multisensory Interactive Technologies for Primary Education: From Science to Technology. *Frontiers in Psychology* 10: 1076. [CrossRef] [PubMed]
- Whalley, Ben. 2019. Fitting Multilevel Models in R | Just Enough R. May 20. Available online: <https://benwhalley.github.io/just-enough-r/fitting-models.html> (accessed on 24 August 2023).
- Wigdorowitz, Mandy, Ana I. Pérez, and Ianthi M. Tsimpli. 2022. Sociolinguistic context matters: Exploring differences in contextual linguistic diversity in South Africa and England. *International Multilingual Research Journal* 16: 345–64. [CrossRef]
- Wigdorowitz, Mandy, Ana I. Pérez, and Ianthi M. Tsimpli. 2023. A holistic measure of contextual and individual linguistic diversity. *International Journal of Multilingualism* 20: 469–87. [CrossRef]
- Wilkinson, Sharon. 1998. On the Nature of Immersion during Study Abroad: Some Participant Perspectives. *Frontiers: The Interdisciplinary Journal of Study Abroad* 4: 121–38. [CrossRef]
- Winter, Bodo. 2020. *Statistics for Linguists: An Introduction Using R*. London: Routledge.
- Wu, Yan Jing, and Guillaume Thierry. 2010. Investigating bilingual processing: The neglected role of language processing contexts. *Frontiers in Psychology* 1: 178. [CrossRef]
- Zheng, Bingham, Sandra Báez, Li Su, Xia Xiang, Susanne Weis, Agustín Ibáñez, and Adolfo M. García. 2020. Semantic and attentional networks in bilingual processing: Fmri connectivity signatures of translation directionality. *Brain and Cognition* 143: 105584. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.