How do native Dutch speakers acquire German grammatical gender? Extending the Incidental L2 Learning Paradigm

Jefta B. Lagerwerf[1]

[1]*Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour; The Netherlands*

This study had one general aim: extending the L2 learning paradigm developped by Brandt, Schriefers and Lemhöfer (2021). This was done in three ways. First, the target population changed from German (L1) to Dutch (L1) speakers, to see if the paradigm could be employed in a different population. Second, previous research indicated that inexperienced participants learned the target grammatical gender structure merely by being exposed to the experiment's stimuli (Brandt, Schriefers & Lemhöfer, submitted). Extending this finding, an analysis of learning over time was performed rather than input-driven learning only. Last, there was an extra category of words that was formally analysed. In previous research, only the *incompatible* nouns – nouns that deviated from the expected grammatical gender – were analysed (Brandt, Schriefers & Lemhöfer, 2021, submitted), while in the current study the *compatible* nouns – nouns that conformed to the expected grammatical gender – were analysed equally. The methods were otherwise as close as possible to Brandt, Schriefers and Lemhöfer (2021) for ease of comparison. Most findings were replicated, showing that Dutch (L1) speakers indeed implicitly learn German articles when input was provided. The time-based analysis showed that inexperienced Dutch (L1) participants learned by merely being exposed to the stimuli, and did so nonlinearly. There were some deviating findings from the previous studies, which implied that there were qualitative differences in experience between the two sampled populations. This probably related to the more classroom-based experience of Dutch (L1) speakers compared to the more naturalistic experience of German (L1) participants.

Keywords: language transfer, dialogue game, grammatical gender, incidental article learning.

Introduction

In a general survey of European citizens called the Eurobarometer, the competences and attitudes regarding multilingualism in Europe were determined (TNS Opinion & Social, 2012). Among the surveyed citizens, 84% posited that every EU citizen should speak at least one foreign language in addition to their mother tongue. This simple finding neatly summarizes the perceived value by Europeans of speaking multiple languages, as it is regarded as a part of European existence by the majority of Europeans. In the same survey, citizens reported they learned their second languages (L2) mostly in formal learning settings such as schools, with reportedly 68% of L2 learning occurring in classroom settings. In agreement with this, a lot of second language acquisition (SLA) research has focused on this classroom setting, ranging from issues with measuring participant's abilities of SLA to the best instructive methods for formal L2 learning. In turn, this has lead to an abundance of literature on which forms of (explicit) corrective feedback are most effective in L2 learning (for reviews see e.g. Norris & Ortega, 2000; Spada & Tomita, 2010; Spada, 2011; Miller & Pan, 2012; Belliveau & Kim, 2013; Plonsky & Brown, 2015; Brown, 2016; Brown, Plonsky & Teimouri, 2018), as corrective feedback is perceived as a useful way to perform L2 language teaching in a formal setting (though see e.g. Best and Tyler (2007) why this classroom focus might be problematic).

It is important to keep in mind that the classroom setting is not the only place where people pick up languages. The previously mentioned survey also reported that Europeans also learned languages by informal conversations with native speakers (15%) and long trips/holidays in other countries (15%; TNS Opinion & Social, 2012). These natural input situations, which usually provide little to no explicit corrective feedback to the learner, are often referred to in the scientific literature as *incidental learning* (e.g., Hulstijn, 2003). In these situations, the focus lies on communicative intent rather than perfect execution of the L2 and learning occurs spontaneously. Since 15% of all reported language learning situations is still a substantial amount, the question arises of how these situations are reflected in the scientific literature? Luckily, this more naturalistic way of learning a language has had its share of dedicated research. However, in contrast to the classroom context, the focus mostly lies on theoretical considerations (e.g. Klein & Perdue, 1997; Ellis, 2006, 2011; Granena & Long, 2013). For naturalistic syntactic learning[1] there seems to be a relative absence of laboratory studies other than artificial grammar studies (Brandt, Schriefers & Lemhöfer, 2021). The absence of laboratory studies is not due to a lack of theoretical interest, but seemingly rather due to a lack of experimental methodologies in which participants remain unaware of the learning-based nature of the study (Brandt, Schriefers & Lemhöfer, 2021).

1 As before in Brandt and colleagues (2021), the learning of grammatical gender of nouns is considered a form of syntactic learning. Although learning an article (combined with a noun) might be considered learning a lexical item, the grammatical gender that is being learned is still a syntactic feature, even if it is learned at the lexical level.

To solve this problem, a method was devised by Brandt and colleagues (2021) based on the confederate scripting technique in dialogue developed by Branigan, Pickering and Cleland (2000). The initial aim of developing this methodology was to see whether stable errors in L2 syntax, resulting from incorrect L1-L2 transfer (see e.g. White, 2003; Hopp, 2010; Antón-Méndez, 2011) between German and Dutch gendered articles, could be reliably induced and measured (Brandt, Schriefers & Lemhöfer, 2021, submitted). This previous work can be summarised as succesfully testing whether the newly developed method is appropriate for investigating the incidental learning of a syntactic feature (grammatical gender) that is coded on the level of lexical items (nouns). This was tested in a German (L1) speaking sample implicitly learning the correct Dutch (L2) articles for a set of given nouns, with the ultimate aim of seeing whether this German (L1) speaking sample would pick up on the gender similarities between German and Dutch, and whether native speakers of German would show incidental learning of Dutch grammatical gender (Brandt, Schriefers & Lemhöfer, 2021, submitted).

The present study aims to extend the findings of this previous work by switching to a Dutch (L1) speaking population. To see how this switch in target population affects the L1-L2 transfer effect as described previously by Brandt and colleagues (2021, submitted), and why this switch is relevant in the first place, it is important to first understand the nature of the German and Dutch article systems, and to understand the relevant differences in German and Dutch populations.

The compatibility of German and Dutch article systems

First, the German and Dutch article systems must be discussed, as otherwise it will be impossible to understand where the L1-L2 transfer effect comes from in the first place, let alone how it might change when the populations are switched around. German and Dutch are closely related languages, and share many words that are close in form and meaning, called *cognates* (Lemhöfer et al., 2008, 2010). The article systems in German and Dutch are also similar, but not quite identical. In German, there are three singular definite articles (der_{masc}, die_{fem}, das_{neu}). In Dutch, there are only two singular definite articles (de_{com}, het_{neu}). Cognate nouns usually have the same gender in German and Dutch. So, whenever a cognate noun corresponds to a masculine or feminine article in German (*die*, *der*), usually the Dutch translation uses the common gender article (*de*). Similarly, whenever a cognate is neuter in German (*das*), it usually corresponds to a neuter translation in Dutch (*het*). A typical example is the German *der Mund* ('the mouth'), which corresponds to the Dutch *de mond*. In this example, two cognates with a similar form and identical meaning indeed have the same gender. Cognates that share the gender across German and Dutch will hereafter be referred to as *compatible cognates* (Lemhöfer et al., 2008, 2010, 2014; Brandt, Schriefers & Lemhöfer, 2021, submitted).

Since this syntactic similarity holds for the majority of cognates, native speakers of German could potentially use the knowledge of their own native article system to accurately predict the correct article in Dutch. Indeed, this is exactly the behaviour that was found in Lemhöfer and colleagues' previous studies (2008, 2010, 2014; Brandt, Schriefers & Lemhöfer, 2021, submitted), and can be accurately described as a generally useful *L1-L2 transfer* strategy (see e.g. White, 2003; Hopp, 2010; Antón-Méndez, 2011). However, in some cases, this strategy fails. Some cognate words with a similar form and identical meaning have a different gender in German and Dutch. These words are referred to as *incompatible cognates* (Lemhöfer et al., 2008, 2010, 2014; Brandt, Schriefers & Lemhöfer, 2021, submitted). For the incompatible cognates, whenever the German noun has masculine or feminine gender (*der* or *die*), it has neuter gender in Dutch (*het*). And whenever an incompatible cognate is neuter in German (*das*), it has common gender in Dutch (*de*). An example is *das Radio* (*the radio* in English) which corresponds to the Dutch *de radio*. Even though the form and meaning are identical in these cognates, the gender of the articles is different. These incompatible cognates are the minority of all German-Dutch cognate pairs (Lemhöfer et al., 2008).

Because of this systematic similarity in gender systems, and the existence of exceptions, the articles of German-Dutch cognates were a perfect target to test out a newly developed method of incidental learning (Lemhöfer et al., 2008, 2010, 2014; Brandt, Schriefers & Lemhöfer, 2021). These studies showed that, while compatible cognates reliably received a correct article in Dutch from native German participants learning Dutch as L2 (with error rates of ~8%), the incompatible nouns induced a high proportion of incorrect articles in Dutch, with error rates as high as 70% (Lemhöfer et al., 2010).

The L1-L2 transfer strategy that these participants consistently pursued was often referred to by the authors as *'mapping*' their L1 grammatical gender knowledge onto the L2 nouns presented in the study. As *mapping* and *mapping responses* are easier terms to work with, they will be used as a shorthand throughout this paper to refer to responses that are in agreement with an L1-L2 transfer strategy. So, whenever a participant's response is guided by their L1 knowledge, it is regarded as a *mapping response*, regardless of whether this response results in an error. Thanks to the reliability of the mapping responses of the participants, the resulting errors were exploited in the design of the ('virtual') confederate-based *incidental learning paradigm* (or '*dialogue game*') developed by Brandt and colleagues (2021). Still, since the effect is so reliable and the paradigm indeed works as intended in a German (L1) sample, why is it relevant for the present study to extend the findings in a Dutch (L1) sample? To understand the importance of this switch in target population, the relevant differences between the Germans and the Dutch need to be examined.

The differences in German and Dutch populations

The first and foremost reason to choose Dutch (L1) participants in this new study, is to extend the proof of concept presented in the paper of Brandt and colleagues (2021): if their results were replicated in a different population, the methods and results can more easily be generalised.

In addition to proving the methodological concept, it is also theoretically interesting to switch these populations around, as the effect of L1-L2 transfer in German (L1) speakers might be qualitatively different from Dutch (L1) speakers. For example, when learning Dutch as an L2, there are only two articles to be distinguished: *de* and *het*. In German as an L2 however, there are three articles: *der*, *die* and *das*. In addition, as already touched upon above, when a Dutch (L1) speaker learns German (L2), the de_{com} category has to be split up into two distinct articles: der_{mas} and die_{fem}.

Although there is – to the knowledge of the author – no theory predicting whether the number of different determiners in a language affects (syntactic) learning, it still intuitively feels more difficult to learn three categories from two than the other way around. The most well known model illustrating this idea is the Perceptual Adaptation Model (Best, 1994; for adult L2 learning see e.g. Best & Tyler, 2007; Tyler, 2019), which typically explains extrapolating two related phonemic categories from one broader category to be more difficult than condensing two categories into one[2]. In essence, this seems similar to the problem Dutch (L1) speakers have to face when converting their determiner system to the German (L2) equivalent. While German (L1) speakers can convert their gendered determiners der_{mas} and die_{fem} into one single category de_{com} when learning Dutch (L2), Dutch (L1) speakers have to split their single category de_{com} into the two categories der_{mas} and die_{fem} when learning German (L2). Although the comparison with the PAM(-L2)'s phonetic category predictions is not perfect, and the PAM(-L2) model has not ever been translated to the syntactic domain, it still serves as an analogy as to why there might be a qualitative difference between Dutch (L1) speakers learning German articles and German (L1) speakers learning Dutch articles.

Given that the German article system is more complex, it might seem like the Dutch (L1) participants are fighting an uphill battle. However, they do have a potential leg up over their German (L1) counterparts due to their schooling, possibly helping them acquire the mapping more easily then their German counterparts. All children in Dutch high school are required to take classes in German for 1 to 6 years, depending on school and education level. One year of German schooling, however, is a mandatory minimum. This is not the case for Dutch in German high schools, where Dutch is only an optional course, if it is at all present in the curriculum. This means that all Dutch (L1) speakers schooled in the Netherlands will have had some formal German

2 Take for example rhotic consonants (see Ladefoged & Maddieson, 1998), which are all contained in the phoneme /r/ in English ('***great***', '***rock***'), while in Spanish there are two distinct phonemic categories: the voiced alveolar trill /r/ (as in '*car**ro***'; English: '*car*') and the alveolar tap /ɾ/ (as in '*car**o***'; English: '*expensive*'). The PAM-L2 predicts that an English (L1) speaker will have more trouble learning the Spanish (L2) rhotics than a Spanish (L1) speaker learning English (L2) rhotics (e.g. Best, 1994; Tyler, 2007; Tyler, 2019).

schooling. The German (L1) speakers of previous research, in contrast, did not receive such mandatory high school classes of Dutch (Lemhöfer et al., 2008, 2010, 2014; Brandt, Schriefers & Lemhöfer, 2021, submitted). This means Dutch (L1) speakers may have had enough exposure to the German article system and thus make use of the mapping prior to the experiment, even if they were not currently immersed in a German context. The German (L1) speakers of previous research mainly got their Dutch knowledge from being immersed in a Dutch context, by studying abroad in Nijmegen (Lemhöfer et al., 2008, 2010, 2014; Brandt, Schriefers & Lemhöfer, 2021, submitted).

So, there is reason to assume that adult Dutch L1 speakers will not acquire the mapping in the same way as adult German L1 speakers. This could either be the case because of the higher complexity of the German article system, or due to the prior experience all Dutch L1 speakers already receive in high school. This, then, explains the relevance of extending the findings for German (L1) speakers learning Dutch (L2) of Brandt and colleagues (2021) to Dutch (L1) speakers learning German (L2). Due to the change in target population, some of the methodology needed to be changed. This was an opportunity not only to extend the findings, but also to extend the methods to gain more interpretable data, and to see whether some of the open questions raised in the previous studies could be answered (Brandt, Schriefers & Lemhöfer, 2021, submitted), In the next paragraph, the necessary methodological changes are discussed, as well as the novel exploratory analyses that is meant to further test the versatility of the incidental learning paradigm.

The present study

The main aim of the present study is to replicate the methods of Brandt and colleagues (2021), to see whether L1-L2 transfer occurs in the same way in Dutch (L1) speakers learning German (L2) as was formerly found in German (L1) speakers learning Dutch (L2). Since no previous work has been done in Dutch (L1) speakers, it was unclear to what extent the L1-L2 transfer had already occurred, and how much this influenced their behaviour. To measure whether participants already showed the expected mapping behaviour as found previously in experienced German (L1) speakers (Lemhöfer et al., 2008, 2010, 2014; Brandt, Schriefers & Lemhöfer, 2021), the article usage in compatible cognates can be used as a baseline.

This signifies an important methodological change in the present study. In the previous experiment (Brandt, Schriefers & Lemhöfer, 2021), compatible cognates were used only as filler counterpart to the critical incompatible cognates, where an erroneous response was expected to occur. Although in a follow-up study, these compatible cognate ‘fillers’ were partially analysed (Brandt, Schriefers & Lemhöfer, submitted), this was a post hoc analysis and the compatible cognates were not analysed in the same way as the incompatible cognates (for details see below).

To be able to give an accurate indication of the strength of *'usage of mapping'*, the present participants' experience needed to be measured in an objective way. Incompatible cognate responses are difficult to interpret in this respect, as there are two opposing forces potentially active: participants that are more experienced are expected to show more mapping responses, but they might also know more nouns that do not follow the mapping strategy. It would thus remain unclear whether errors are due to experience or inexperience with the L2. In contrast, since there is no difference in correct article responses and mapping responses[3] in compatible cognates, the pattern of article usage for compatible cognates is simple and straightforward to interpret as a reflection of already established knowledge in L2.

By taking a look at the general article usage pattern in these compatible cognates for each participant, an assessment could be made per participant whether they had already started mapping their (Dutch) L1 grammatical gender knowledge onto their (German) L2 nouns. This, then, could be used as an indication of 'experience' with the mapping, akin to the experienced German participants in previous studies, as it would be an indication that the participants already possessed (unconscious) knowledge of the beneficial Dutch-German L1-L2 transfer. To be able to perform this analysis properly, compatible cognates needed to be recorded in the same way as incompatible cognates. And, as a secondary benefit, adding the compatible cognates to the analysis simply gives a more complete view of the entire article usage pattern of every participant, making a direct comparison between compatible and incompatible nouns possible.

Other than methodological changes, one additional analysis was performed in the present study. This analysis followed from a surprising finding in Brandt and colleagues (submitted), where relatively naive, or inexperienced, participants seemed to acquire the mapping throughout the experiment. At the start of the experiment they did not show any evidence of mapping their (German) L1 grammatical gender knowledge onto the corresponding (Dutch) L2 nouns. But in the posttest, performed at the end of the experiment, they did show a bias for mapping responses. So, over the time course of the experiment these inexperienced participants learned to use the mapping simply by being exposed to the stimuli. This was unexpected as this was a form of learning that was not seen in experienced participants (Brandt, Lemhöfer & Brandt, 2021).

The only contrast tested in the study of inexperienced participants (Brandt et al., submitted) was the baseline '*usage of mapping*' at the start of the experiment versus the '*usage of mapping*' at the end of the experiment. Throughout the experiment, participants received native input to learn the correct articles for all critical words (details will follow below), but in this contrast of '*usage of*

3 Except when there was an error by using *die* instead of *der* or vice versa, which was resolved by counting these articles as the same response type ('*common gender*', as the Dutch de_{com}) for the mapping response analyses. Although these responses could be incorrect, they still followed the same mapping strategy, so were in line with the L1-L2 transfer strategy. The choice for this method of analysis will be further elaborated upon in the methods section.

mapping', only productions without input were taken into account. This means that the effect of an increase in mapping responses in the posttest could only have been caused by mere exposure to the stimuli (Brandt et al., submitted).

Since Brandt and colleagues (submitted) only contrasted the start of the experiment with the end of the experiment, while there was no continuous measurement or analysis, it is unknown what this learning looked like over time. However, such an analysis could be integrated within the general experimental framework relatively easily. Both the former and the current experiment were performed in four different blocks, with randomised stimuli. Other than the exact stimuli, the blocks were identical to one another. After these blocks, the posttest was performed. So, in the current experiment, the blocks could be used to provide a more fine-grained time course analysis of this learning effect. Instead of looking at only the start of the experiment and the posttest, the current experiment will contrast each block with the others, so the time course of the learning effect (expected to be an increase in mapping responses) will become clear. This will show whether the tendency towards mapping responses builds up gradually over the course of the experiment, or whether the effect looks entirely different. As this is the first time the time course of this learning effect is analysed, what the exact (temporal) form looks like is an open question.

As no previous work has been performed on Dutch (L1) speaking participants, for all analyses (time course or otherwise) Dutch participants might show an attenuated effect or show no effect at all, as the German article system could prove too complex for the L1-L2 transfer effect to occur in the same way as with German (L1) participants (e.g. Best & Tyler, 2007; Baten, 2013; Tyler, 2019). However, based on the strong tendency of the German (L1) participants to adhere to the mapping strategy, and the resulting stable error pattern for incompatible cognates (Lemhöfer et al., 2008, 2010, 2014; Brandt, Schriefers & Lemhöfer, 2021), as well as the definite learning effect shown by German participants after only a single input per word (Brandt, Schriefers & Lemhöfer, 2021), a replication of these results iss expected: a positive effect of input on correct article production for incompatible cognates, and no change in behaviour when there was no input. Whether Dutch (L1) speakers would prove to be more or less proficient in their German L2 due to their previous exposure compared to German (L1) speakers who were 'immersed' in their Dutch L2 environment at the time of testing remains purely explorative, however, and the present study will serve as a guide in Dutch participants' behaviour for following studies. As the time course and compatible cognate analyses have not yet been performed either, these also serve as explorations into the possibilities of the Brandt and colleagues' (2021) incidental learning paradigm.

Before moving on to the methods section, the paradigm as developed by Brandt and colleagues (2021, submitted) needs to be understood. As this largely an effort in replicating the previous studies, it uses the same paradigm. Since the paradigm uses many new terms as a shorthand for its

ideas, and due to its structure, these will be addressed first, before moving on to the details of the methods section. The paradigm itself will be referred to as the '*incidental learning paradigm*' or *'(L2) dialogue game'*.

The incidental learning paradigm (or L2 Dialogue Game)

The *'incidental learning paradigm'* (or '*L2 Dialogue Game*') paradigm was designed by Brandt and colleagues (2021) to assess whether participants could learn articles by a native speaker's input without being aware they were learning anything. The purpose was to simulate a naturalistic, immersive context wherein the participant would 'incidentally' learn the correct grammatical gender of nouns by virtue of the input of a native German speaker. To be able to simulate learning from a natural conversation, the explicit nature of direct feedback needed to be masked, as in a natural conversation the aim usually is not to explicitly learn a language. To achieve this, the participants were told they would perform a memory task in German, so that memory could be observed while participants would use a second language. This way, the actual aim of learning was masked by an unrelated memory task.

With the purpose behind the paradigm clear, now we turn to the implementation in the present experiment. For each trial, a participant had to learn a pair of nouns to later use in a sentence. Both the nouns and the sentences were produced in German, the target language of the experiment. Each pair of nouns was learned through presentation of two pictures, which were presented side by side without any accompanying text. When all pairs were learned through this exposure phase, the sentence-production part of the experiment began. For every sentence-production trial, participants were given four pictures. One on the left-hand side of the screen, and three on the right-hand side of the screen. Participants had to combine the left-hand side picture with one of the right-hand side pictures, corresponding to the pairs they learned previously in the exposure phase. The participants selected their right side picture of choice through combining the pairs of nouns in the form of a sentence. Each time they had to produce a sentence, the participants were asked to use a fixed structure. Only the nouns and their appropriately gendered article would change across trials. An example of a sentence that the participants were asked to produce is, in German: '***Das Boot*** *und* ***die Blume*** *gehören zusammen' ('**The boat** and **the flower** belong together'* in English). These sentence-production trials of participants are referred to as '*production trials*' throughout this paper.

In this way, participants were indeed tested on their memory skills. Although memory was not part of the analysis in the present paper, the correct memory responses were recorded and remain available upon request. In addition, the participants were forced to use self-generated articles for the appropriate nouns due to having to produce their own sentences, making the paradigm more natural as a result. Each noun pair was presented to the participant a total of three times throughout the

experiment. This means that the same pair of nouns was repeated in a production trial a total of three times, so that three measurements of the same noun pair were produced by each participants.

Crucially, the participants were not alone in producing their sentences. After each production trial, a German (L1) speaker (in the form of a recorded voice) produced a sentence while the participant had to judge whether the voice had correctly identified the right-hand side picture on the screen, in accordance with the correct noun pair. This voice of the German (L1) speaker was the '*(virtual) dialogue partner*' of the participant, and will be referred to as such throughout the rest of the paper. The virtual dialogue partner sometimes made an error in the pairing, but always provided the correct article for any given noun. These trials, in which the participants had to judge the correctness of the memory of the virtual dialogue partner, will be referred to as '*listening trials*'.

For half of all noun pairs, the virtual dialogue partner would provide completely correct input in the form of a listening trial. This input was provided in between the 2nd and 3rd time that participants produced a sentence with the same noun pair. In this way, the first two production moments were always input-free and served as a baseline for the participant's natural article production. After input, it was possible to see whether the participant picked up the correct article (if they were erroneous in their first two productions) in their 3rd production of the same sentence. As a way to compare 'input' to 'no input', half of the noun pairs did not receive any input from the virtual dialogue partner for a given participant. Instead, they were provided with a '*filler trial'* (see below) that gave no input on the articles of a critical noun pair.

One important difference from the present paper's version of the paradigm compared to the Brandt and colleagues' (2021) experiment was the way the compatible cognates were treated. As the virtual dialogue partner would produce a sentence every other trial, and only gave correct input once for every three production trials (and even then, only in the input condition), many trials could not consist of the actual experimental noun pairs. To fix this problem, '*filler trials*' were constructed to fill in these gaps. Brandt and colleagues used both compatible cognate pairs and noncognate pairs for this purpose. Noncognates were defined as nouns with the same meaning that have no similarity in form between German and Dutch, such as '*(der) Bahnhoff*' in German ('*(the) railway station*' in English) and '(*het) treinstation*' in Dutch. The present paper uses instead only noncognate pairs as filler trials. Each compatible noun was coupled in pairing to an incompatible noun, and they were treated in the same way throughout the entire experiment. Not only did this provide an accurate baseline for experience in German article production, but it also resulted in a more complete picture of article usage in German (L2) by Dutch (L1) participants. By extending the analyses to include compatible cognates, this change was another exercise in extending the existing methodology devised by Brandt and colleagues (2021).

Methods

Pilot

Participants

Twenty-four native Dutch (L1) speaking participants were selected for a pilot study in which German translations of the stimuli used in Brandt, Schriefers and Lemhöfer (2021; see below) were tested to see if they were suitable for Dutch (L1) participants. These participants were recruited through a convenience sample and received no compensation. They were selected on the basis of being between 18 and 30 years old, with a higher education: i.e. similar to the most likely participants for the main experiment. One additional constraint was that all participants had chosen German as a course in the higher classes of high school, to ensure they had some prior knowledge of German. Of the twenty-four participants, one's recorded data were unusable and therefore could not be included, while three others gave only partial responses but could still be included. The remaining 23 participants were on average 21.4 years old (*SD* = 2.39, 16 identified as female).

Experiment

Participants

For the main experiment, a new sample of 62 Dutch native speakers was recruited through a combination of convenience sampling and internal recruitment in the Radboud University Nijmegen[4] via posters and the internal SONA-system. All participants were either rewarded with participant credits or vouchers with a value of €20. They had normal hearing, normal or corrected-to-normal vision and none were diagnosed with dyslexia or any reading-related problems. Of the recruited participants, the first three were used for a technical pilot: their data were not analysed. Of the remaining 59 participants, two more had to be excluded due to errors during the experiment, resulting in a presentation of the incorrect experimental list or failure to complete the entire experiment. A third participant was excluded from analysis due to exceeding the a priori threshold of maximally 25% missing trials throughout the experiment.

Of the remaining 56 participants, 40 identified as female and the rest as male, they had a mean age of 26.52 years of age (*SD* = 9.51, ranging from 18 to 65 years, one participant did not fill in their age). To ensure all participants had had enough exposure to German, they were selected on the basis that they had chosen German as a course in their final year of Dutch high school. They were also selected to have no bilingual German background. No further constraints were placed on the participants' selection.

All participants reported knowing at least one foreign language, while on average they reported knowing 2.05 foreign languages. One participant did report knowing a foreign language, but did not

4 Radboud University Nijmegen, The Netherlands.

report which one. The three most reported foreign languages were English (all 55), French (35) and Spanish (9), with one participant reporting sign language as a spoken foreign language. All participants lived in the Netherlands at the time of testing[5], while three reported to have lived in Germany for at least a month at some point in their life, with one participant reporting a plan to emigrate to Germany. Four participants, due to technical difficulties, could not participate in the posttest, so only 52 participants were included in the posttest. In the delayed posttest, even less participants were included because only 18 participants responded to an email prompt sent 6 weeks after the original experiment. Of these participants, 13 identified as female and their mean age was 26.11 years of age (*SD* = 8.897) ranging from 18 to 58 years of age. No additional rewards were provided for participating in the delayed posttest.

An 'immersion score' was calculated as in Brandt, Lemhöfer and Schriefers (2021, see also Mickan & Lemhöfer, 2020). This score consisted of five questions, for which each 'yes' would add one to the score (range 0-5). The questions were whether participants lived in Germany, had German roommates, German friends, a job in Germany, and/or a relationship with a German native. No participant had a score higher than 3. Participants were also asked to self-rate their overall proficiency in German on a 7-point Likert scale and on four subscales regarding writing, listening, reading and speaking. All subscales consisted of one question, except listening which consisted of two questions, all with a 5-point Likert scale. Finally, the LexTALE (Lemhöfer & Broersma, 2012) task was run in German for an indication of the participants' vocabulary size in German.

A summary of these measures, as well as the gender and age information of all participants and their subgroups is provided in Table 1. The subgroups refer to a division in participants that appeared to 'follow the mapping' (or: showed evidence of L1-L2 transfer) at the start of the experiment, and those who did not show any evidence of 'following the mapping' at the start of the experiment. The former are referred to the 'Follow subgroup' (N = 38) while the latter are referred to as the 'NoFollow subgroup' (N = 18). For details on defining these groups, please see the Results section.

5 See Appendix B for the exposure to regional languages of the Netherlands reported by the participants.

Table 1. *Means and Standard Deviations of participant characteristics regarding Age, Linguistic Background (questionnaire) and L2 vocabulary (LexTALE) for all participants (N = 56) and for the Follow (N = 38) and NoFollow (N = 18) subgroups.*

	All participants	Follow subgroup	NoFollow subgroup
	M (*SD*)	*M* (*SD*)	*M* (*SD*)
Age	26.52 (9.51)	27.34 (11.25)	24.83 (3.88)
Immersion score	0.40 (0.655)	0.49 (0.692)	0.22 (0.548)
L2 vocabulary size (LexTALE)	62.25 (7.417)	63.71 (7.742)	59 (5.545)
Average self-rated proficiency:	3.19 (0.90)	3.39 (0.92)	2.71 (0.67)
Self-rated proficiency: speaking	2.64 (1.09)	2.87 (1.17)	2.17 (0.71)
Self-rated proficiency: listening	3.55 (0.96)	3.74 (0.98)	3.17 (0.80)
Self-rated proficiency: writing	2.45 (1.10)	2.71 (1.11)	1.89 (0.83)
Self-rated proficiency: reading	3.66 (0.98)	3.89 (0.92)	3.17 (0.92)

Note. LexTALE scores are given in a mean score of the standard score used in the LexTALE software (Lemhöfer & Broersma, 2012).

Materials

Stimuli

In total, 142 Dutch nouns were used that had already been selected in earlier studies (Lemhöfer, Spalek & Schriefers, 2008; Lemhöfer, Schriefers & Hanique, 2010; Lemhöfer, Schriefers & Indefrey, 2014; Brandt, Schriefers & Lemhöfer, 2021, submitted). These nouns were high-frequency Dutch words (CELEX database; Baayen, Piepenbrock & Gulikers, 1995) and corresponded as much as possible with the nouns used in Brandt, Schriefers and Lemhöfer (2021) to ensure comparability of the studies. The Dutch nouns were translated into German and were checked by two native German speakers. In the pilot study, through a Qualtrics questionnaire, it was assessed how familiar these German nouns were to our target population of Dutch native speakers.

Of the 135 nouns, 48 were German-Dutch cognates with incompatible grammatical gender across the two languages (16 *der*-, 16 *die*- and 16 *das*-words) like 'das_{neu} *Boot'* (German; 'de_{com} *boot'* (Dutch); '*the boat'* (English)) or der_{masc} *Park* (German; 'het_{neu} *park'* (Dutch); '*the park'* (English)). These were dubbed the *incompatible cognates*. Additionally, 48 German-Dutch cognates were selected with compatible grammatical gender across the two languages, like 'das_{neu} *Gewehr*' (German; 'het_{neu} *geweer*' (Dutch); '*the rifle*' (English)) or die_{fem} *Kirche* (German; 'de_{com} *kerk'* (Dutch); '*the church*' (English)). These were in contrast called the c*ompatible cognates*.

Throughout most of the experiment, these target nouns were presented to the participants in pairs. Each pair consisted of one compatible and one incompatible noun, making 48 cognate pairs in total. The incompatible and compatible cognates were paired in the same way as in Brandt and colleagues (2021), with the only differences being that compatible words also appeared in the '*left-side*' position – on the left of the screen and first in the sentence – while incompatible words appeared in the '*right-side*' position – on the right of the screen and second in the sentence. This change in presentation allowed for both compatible and incompatible categories to be analysed in the present study. This further meant that all cognates used in the present study were '*critical items*' (see e.g. Brandt et al., 2021), while what has formerly been described as '*accompanying items*' ceased to be relevant for the present study.

Still, there remained a need for filler trials, for which 39 non-cognate words (extracted from the same source) were added. Of these fillers, 16 were *incompatible noncognates* (9 *der*-, 5 *die*-, and 8 *das*-words) and 23 were *compatible noncognates* (8 *der*-, 7 *die*-, and 8 *das*-words). Since these fillers were noncognate words the compatibility was unlikely to influence the participants' mapping behaviour, as there was little to no similarity between the Dutch and German nouns. For this reason, the noncognate filler pairs were paired strictly randomly, without taking gender-compatibility into account. Participants never produced sentences with these filler words. Rather, these filler-trials

were only used in '*listening trials'* and were added to mask for the participants that the experiment exclusively concerned cognates. For a complete list of the stimuli, see Appendix B.

For all words, fillers and critical, a corresponding picture was selected from the same database as used in Brandt, Schriefers and Lemhöfer (2021). The pictures were squares of 425 x 425 pixels with a resolution of 72 dpi. In total, 96 target items and 39 filler items were presented to each participant, each with a unique picture and all in unchanging pairs, totaling 135 items presented as 68 unique item pairs. One filler item was accidentally used for two item pairs, but it was deemed unlikely that the participants' behaviour regarding the articles was influenced by this error.

Sound recording of auditory stimuli

For the memory game (see below), the sentences produced by the "interaction partner" were recorded by a female 29 year old native German speaker with no apparent regional accent, in a sound recording laboratory at the Radboud University using the software Audacity® (Version 2.2.1.; Audacity Team, 2017). The recordings were spliced into separate audio-files for each sentence and were normalised to a similar volume

Setup & Software

The participants were seated in a sound-proof booth with a monitor, keyboard and microphone placed on a desk in front of them. Participants heard the interlocutor's sentences via headphones at a constant volume across participants. The participant responded verbally and through two response button presses on the keyboard. The familiarization phase was programmed in Delphi® software (Version XE5 Update 2; Embarcadero Technologies Inc. Retrieved from: https://www.embarcadero.com/products/delphi), while the 'pair learning task' and main experiment were programmed in Python (Version 3.6; Van Rossum & Drake, 2009).

A second computer was situated outside of the sound-proofed booth from which the experimenter ran the experiment. A baby monitor was used to hear participant's responses online, but the responses were definitively coded offline using the audio recordings, which were done using the recording software of Audacity® (Version 2.2.1., Audacity Team, 2017. Retrieved from: https://audacityteam.org/) and the spreadsheet software of LibreOffice Calc (Version 5.4.0.1., The Document Foundation, 2017; Retrieved from: https://www.documentfoundation.org/).

All parts of the experiment were performed in the sound-proofed booth, except for the posttest and LexTALE (Lemhöfer & Broersma, 2012) tasks, which were performed at the experimenter's computer using Presentation® software (Version 18.0, Neurobehavioral Systems, Inc. Retrieved from: https://neurobs.com/). All questionnaires were administered in printed form. The delayed posttest was administered via Qualtrics software (Qualtrics, 2017; Retrieved from:

https://qualtrics.com/), for which participants received an invitation on their email-adress containing a private link six weeks after the formal experiment was concluded.

Procedure

Ethical Approval

As this study was planned and designed in the context of a PhD project led by Annika Brandt, it fell under the ethical approval of the Ethics Committee Social Science of the Radboud University already granted to the project. Participants were given a standardised informed consent form for behavioural studies as issued by the Donders Institute (Donders Institute for Brain, Cognition and Behaviour, Radboud University, the Netherlands) which had to be signed prior to the start of the experiment.

General Procedure

The set-up of this study was similar to Brandt, Schriefers and Lemhöfer's (2021) study, specifically their 2nd experiment. In the present experiment there were three consecutive phases: a familiarisation phase, a learning phase, and a retrieval phase, with the latter phase being the actual '*L2 dialogue game*'. This last phase has already been outlined in the introduction, in this section all phases are discussed in detail. The experiment consisted of 4 blocks, each consisting of 12 unique critical item pairs and 5 unique filler item pairs (one of the four blocks comprised 6 unique filler item pairs). Each of the four experimental blocks consisted of three phases (familiarisation, learning and retrieval).

Familiarisation

Each of the four blocks started with a familiarisation phase to ensure all critical nouns could adequately be named. This was necessary so that participants would recognise and use the expected names for the stimuli, as the stimuli were specifically selected to have a correct balance of articles. This familiarisation served to encourage the participants to use the intended nouns and not synonyms, or other words altogether.

In this phase, all of the block's items (cognates and fillers) were shown as a picture in the middle of a white screen, and participants were instructed to name each item in German. If an item were unknown, the participant was told to guess. After a response of the participant, the experimenter pressed a button to reveal the target noun (without an article) in black text under the picture. The participant was asked to read aloud this noun – regardless of their initial response – after which the experimenter continued to the next item.

The experimenter noted for each cognate whether it was named correctly, or at least whether it was easily interpretable as the target noun. If a given participant's initial responses were not (interpretable as) the target for more than 75% of all cognates in a block, the familiarization was repeated. Since the noncognates (fillers) would not be produced by the participants in the critical trials, these were ignored in the familiarisation phase's coding. The large majority of participants had to repeat the familiarisation for at least one of the blocks. After maximally one repetition per familiarisation, all participants answered with the target response at least 75% of the time.

The 24 cognates per block were equally distributed over gender class: 8 were '*der*-words', 8 were '*die*-words' and 8 were '*das*-words'.

Learning of Picture Pairings

After the familiarisation phase for a block, the participants were shown all pairs of pictures of the respective block, side by side, on a grey background. They were told that the objective of this phase was to remember that the pictures of a pair 'belong together', and that this would be tested in the next phase of the experiment. This stressed the idea that the experiment concerned memory retrieval, not article learning, and served to give the participants a task to perform for the rest of the experiment. Each picture-pair was shown for 2 seconds before continuing to the next pair. Each pair consisted of a gender-incompatible and a gender-compatible object, and always was either a cognate-cognate (critical) pair or a noncognate-noncognate (filler) pair. An example trial of the Pair Learning task is shown in Figure 1.

In total, 17 or 18 pairs were presented per block in this phase. Of those pairs, 12 were cognate-pairs, and 5 or 6 were noncognate filler-pairs. Of the 12 cognate-pairs, which were always a combination of one compatible and one incompatible cognate, 6 would have the compatible cognate on the left side of the screen and 6 would have the incompatible cognate on the left side of the screen.

Figure 1. An example of a learning phase trial as used in the experiment. This figure was originally used in Brandt, Schriefers and Lemhöfer (2021). Due to copyright, the pictures were not identical to the ones used in the experiment.

Main experiment: The L2 Dialogue Game (Retrieval Phase)

After the pair learning phase, the participants participated in the main '*(L2) dialogue game*' part of the experiment. This 'game' was first employed in a previous study by Brandt, Lemhöfer and Schriefers (2021). As previously mentioned in the introduction, there were two types of trials in the dialogue game (or *incidental learning task*): '*production trials*', where the participant themselves produced a sentence containing two target nouns and their articles, and '*listening trials*' where the '*(virtual) dialogue partner*' produced a sentence.

In the production trials, the picture corresponding to one of the two target nouns was presented on the left side of the screen, and three other pictures were presented vertically stacked on the right side of the screen (see Figure 2 below). One of the three right-side objects had been paired with the left-side object during the learning phase, while the other two right-side pictures served as distractors. The participant's task was to identify the correct picture from the three right-side alternatives. To indicate their choice, the participants were asked to use a standard sentence like *'**Das Boot** und **die Blume** gehören zusammen'* (*'**The boat** and **the flower** belong together'* in English). In this sentence, '***Das Boot**'* indicated the left-side item, and '***die Blume***' the right-side alternative. The sentence format was chosen such that both nouns in the sentence occurred in nominative case in order to avoid any additional complications of the German case marking system[6]. By having all articles in the German nominative case ($der_{masc}/die_{fem}/das_{neu}$), the complications of case inflections could be wholly avoided.

6 While Dutch does not have any case-system altering the form of articles, German has an extensive case-system that demands the articles change their form depending on what function a noun has in a sentence (Baten, 2013). The nominative case is the case closest to the Dutch sentence structure, which is why the current sentence form was used.

The remaining half of the trials were '*listening trials*', in which a '*virtual dialogue partner*' (pre-recorded German native voice) produced a sentence that was either an *input trial* or a *non-cognate filler trial*. The input trials were identical to the sentence structure that the participant produced ('*Das Boot und die Blume gehören zusammen'*). In fact, these input-trials would contain the same noun pairings as those in the sentences that the participant produced earlier. In all listening trials, the participants were asked to judge whether the dialogue partner made a mistake in combining the two paired nouns (e.g. chose an incorrect alternative). They made their choice by pressing one of two buttons (labeled as correct vs. incorrect). To make this task engaging, the dialogue partner would provide an incorrect alternative in 25% of the filler trials. The errors only concerned incorrect choice for a right-side noun, never an incorrect article.

For all trials, the visual set-up was identical. The right-side pictures were randomised in terms of location (*upper*, *middle* or *lower* position). Of the three right-side pictures, one would always be the correct pairing, one would be incorrect but consistent across repetitions of that noun-pair, and one would be incorrect and inconsistent across repetitions. Each pair would be repeated three times by the participant during production trials (at *production moment 1*, *2*, and *3*) and half of all pairs received one *input trial* during one of the listening trials, produced by the virtual dialogue partner (see the *Trial Scheme* below). Figure 2 shows an example of a typical trial set-up for both the production and the listening trials.

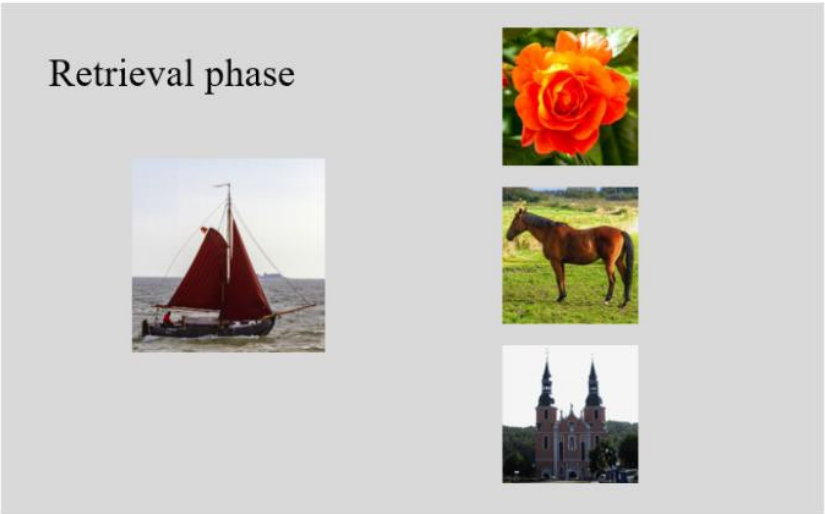

Figure 2. An example of a retrieval phase (dialogue game) trial as used in the experiment. Production and listening trials were visually indistinguishable. This figure is was originally used in Brandt, Schriefers and Lemhöfer (2021). Due to copyright, the pictures are not identical to the ones used in the experiment.

Trial Scheme

The dialogue game used a specific ordering of trials, critical to the input-based learning nature of the game. This trial scheme was slightly modified from the former study (Brandt, Schriefers & Lemhöfer (2021), but followed the same general structure.

The previously learned item-pairs followed a predetermined pattern of repetition, where the participant and the virtual dialogue partner (the prerecorded German voice) took turns in producing a sentence. The sentences produced by the participant and by the (virtual) dialogue partner always had the same structure: *'**Das Boot** und **die Blume** gehören zusammen'* (*'**The boat** and **the flower** belong together'* in English). The idea, in principle, was that the participant produced a sentence containing a noun-pair three times in total, and the dialogue partner provided input (in the *Input* condition, see below) by correctly providing this same sentence in between the second and third production moment. This way, the participant could potentially learn from the input after first providing a baseline measure of article usage.

In between the first and second production of a given sentence, five intervening trials occurred. Then, after the second production of the same sentence by the participant, two more intervening trials occurred. The virtual partner then either provided the input used for learning by producing the same sentence with the correct articles (in the *Input* condition), or simply provided a (noncognate) filler trial (in the *No Input* control condition). Half of the critical sentences of each block were provided in the Input condition and the other half in the No Input condition. Then, two more trials followed before the participant produced the sentence for the third and final time. In the Input condition, participants first provided two productions of an article coupled to a noun, which could be grammatically erroneous. Then, correct native input was provided by the dialogue partner, from which the participant could gather the correct article for said noun. In the third production of the same noun, the participant could then show whether they changed their previous (erroneous) article production, showing whether they (implicitly) learned from the native input they received. A visualisation of this trial scheme can be found in Figure 3. The No Input condition simply changed the input provided by the dialogue partner to a filler trial, providing no informative native input for the participant to learn from, and served as a control condition to later contrast with the Input condition.

By having two productions before the experimental manipulation, a calculation could be performed to ascertain whether a participant was already familiar with the German-Dutch mapping before any input was received. Comparing the first two production moments to an expected response pattern gave a measure of stability of the article productions and the tendency of participants to answer in a systematic way (see the Results below). The experimental manipulation

of providing Input/No Input made sure that structural changes in answering patterns over time could be disentangled from direct learning effects from input.

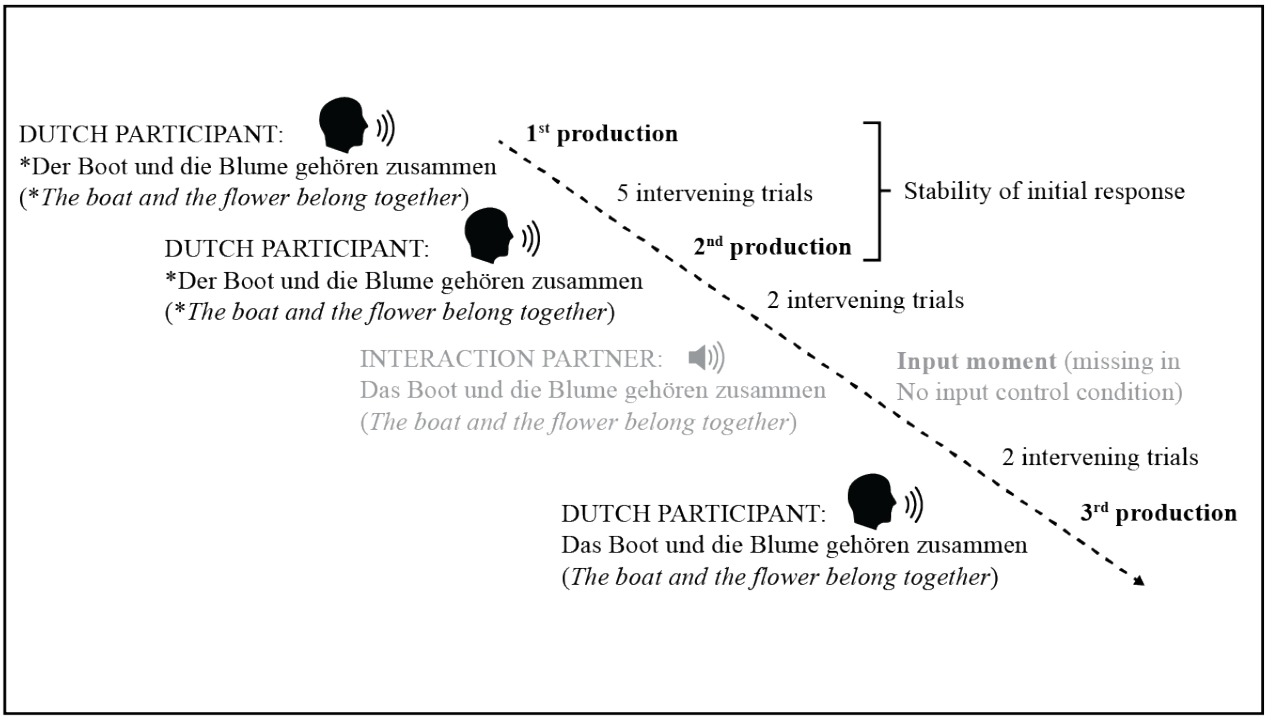

Figure 3. Overview of general structure of each block in the Retrieval Phase. In the example, the participant is shown to correctly learn the article of the incompatible '*Das Boot*' ('*De boot'* in Dutch; '*The boat*' in English), while the compatible *'Die Blume*' ('*De bloem*' in Dutch; '*The flower*' in English) is produced correctly throughout all productions. This figure was adapted from Figure 3 in Brandt, Schriefers and Lemhöfer (2021).

There were a total of 72 trials per block, plus three practice trials at the beginning of each block. These practice trials always consisted of noncognate filler-trials, one of which was a production trial. This was the only instance per block where the participant had to form a sentence with noncognates. Of all 24 target items in a block, 12 received input (after having occurred twice in production trials) and 12 did not.

To counterbalance for order effects and to make sure that each block could be analysed separately, 8 lists were constructed. Across these 8 lists, all target items were used in both Input and No Input conditions. These lists were then randomly presented to the participants, making sure they occurred equally frequently.

After each block, the participant could take a break, after which the next block would start with its familiarisation phase. The familiarisation phases and retrieval (or 'dialogue game') phase were audiotaped separately, and each participant's responses were later coded with regards to articles,

their accuracy (correct or incorrect?), mapping (in agreement with German-Dutch article mapping?) and correct noun usage (was the noun correct and recognisable?). The percentage of correct noun productions varied for the left-side and right-side nouns, as the right-side noun had to be (correctly) chosen from one of three alternatives. The left-side noun was not recognisable as the correct noun 2% of all trials, while the right-side noun was not recognisable as the correct noun 8.4% of all trials[7]. Each participant's responses in the listening trials (did the 'virtual partner' make a mistake?) were recorded by button press, but were not analysed in this study as they were used as a distractor task.

Posttest, Debrief & Delayed Posttest

After the main experiment, the participants were asked about their intuition of the study's aim by the experimenter in a structured-interview format. When asked whether they tried to learn from the input, 45 out of 56 participants reported they had done so, but did not necessarily spontaneously name 'learning the correct articles' as the object of their attention. More often they would report pronunciation or an unspecific 'the words' as their main learning objective. Of these 45 participants that reported they tried to learn something, when prompted whether they tried to learn the articles specifically, 37 reported doing so. A mean of 66.9% (SD = 30.4%) of trials was reported to be attended to by participants when asked to give a percentile number to how often they attended to the input. So, it seems as if the participants were often aware – or paying attention – to the core idea of the experiment, but not necessarily for all trials. It remains unclear, however, how Dutch (L1) participants hold up to the German (L1) participants of Brandt and colleagues (2021), due to the lack of a direct comparison within the same study. Seemingly, Dutch (L1) participants were more actively trying to learn from the native input when confronted with the German stimuli than their German (L1) counterparts, perhaps because of their prior schooling experience. Future research should take care in avoiding leading questions regarding the manipulation, as participants might have been influenced by the prompt of whether they learned the articles specifically and answered 'yes' while actually they might not have payed much attention to them at all. At present, it remains an open question whether (and if so, how) Dutch (L1) and German (L1) participants differ in their approach to L2 (experimental) stimuli in terms of attentiveness.

When this structured interview was completed, a German version of the LexTale (Lemhöfer & Broersma, 2012) task was administered on a computer. After the LexTALE task, the *posttest* was conducted. The posttest was used as a measurement of article usage after the incidental learning

7 The inclusion of the right-side word in the analysis resulted in a more than four times higher error rate, as was described here, but also resulted in double the amount of production moments, as twice as many words per sentence could be analysed. No qualitative differences in article production or any other factor was found between left-side and right-side words.

task, to see if an effect of input – if observed – was stable after the main experiment was completed. Participants were seated in front of a computer and shown all the nouns that were presented in the experiment one by one. Each word was presented in black letters on a white screen. Around each word, the articles 'Der', 'Die' and 'Das' were presented in black letters with a circle around them. The participants were asked to click on the correct article for each word (both critical nouns and fillers). When the posttest was completed, participants were asked to fill in a questionnaire about their language background and language proficiency (see also Brandt, Schriefers and Lemhöfer (2021).

Finally the participants were debriefed, received compensation and were asked if they wanted to participate in the *delayed posttest*, which would be sent to them by mail in the form of a Qualtrics questionnaire, again assessing their knowledge of articles. This delayed posttest largely mirrored the posttest as described above, with exactly the same stimuli being used, only differing in the use of the Qualtrics framework. The mail containing the link to the delayed posttest was sent six weeks after the experiment took place. Only 18 of the 62 original participants responded with usable data to the Delayed Posttest email. In total, the experiment took between 1½ to 2 hours, and the pelayed posttest took another 15 – 30 minutes.

Results

Defining the Mapping subgroups

First, participants were identified who showed evidence of mapping at the start of the experiment, that is in the first block, by calculating a chi-squared test statistic for each participant. For this purpose, the compatible cognates were evaluated, as those words were expected to provide the most robust and reliable evidence for a potential use of article mapping by a participant. First, each article production needed to be quantified in terms of 'mapping'. This was done as follows: if a participant coupled a production of '*die'* or '*der'* to a noun that would have '*de'* as the correct article in Dutch (e.g., *'De*$_{com}$ *mond'* in Dutch/*'Der*$_{masc}$ *Mund'* in German*; 'The mouth'* in English) this was coded as an instance of '*mapping followed'*. These '*mapping followed*' responses were congruent with the expectation that Dutch (L1) people apply their grammatical gender knowledge of Dutch to German nouns: nouns get assigned the same gender in both languages.

Similarly to the '*der'/'die'* example above, whenever a participant responded with '*das'* to a word that would need the article '*het'* in Dutch (e.g., *'Het*$_{neu}$ *orkest'* in Dutch/*'Das*$_{neu}$ *Orchester'* in German*; 'The orchestra'* in English), the response was also counted as an instance of '*mapping followed'*. All other productions of an article were counted as '*mapping not followed*' – meaning the response was incongruent with a Dutch-like article response – except for productions where no article could be clearly identified. These latter cases were not included in the calculations.

A chi-square analysis can show whether an observed distribution is significantly different from a distribution that is expected by a given expected (chance) value (Field, 2013). This expected value, in the current case, was calculated based on the experimental list in block 1, separately for each participant. The chance of randomly following the mapping is not equal across article types. Since the German '*der*$_{masc}$' and '*die*$_{fem}$' both map to the Dutch '*de*$_{com}$', while the German '*das*$_{neu}$' is the only corresponding article to the Dutch '*het*$_{neu}$', there were two response options resulting in "following mapping" for the Dutch article '*de*$_{com}$' and only one for the the Dutch '*het*$_{neu}$'. The chance that a randomly selected article would correspond to the '*de*$_{com}$*-mapping'* was, then, twice as high as the chance that a randomly selected article would correspond to the '*het*$_{neu}$*-mapping'*.

The calculation of the expected values was therefore performed as follows: if a participant would give random articles for any noun (by way of unbiased guessing), all articles ('*der*$_{masc}$', '*die*$_{fem}$', and '*das*$_{neu}$') would have an equal chance of occurring; 1/3rd of the time. For any given compatible '*der*$_{masc}$*/die*$_{fem}$*-noun*', the chance that either '*der'* or '*die'* is used by the participant is 2/3, since there are 3 answer categories in total. This is the chance a *mapping followed* response would occur randomly for any compatible '*der*$_{masc}$*/die*$_{fem}$*-noun*'. For a compatible '*das*$_{neu}$*-noun*', only a '*das'* response counts as *mapping followed*, meaning that there is only a 1/3 chance of following the mapping when answering randomly for any compatible '*das*$_{neu}$*-noun*'.

Given these expectations per word (either 33% or 66% chance of a *mapping followed* response), the exact expectation of how often a *mapping followed* response occurred can be calculated for any given list. For the estimation of whether participants already followed the expected mapping at the start of the experiment, only the first block was considered as there is evidence that participants could learn the mapping through mere exposure to the experiment (Brandt, Schriefers & Lemhöfer, submitted). Only compatible nouns were selected to make interpretation straightforward. So, for every compatible noun of block 1, the expected value (33.3% for *das*$_{neu}$*-nouns* and 66.6% for *der*$_{masc}$*/die*$_{fem}$*-nouns*) was summed together. This sum of values is the expected frequency of *mapping followed* responses that any randomly guessing participant would produce throughout the first block.

The expected value was then compared to the observed value of *mapping followed* responses of each participant. As each noun was produced two times before input occurred within the first block, both the 1st and 2nd productions were used for the expected and observed values. Whenever a participant made an unintelligible article/noun or incorrect noun response (as judged by the coder, the first author of this paper), the response was treated as missing in *both* expected *and* observed values so there would be no bias for *mapping not followed* responses. As all words were cognates, the Dutch version of a word could be so similar to the German version that it was treated as

intelligible (in German) even though the pronunciation was 'Dutch-sounding'. For example: if the Dutch *'Hond'* was used instead of the German *'Hund',* this was treated as an intelligible utterance. Similarly, nouns that contained an error in pronunciation, but were nonetheless identifiable as the correct noun (as judged by the coder), were included in all analyses.

A chi-squared test was used to compare the expected value to the observed value of each participant. The result from this test[8] was a judgment of significance per participant, with a critical value of $p = 0.05$. This test served as an indication for which participants scored consistently different from the expected distribution, with participants that invariably (except for 1, which was later excluded for other reasons) having more *mapping followed* responses than expected when their responses indeed differed from the expected distribution. These participants comprised the *Follow* subgroup (N = 38), meaning that at the start of the experiment (in block 1) they already showed evidence for using their L1 knowledge of articles to guide their answers in their German L2: they showed evidence of '*using the mapping*'. The remaining participants were added in the *NoFollow* subgroup (N = 18), as these participants showing no significant difference from the expected random distribution of *mapping followed* responses. The total amount of participants was, then N = 56.

Production 1 and 2 Over Blocks

As a reminder, there were four consecutive blocks in the experiment, with each block introducing a new set of (compatible and incompatible) items. Within each block, there were two productions of a corresponding item before a participant received correct input on this item (in the input condition). This means that the development of the proportion of mapping responses in production 1 and 2 over the four blocks could be used as an index of learning that was not item-specific. Put differently, if the proportion of responses following the mapping increases for specifically production 1 and 2 over blocks, then this increase can only be due to the fact that the participants have been learning the general mapping between the Dutch and German gender systems over the course of the experiment, and not due to learning the specific article belonging to a specific German noun.

Because the nouns in the 1st and 2nd production had not yet received any input, the participants' responses over blocks reflect a general change in responses over time. Due to the fact that no input was provided in this analysis, the *Input* and *No Input* conditions were collapsed as they had no discernible difference. Then, before any analysis was performed, an average of the amount of *mapping responses* per *Block* for both *Compatible* and *Incompatible* nouns was computed for each participant. *Mapping responses* refer to the responses that were congruent with the expected

8 A more in-depth explanation for the test is provided in Appendix C.

responses based on grammatical gender knowledge of Dutch nouns, *Block* refers to the experimental block (1 through 4), and *Compatible and Incompatible* refer to whether the grammatical article of a German noun is in agreement with the mapping expectation (*Compatible*) or not (*Incompatible*).

In Table 2, the mean percentage of *mapping followed* responses across production 1 and 2 is reported per Subgroup (*Follow, NoFollow*), per Block (1 through 4), per Word Category (*Incompatible, Compatible*), as well as the totals of all participants regardless of their subgroup (*Total*). Figure 4 contains a plot of the development of the mapping responses of the *Follow* and *NoFollow* subgroups.

As Brandt and colleagues (submitted) previously found that participants learned the mapping throughout the experiment, a net increase in mapping usage from the first (block 1) to the last block (block 4) was expected. It is unclear whether this will hold for both subgroups, or whether it will hold only for the relatively naive group with no prior mapping knowledge (i.e. the *NoFollow* group, see Brandt et al., submitted). The net increase in mapping responses should be observed in both compatible and incompatible nouns, as the strategy of using the mapping for a correct article should be independent of Word Category. However, it might be the case that participants operate on prior knowledge, especially for the *Follow* group, which might lessen the effect in the incompatible nouns as their previous experience might lead them to correctly recognise specific nouns as exceptions to the rule, i.e. as nouns not following the mapping.

Table 2. *Mean percentage of mapping productions in production moments 1 and 2 (averaged), given per Block (1 through 4, aggregated per block), per Group (Total, Follow and NoFollow) and per Word Category (Incompatible and Compatible), with the SD and Range after aggregating.*

	Block							
	Block 1		Block 2		Block 3		Block 4	
	M (SD)	*Range*	*M (SD)*	*Range*	*M (SD)*	*Range*	*M (SD)*	*Range*
Incompatible								
Follow	61.8 (15.5)	31.8 – 94.4	65.7 (15.3)	33.3 – 100	66.1 (15.0)	37.5 – 90.9	61.2 (13.2)	36.4 – 95.83
NoFollow	51.1 (19.1)	20 – 81.8	50.3 (16.0)	27.3 – 83.3	52.6 (15.5)	29.17 – 75	56.6 (16.8)	25 – 83.3
Total	58.4 (17.3)	20 – 94.4	60.8 (17.0)	27.3 – 100	61.8 (16.3)	29.2 – 90.9	59.7 (14.5)	25.0 – 95.8
Compatible								
Follow	89.8 (7.8)	25 – 75	90.3 (9.4)	45.5 – 54.6	89.2 (10.8)	41.7 – 58.3	86 (10.7)	41.7 – 58.3
NoFollow	57.8 (15.8)	11.1 – 75	72.4 (16.6)	25 – 100	72.9 (16.3)	41.7 – 100	69.7 (13.6)	41.7 – 91.7
Total	79.5 (18.6)	11.1 – 100	84.6 (14.7)	25 – 100	84 (14.8)	41.7 – 100	80.7 (13.9)	41.7 – 100

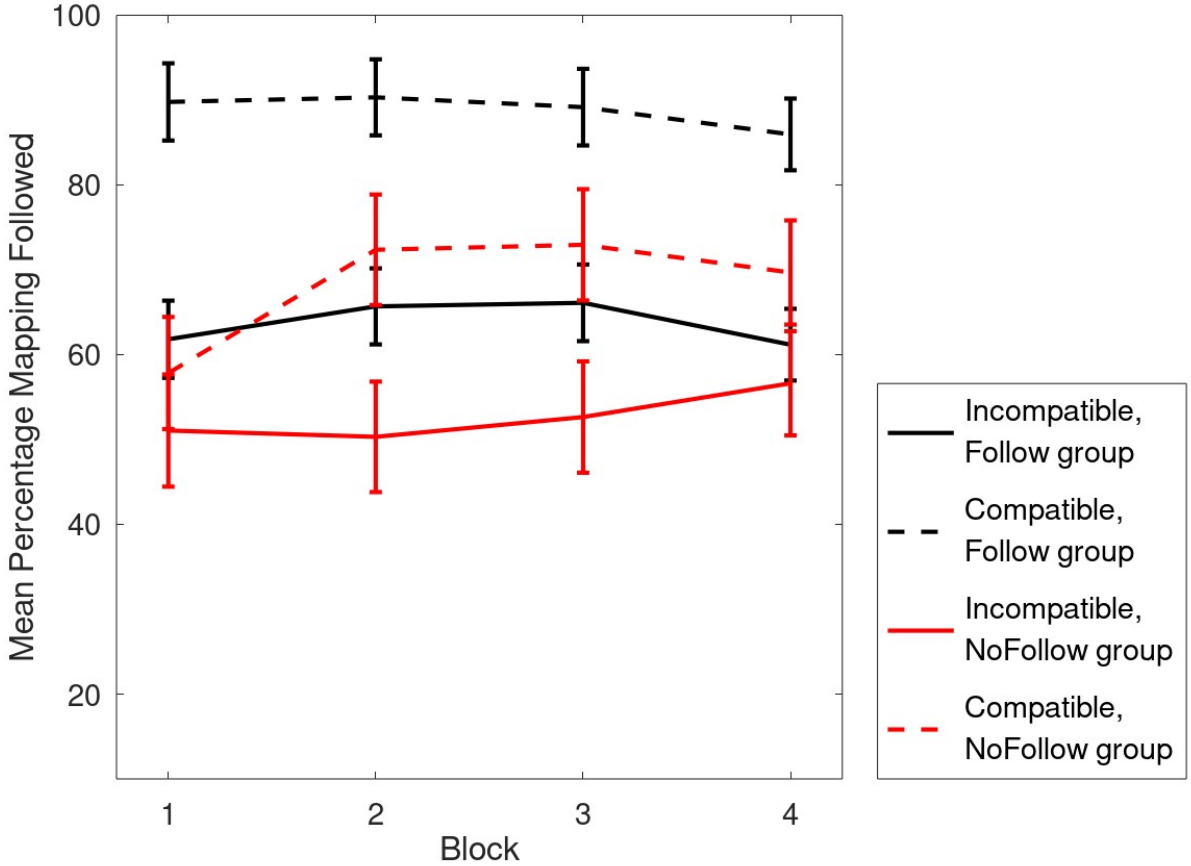

Figure 4. Mean percentage of mapping productions in production moments 1 and 2 of both subgroups (Follow and NoFollow), given per Block (1 through 4), and per Word Category (Incompatible and Compatible), with the confidence intervals as error bars.

First, it was necessary to see whether the *Follow* and *NoFollow* groups were, as expected, significantly different in their usage of mapping over blocks. An ANOVA was run with the within-subjects factors of Block (1, 2, 3 and 4) and Word Category (*Compatible* and *Incompatible*), and a between-subjects factor of Subgroup (*Follow* or *NoFollow*) in SPSS Statistics (version 27). The main effect of Subgroup was significant for both Compatible ($F(1, 54) = 59.86$, $p < 0.001$, $\eta_p^2 = 0.526$) and Incompatible ($F(1, 54) = 11.31$, $p = 0.001$, $\eta_p^2 = 0.173$) categories. In all blocks the Follow group scored higher than the NoFollow group (see Table 2). This confirms that the subgroups behaved differently, with the Follow group having higher mapping responses overall, so the subgroups were analysed independently in all further analyses. A main effect of Block was also observed for the Compatible ($F(3, 162) = 8.14$, $p < 0.001$, $\eta_p^2 = 0.131$), but not for the Incompatible ($F(3, 162) = 567$, $p = 0.637$, $\eta_p^2 = 0.01$) category. However, these contrasts were not yet very meaningful as it was unclear if only one group drove the effect in the Compatible category, or if the other obscured an effect in the Incompatible category. The uncertainty of the Compatible category's main effect was exacerbated by a significant interaction effect between Block and Subgroup for the Compatible category ($F(3, 162) = 9.528$, $p < 0.001$, $\eta_p^2 = 0.150$), which was not found for the

Incompatible cateogry ($F(3, 162) = 1.884$, $p = 0.134$, $\eta_p^2 = 0.034$), implying that the groups had a significantly different development of mapping for Compatible cognates.

When inspecting Figure 4, this pattern of results seems to stem from the obvious difference between Compatible and Incompatible categories in both groups, where responses were more often in line with the mapping for compatible nouns in both groups, but also due to a difference in development of the mapping responses over time for each group. That there is a difference between Compatible and Incompatible categories in both groups is quite surprising, as it means that the participants must have had some prior knowledge of the difference between the compatible and incompatible word categories, because, without any prior knowledge about these categories, they were indistinguishable. This finding will be revisited in more detail when discussing the subgroups separately.

When looking selectively at the Follow group, there seemed to be no obvious change in mapping responses over time. Interestingly, while the blocks did not seem to differ too much from each other, the lowest values for both Compatible (86%) and Incompatible (61.2%) cognates occurred in the 4th block. When inferential statistics were used to examine these effects more closely using an ANOVA (with the same factors as before except Subgroup) for just the Follow group there turned out to be no significant main effects of Block in either the Compatible ($F(3, 111) = 2.32$, $p = 0.079$, $\eta_p^2 = 0.059$) or Incompatible ($F(3, 111) = 1.94$, $p = 0.128$, $\eta_p^2 = 0.05$) categories, meaning that participants' answering patterns showed no significant change over time during the experiments. This finding was further supported by an absence of significant differences in the Bonferroni pairwise comparisons (for all contrasts $p > 0.1$).

Now it is time to look at the NoFollow group. The NoFollow group had a very interesting development immediately at the start. Specifically for the compatible nouns, the percentage of *mapping followed* responses jumps from 57.8% in the 1st block to 72.4% in the 2nd block (ΔM = 14.6%.). An increase for incompatible nouns was also numerically observed and started in the 2nd block after a small initial decline and developed more gradually from 50.3% to 56.6% (ΔM = 6.3%.), so it seems like the NoFollow group picked up the mapping for the compatible nouns almost immediately, while there was a far slower (possibly insignificant) increase in *mapping responses* for the incompatible nouns. As the analyses until now only focused on the productions without any input, the different treatment of Compatible and Incompatible categories by the participants implies that the NoFollow group did have prior knowledge of a difference between compatible and incompatible cognates, even though this knowledge seemed only to become more readily available from the 2nd block onwards.

When an ANOVA was run, the NoFollow group turned out to indeed behave quite differently from the Follow group. The Follow group did not show any significant change in mapping

responses over time, regardless of category. Not so for the NoFollow group. The ANOVA gave a main effect of Block for the Compatible category ($F(3, 51) = 9.06$, $p < 0.001$, $\eta_p^2 = 0.348$), but not for the Incompatible category ($F(3, 51) = 0.8$, $p = 0.5$, $\eta_p^2 = 0.045$), meaning that the percentage of mapping responses did change significantly between blocks, specifically for the compatible nouns. In the Bonferroni pairwise comparisons for the Compatible category, the only block that differed significantly from the others was the 1st block ($M = 57.84$, $SE = 3.73$). In fact, it differed significantly from the 2nd block ($M = 72.36$, $SE = 3.91$, $\Delta M = 14.53$, $p = 0.002$), 3rd block ($M = 72.95$, $SE = 3.85$, $\Delta M = 15.1$, $p = 0.006$) and 4th block ($M = 69.69$, $SE = 3.22$, $\Delta M = 11.85$, $p = 0.021$). Interestingly, only the 1st block differed significantly from the other blocks, indicating that the effect might occur quite early in the experiment, which is supported by the visualisation in Figure 4 for at least the compatible nouns. The Incompatible category showed no significant changes, nor were any of the Bonferroni pairwise comparisons significant.

As found for the Follow group before, the difference between compatible and incompatible categories can only be explained by the inexperienced participants of the NoFollow group having some (perhaps very rudimentary) form of prior knowledge about incompatible cognates.

While the Incompatible category showed no change over time, the Compatible category certainly did. And as was already expected from Figure 4, the effect was contained completely within the 1st block. In fact, the 1st block was significantly different from all other blocks. This showed, conclusively, that the effect of mapping changing over time is driven mainly by a 'jump' in mapping responses in the 1st block , though it was only found for the 'easy to map' compatible nouns. This points towards a rapid form of learning for the inexperienced (NoFollow group) participants, likely dependent on some form of interaction with their prior knowledge (at least at the group level).

It is not completely clear where this immediate learning came from. Perhaps because all participants were somewhat familiar with German from the Dutch schooling system, it might have been a recovery of mapping knowledge they learned in school, which would explain the speed of the effect. The discrepancy in learning effect between incompatible and compatible nouns indicates that at least, the participants had some access to prior knowledge of the difference between these categories. Remember, the categories should be indistinguishable for a completely naive learner of German. The participants therefore could not have been completely naive, even though they did not possess mapping knowledge at the start of the experiment. The most likely way in which the participants learned this distinction, would be in Dutch high school.

Although a discrepancy was found for both the Follow and NoFollow groups, the amount of discrepancy differs per group, where the Follow group has a bigger difference ($\Delta M = 28\%$., see Table 2) between compatible and incompatible noun categories in the first block than the NoFollow

group (ΔM = 6.7%., see Table 2), reflecting that the Follow group presumably starts out with more knowledge about which particular nouns are incompatible.

It is important (especially for the NoFollow group) to realise that the described effects have only been investigated at the group level, as individual differences could well explain part of the effect: some participants start out with usable mapping knowledge, and others might not. The participants with access to prior knowledge might have driven up the average at the group so much – especially in the NoFollow group with it's smaller size (N = 18) – that the average reflected their behaviour, while an analysis of individual differences might tell a different story. Such an analysis is, however, outside of the scope of the present study. It remains unclear where the (inexperienced) participants gained their previous knowledge. Schooling is a likely option but other sources of exposure to the incompatible 'exceptions to the rule' in German could also have provided the necessary knowledge of incompatibles. One could think of examplars like 'Das Auto.' from the well known Volkswagen commercials in which the article receives special stress (see e.g. Volkswagen Nederland, 2011), and the slogan is in itself an incompatible noun ('*Das*$_{neu}$ *Auto*' in German; '*de*$_{com}$ *auto'* in Dutch; '*the car'* in English. The slogan remained unstranslated while it was in use in the Netherlands).

In the previous study of Brandt, Lemhöfer and Schriefers (submitted) it was found that the naive German (L1) participants seemed to 'learn' the mapping throughout the experiment, having a higher percentage of responses that followed the mapping at the end of the experiment than at the beginning. In the present study, a more precise timing of the effect was examined. Instead of a linear increase, the entire effect was confined to the first quarter of the experimental procedure. In addition, this increase could only conclusively be observed in the relatively inexperienced NoFollow group. Unfortunately, as compatible cognates were not measured in previous research,it is unclear whether (relatively naive) German (L1) participants showed the same discrepancy as Dutch (L1) participants did in the present study.

In any case, the trajectory of the learning effects in the present experiment make one thing clear: it is decidedly not the case that there is a steady increase in the proportion of following the mapping responses in all Dutch participants, as it differs per category and per group, and in no case shows a clear (linear) over-time increase. Therefore it could be worthwhile for further research to zoom in on individual differences, as these group effects might obscure the real underlying story. It might also be worthwhile to compare Compatible and Incompatible cognates over time in naive German participants, to see whether the discrepancy and learning trajectory as found in the present study can be replicated.

Until now, we have only discussed the data in terms of *mapping followed* responses as the dependent variable. This implies that also responses that are actually incorrect articles can be counted as *mapping followed*. For example, for the German noun '*Mund'* both '*die*$_{fem}$*'* and '*der*$_{masc}$*'*

would count as *mapping followed* while only '*die'* would count as a *correct article*. The choice of 'mapping followed' as dependent variable is determined by the fact that we were, in this part of the paper, interested in the learning of the mapping, and not in the learning of the specific correct article.

Learning the articles by explicit input

Having discussed the general changes in mapping over time, we now turn to the effect of native input on article production. This input, provided by the 'virtual dialogue partner', was expected to cause item-specific learning in an incidental fashion. The 'incidental' character of the item-specific learning was ensured by providing this input while the focus the of the participants was put only on retaining nouns in their memory, not on the input itself or learning of language in general.

For the analysis of item-specific learning, the dependent variable needed to be changed to '*correct article responses'* (whether the participant selected the correct German article), instead of the previously used '*mapping followed responses*' (whether the participant followed the mapping-based strategy in selecting an article)[9]. As explained at the end of the preceding section, for the mapping of Dutch article knowledge onto German articles, 'following the mapping' does not always and necessarily lead to the choice of the *correct article*. It should be noted that this is different for the reverse mapping from German article knowledge onto Dutch articles (as studied in e.g. Brandt, Schriefers & Lemhöfer, 2021). In this latter direction of the mapping, 'following the mapping' will always lead to a choice of the correct article.

For the analyses on item-specific learning, all data were aggregated across blocks. This appears to be justified given that no substantial non-item specific learning across blocks was observed, at least not in the majority of participants (the Follow group). In Table 3, the means, standard deviations and range for all participants (N = 52), as well as for the Follow (N = 35) and NoFollow (N = 17) groups are given separately for noun category and production moment. Note that 4 participants – 1 from the NoFollow group, 3 from the Follow group – had to be excluded due to a failure to register their posttest data.

Figure 5 plots the corresponding data as line graphs for the both subgroups to visualise the development of the percentage of correct article responses over production moments.

9 A summary of the descriptive statistics of the *'mapping followed'* results, as well as their visualisations, are presented in Appendix D, but do not include inferential statistics and are not discussed further in this paper..

Table 3. *Means, Standard Deviation and Range for the Follow and NoFollow subgroups, for all Word Categories (Incompatible and Compatible) and all production moments (1, 2, 3 and Posttest).*

	Production Moment							
	Production 1		Production 2		Production 3		Posttest	
	M (SD)	*Range*	*M (SD)*	*Range*	*M (SD)*	*Range*	*M (SD)*	*Range*
Incompatible with Input								
Follow	27.6 (11.7)	5.3 – 52.2	28.0 (12.8)	4.3 – 47.8	52.4 (18.7)	11.8 – 100	41.2 (14.5)	13.6 – 72.7
NoFollow	33.8 (8.5)	19.0 – 47.6	33.6 (8.0)	14.3 – 45.0	40.0 (13.2)	25.0 – 65.0	41.4 (16.7)	9.1 – 72.7
Total	29.6 (30.2)	5.3 – 52.2	29.8 (11.7)	4.3 – 47.8	48.4 (17.9)	11.8 – 100	41.3 (15.1)	9.1 – 72.7
Incompatible without Input								
Follow	30.4 (13.4)	10.5 – 68.0	29.8 (13.5)	0 – 64.0	32.6 (13.5)	13.0 – 72.0	39.3 (12.2)	11.1 – 66.7
NoFollow	31.9 (8.8)	13.6 – 42.9	33.4 (8.8)	13.6 – 50.0	38.4 (11.1)	18.2 – 56.0	37.5 (14.3)	14.8 – 66.7
Total	30.9 (12.0)	10.5 – 68.0	31.0 (12.2)	0 – 64.0	34.5 (32.7)	59.0 – 72.0	38.7 (16.3)	31.6 – 87.5
Compatible with Input								
Follow	63.4 (10.5)	45.5 – 86.4	66.7 (9.1)	50.0 – 86.4	73.5 (12.3)	47.8 – 100	69.3 (10.7)	43.5 – 87.0
NoFollow	45.4 (38.9)	31.8 – 69.6	46.2 (14.3)	23.8 – 71.4	50.2 (15.8)	27.8 – 80.0	53.7 (15.3)	21.7 – 73.9
Total	57.9 (14.2)	31.8 – 86.4	60.0 (14.6)	23.8 – 86.4	65.9 (17.3)	27.8 – 100	64.2 (14.3)	21.7 – 87.0
Compatible without Input								
Follow	69.5 (13.4)	31.6 – 87.5	69.6 (11.5)	36.8 – 91.7	70.5 (11.6)	36.8 – 87.0	70.1 (9.9)	52.0 – 96.0
NoFollow	48.7 (12.4)	31.6 (76.0)	47.9 (13.2)	26.3 – 72.0	48.0 (14.2)	25.0 – 66.7	54.8 (11.7)	36.0 – 76.0
Total	62.7 (16.3)	31.6 – 87.5	62.5 (15.8)	26.3 – 91.7	63.2 (16.3)	25.0 – 87.0	65.1 (12.7)	36.0 – 96.0

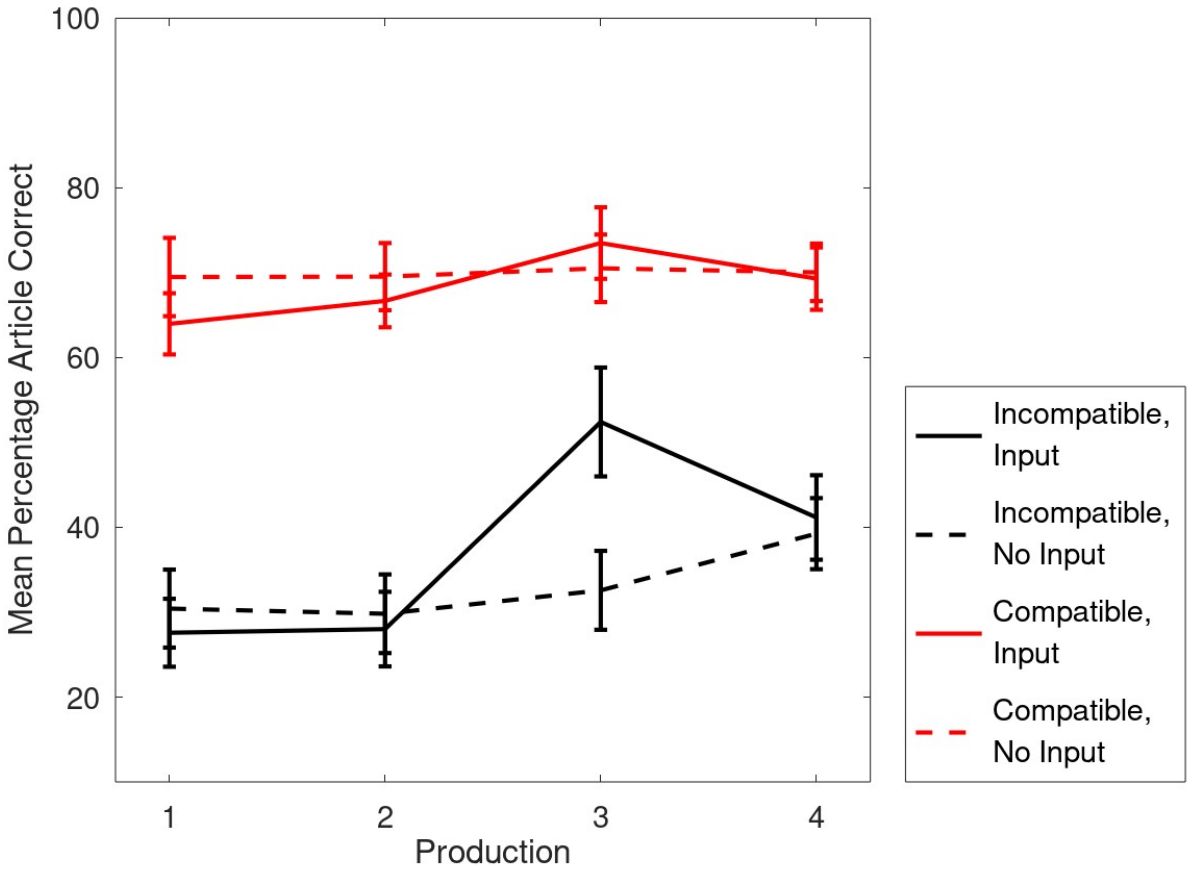

Figure 5. Mean percentage of correct article responses of only the Follow subgroup, given per Production Moment (1, 2, 3 and Posttest), and per Word Category (Incompatible and Compatible), with the confidence intervals as error bars.

A repeated-measures ANOVA was run with the within subjects factors of Production Moment (4 levels: Production moment 1, 2, 3 and Posttest) x Input (2 levels: Input and No Input) x Word Category (2 levels: Compatible and Incompatible). It was deemed helpful for interpreting the results to combine the Word Category and Input factors into four distinct categories: Incompatible with Input (1); Incompatible without Input (2); Compatible with Input (3); Compatible without Input (4). Each category was expected to have a distinct developmental pattern over production moments, which is why it made sense to discuss them as four different categories.

The ANOVA was run separately for the Follow and NoFollow groups because the groups were expected and shown to behave differently in terms of mapping behaviour in the previous analyses. In general, the most important prediction was a significant positive effect of Production Moment on correct article responses in the two conditions in which input was provided (*Compatible with Input* and *Incompatible with Input*), as was found before in the incompatible nouns of Brandt and colleagues (2021). Based on the same study's results, this effect should be absent for the two categories without input (*Compatible without Input* and *Incompatible without Input*).

Additionally, the effect of input was expected to occur between the 2nd and 3rd production moment, as input was provided in between these production moments. So, if there indeed was an effect of Production Moment in the word categories that received input, the effect should occur in between the 2nd and 3rd production moment, as only then the effect could surely be attributed to the native input provided in the dialogue game (or *incidental learning task*). It was checked whether this was indeed the case by looking at Bonferroni pairwise comparisons between production moments.

Lastly, given the previous results of the analysis across blocks, the contrast between compatible and incompatible nouns was assessed. The Follow group was expected to show this difference between categories most clearly, as they were selected for following the mapping. This should result in a lower percentage of correct article responses in the incompatible categories than the compatible categories, as participants who tend to follow the mapping were expected to make mistakes in the incompatible categories (see e.g. Lemhöfer et al., 2010). The NoFollow group, too, showed a significant difference between incompatible and compatible categories in the previous analyses. So for the present analyses, a lower percentage of correct article responses in the incompatible categories was also expected, although this would likely be a smaller difference with the compatible categories than in the Follow group.

Starting with the Follow group, there was no main effect of Input ($F(1, 34) = 1.330$, $p = 0.257$, $\eta_p^2 = 0.038$). As this main effect does not take into account the different production moments, a null effect is not very surprising. Indeed, both the main effect of Production ($F(2.5, 86.4) = 4.77$, $p < 0.001$, $\eta_p^2 = 0.574$, *Greenhouse-Geisser corrected*) and the interaction effect between Input and Production were found to be significant ($F(2.4, 80.6) =32.507$, $p < 0.001$, $\eta_p^2 = 0.489$). This showed that without taking the word categories into account, an effect of input was detected in some production moments, but not necessarily in all. This was as expected, as input was only provided after the 2nd production moment, and the effect could only occur after input was provided.

When the word categories were also taken into account, it became clear that they indeed needed to be discussed separately. There was a large main effect of Word Category ($F(1, 34) = 429.289$, $p < 0.001$, $\eta_p^2 = 0.927$), meaning the compatible and incompatible nouns indeed showed very different percentages of correct article responses. Supporting the notion that the participants treated (in-)compatible words as separate categories, the interaction effect between Input and Word Category ($F(1, 34) = 8.868$, $p = 0.005$, $\eta_p^2 = 0.207$) as well as the interaction effect between Input, Word Category and Productions were also significant $F(2.5, 85.2) = 10.307$, $p < 0.001$, $\eta_p^2 = 0.233$, *Greenhouse-Geisser corrected*). Because these effects clearly showed that participants behaved differently regarding the two word categories, the categories were separated in the ANOVA analyses below.

To get an overview of what to expect for each different word category (Incompatible with Input (1); Incompatible without Input (2); Compatible with Input (3); Compatible without Input (4)), it is useful to first consult the descriptive plot of Figure 5. The incompatible nouns seem to have indeed benefited from input, if it was provided. There seems to be a large spike in correct article responses from the 2nd (M = 28.0, SD = 12.8) to the 3rd (M = 52.4, SD = 18.7) production moment for the Incompatible with Input category. By contrast, such an effect seemed absent for Incompatible without Input category. A similar pattern emerges for the compatible nouns, though the relative increase in correct article responses from the 2nd to the 3rd production moment – which indicates the effect of input – appears much smaller than for the incompatible nouns.

When running the ANOVA for only the incompatible cognates, dropping the Word Category variable, there were indeed a significant effect of Input ($F(1, 34) = 9.231$, $p = 0.005$, $\eta_p^2 = 0.214$) and of Production ($F(2.4, 80.9) = 55.435$, $p < 0.001$, $\eta_p^2 = 0.62$, *Greenhouse-Geisser corrected*), as well as a significant interaction effect between these two factors ($F(2.4, 82.8) = 36.64$, $p < 0.001$, $\eta_p^2 = 0.519$, *Greenhouse-Geisser corrected*). These effects show that, as Figure 5 suggested, there was indeed an effect of input present in the incompatible cognates over production moments. To discover whether this effect was selectively present only for the words that received input, and whether it occurred at the expected time (indicated by an increase in correct articles from the 2nd to the 3rd production moment), another ANOVA was run for the Input and No Input conditions separately. To check in between which production moments exactly an effect of input might have occurred, Bonferroni-corrected pairwise comparisons were assessed for each significant main effect.

Starting with the Incompatible with Input category, there was indeed a significant main effect of Production ($F(2.1, 72.4) = 58.864$, $p < 0.001$, $\eta_p^2 = 0.634$, *Greenhouse-Geisser corrected*). When consulting the pairwise-comparisons, the effect of input seemed to follow the expected pattern, with the 1st (M = 27.6, SE = 2.0) and 2nd (M = 28.0, SE = 2.2) production moments showing no significant difference ($p = 1$), while there was a significant difference between the 2nd and 3rd (M = 52.4, SE = 3.2, ΔM = 24.4, $p < 0.001$) production moments. The posttest (M = 41.2, SE = 2.5) also showed a significant difference from the 1st and 2nd ($p < 0.001$) production moments, suggesting that the input effect was at least somewhat stable over time. However, the amount of correct article responses in the posttest was significantly lower than in the 3rd production moment ($\Delta M = 11.3$, $p < 0.001$) as well, as is illustrated in Figure 5. This might simply reflect a loss of learning between the 3rd production moment and the posttest, but it might also reflect some other general effect, that could be due to the difference in method between the posttest and the other production moments. This problem will be elaborated upon below.

When consulting the inferential statistics for the Incompatible without Input category, it turns out that this category, too, showed a significant main effect of production moment ($F(2.2, 76.0) = 18.628$, $p < 0.001$, $\eta_p^2 = 0.354$, *Greenhouse-Geisser corrected*). This effect was wholly unexpected: How could participants learn the correct articles for each noun in the absence of input of said articles?

In the pairwise comparisons the only production moment that showed any significant differences at all was the posttest ($M = 39.3$, $SE = 2.1$). It was significantly different from the 1st ($M = 30.4$, $SE = 2.3$, $\Delta M = 8.8$, $p < 0.001$), 2nd ($M = 29.8$, $SE = 2.3$, $\Delta M = 9.4$, $p < 0.001$) and 3rd ($M = 32.6$, $SE = 2.3$, $\Delta M = 6.7$, $p = 0.002$) production moments. No other significant differences between production moments were observed (all $p > 0.14$). This means that the effect seemingly occurred only after the formal experiment had ended. But what is the explanation for this effect? Is the effect due to a decrease in overall mapping responses over time? This appears unlikely as the previous analysis of mapping development over timee showed no increase or decrease of mapping responses for the Follow group.

Another possibility is that the posttest was so different from the rest of the experiment that it focused participants' attention on whether the articles were correct, rather than the (distractor) memory task of the incidental learning paradigm. The posttest responses were recorded through physical clicks on one of three articles on screen, perhaps giving the 'article learning' nature of the experiment away. The most pessimistic interpretation of this finding is that the methods drove all of the effect. But this might not necessarily be the only explanation. In particular, the previous research of Brandt and colleagues (2021, submitted) did not find any such effect. This shows that somehow the change from German (L1) to Dutch (L1) participants must have had an influence on the results.

What, then, could the other explanations be? Learning could not have happened due to input, as for this contrast there was none, nor can it have been based on 'learning the mapping', as that would have resulted in more errors in the articles for incompatible nouns, instead of more correct responses. What is left, then, is some sort of repetition effect. Each repeated production might make a participant more likely to switch selectively to the correct article, without changing the articles which they already know to be correct. Or, perhaps, the change from implicit learning and spontaneous generation of articles to explicit retrieval of articles made previously learned knowledge of articles available, which might explain the difference between Dutch (L1) and the previously studied German (L1) participants. Based on the present study, any explanation of the effect is highly speculative. Still, it is clear that there is a lot for future research to unpack this unexpected finding. And as we will see for the NoFollow group, this is not the only unexpected learning effect either.

First, to finish up the Follow group's analyses, we turn to the compatible categories. For the compatible categories, the same procedure was followed as for the incompatible categories. The expectation for both the input effect was the same as for the Incompatible words: a positive effect of input on the percentage of correct article responses, located in between the 2nd and 3rd production moments, and no effect of an absence of input. Although the first ANOVA did not find a main effect of Input ($F(1, 34) = 0.861$, $p = 0.36$, $\eta_p^2 = 0.025$), there was a main effect of Production ($F(2.5, 83.1) = 6.598$, $p = 0.001$, $\eta_p^2 = 0.163$; *Greenhouse-Geisser correction*) and a significant interaction effect between input and production moment ($F(2.6, 87.4) = 4.948$, $p = 0.005$, $\eta_p^2 = 0.127$; *Greenhouse-Geisser correction*). As for the incompatible categories, this means that the Input and No Input conditions indeed behaved differently from each other in terms of production moments, even though there was no clear effect of input over all production moments. Again, as before, two separate ANOVAs and their respective Bonferroni-corrected pairwise comparisons were consulted to disentangle the differences in production moments for the Input and No Input conditions.

Only the Compatible with Input category showed a significant effect of Production ($F(2.2, 84.8) = 10.389$, $p < 0.001$, $\eta_p^2 = 0.234$; *Greenhouse-Geisser correction*). There was no effect for the Compatible without Input category ($F(2.295, 78.020) = 0.184$, $p = 0.860$, $\eta_p^2 = 0.005$; *Greenhouse-Geisser correction*). Thus, so far, the expectations were confirmed: an effect was only found in the category where input was received.

Unsurprisingly, when consulting the pairwise comparisons for the Compatible without Input category, the production moments showed no significant differences from one another ($p = 1$ for all contrasts). For the pairwise comparisons for the Compatible with Input category, however, the 3rd ($M = 73.5$, $SE = 2.1$) production moment was significantly different from the 2nd ($M = 66.7$, $SE = 1.5$, $\Delta M = 6.8$, $p = 0.004$) and 1st ($M = 64.0$, $SE = 1.8$, $\Delta M = 9.5$, $p < 0.001$) production moments, but not significantly different from the posttest ($M = 69.3$, $SE = 1.8$, $\Delta M = 4.2$, $p = 0.268$). The posttest only differed from the 1st production moment significantly ($\Delta M = 5.3$, $p = 0.011$). This finding indicates a decrease in correct article responses in the posttest after the 3rd production moment (see Figure 5). Thus it seems that there is a relatively small but robust effect of input (2nd vs 3rd production moment) that did not persist over time, i.e. until the posttest. This is in agreement with the conclusion for the incompatible categories before. There seems, however, to be no evidence of the posttest behaving differently from the other production moments. No other significant differences were observed in the pairwise comparisons (all $p > 0.27$).

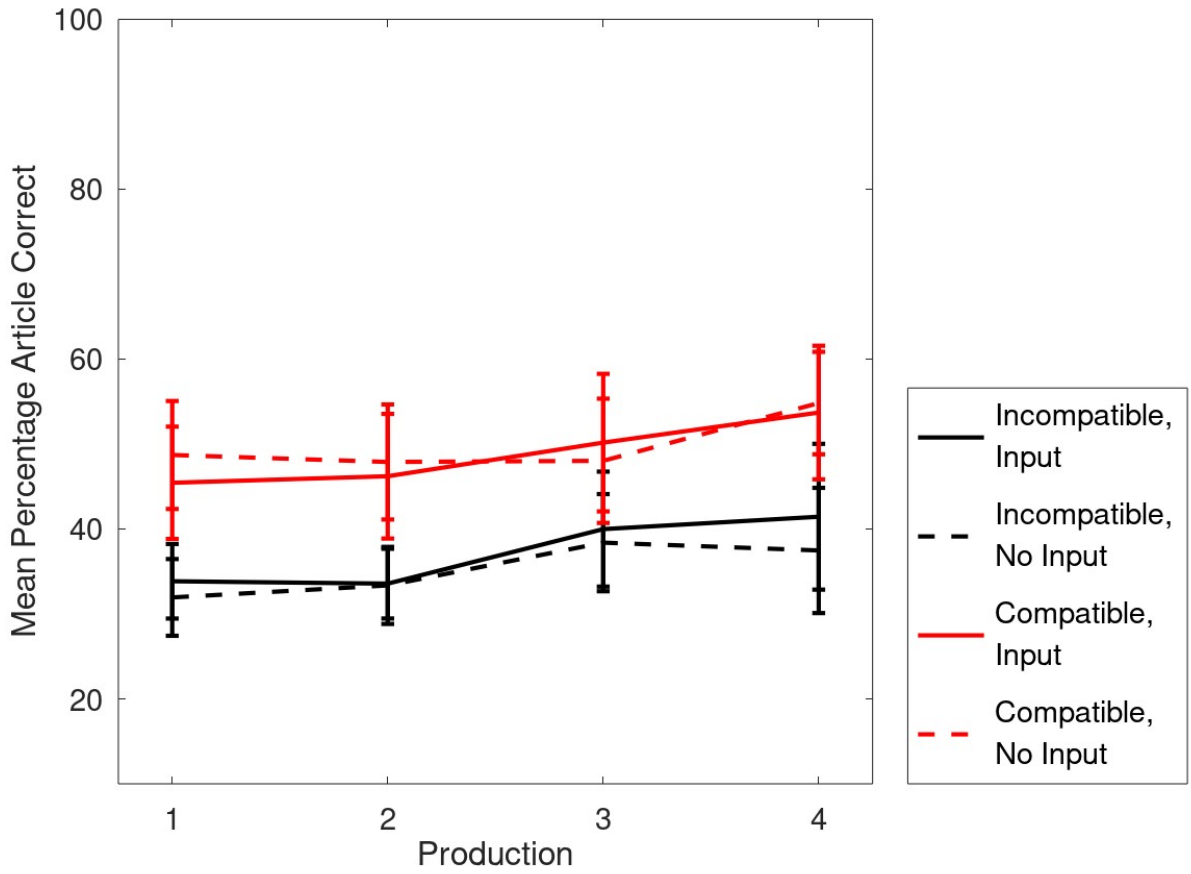

Figure 6. Mean percentage of mapping productions of only the NoFollow subgroup, given per Production moment (1, 2, 3 and Posttest), and per Word Category (Incompatible and Compatible), with the confidence intervals as error bars.

Figure 6 shows the results for the subgroup of participants that did not follow the mapping at the beginning of the experiment (the NoFollow group). The same procedure as before was followed. The only descriptive change across production moments present in Figure 6 appears to be a small general increase in correct article responses. This increase seems to be present for both compatible and incompatible words and regardless of the presence or absence of input. A priori there were no clear expectations for the NoFollow group, so the best way to examine this group was to see what differences there were compared to the Follow group.

To see whether the NoFollow group showed the same patterns as the Follow group, the same analysis strategy was followed. The first ANOVA showed no significant main effect of Input ($F(1, 16) = 0.155$ $p = 0.699$, $\eta_p^2 = 0.1$), but did show a significant main effect of both Word Category ($F(1, 16) = 18.096$, $p < 0.001$, $\eta_p^2 = 0.531$) and Production ($F(1.6, 25.9) = 8.238$, $p = 0.003$, $\eta_p^2 = 0.34$; *Greenhouse-Geisser correction*). This is indeed parallel to what was found in the Follow group above with largely the same interpretation. However, in partial contrast to the Follow group, there was no significant interaction effect of Input and Word Category ($F(1, 16) = 0.745$, $p = 0.401$, $\eta_p^2 = 0.044$), nor one of Production and Word Category ($F(3, 48) = 1.366$, $p = 0.264$, $\eta_p^2 = 0.079$),

nor one of Input and Production ($F(1.7, 27.8) = 0.003$, $p = 0.562$, $\eta_p^2 = 0.034$; *Greenhouse-Geisser correction*). Since the three-way interaction effect of Input, Word Category and Production was also not significant ($F(1.9, 29.7) = 0.542$, $p = 0.656$, $\eta_p^2 = 0.033$; *Greenhouse-Geisser correction*), this showed that while incompatible and compatible nouns definitely elicited different amounts of correct responses, the development of the response patterns over production moments was similar, regardless of input. Since the incompatible and compatible categories still showed a significant dissimilarity in values, they were nonetheless investigated further with separate ANOVAs.

The Incompatible category showed no main effect of Input ($F(1, 16) = 0.721$, $p = 0.408$, $\eta_p^2 = 0.043$, but did show a significant main effect of Production ($F(1.7, 27.8) = 0.003$, $p = 0.562$, $\eta_p^2 = 0.034$; *Greenhouse-Geisser correction*). The interaction effect between Input and Production, however, was nonsignificant ($F(1.7, 27.8) = 0.003$, $p = 0.562$, $\eta_p^2 = 0.034$; *Greenhouse-Geisser correction*), which means that there was no difference between the Input and No Input conditions over production moments. But without input, how can there have been any learning? Since the productions were aggregated over blocks, the passage of time cannot explain the effect, so the effect must somehow be explained just by the productions themselves.

To disentangle this finding, Bonferroni paired-comparisons were assessed. These showed that for the incompatible nouns, only the 3rd ($M = 39.2$, $SE = 2.5$) production moment is significantly different from the 1st ($M = 32.9$, $SE = 1.7$; $\Delta M = 6.3$, $p = 0.004$) and 2nd ($M = 33.5$, $SE = 1.6$; $\Delta M = 5.7$, $p = 0.024$) production moments. No other comparisons were significantly different, including comparisons with the posttest (all $p > 0.14$). Although at first glance this looks like an effect of input, remember that the Input and No Input conditions show no difference.Since the passage of time, and with time more exposure to the experimental stimuli, was controlled for in this analysis, the only remaining explanation is a sort of repetition effect. The Production factor is, in essence, only the repetition of a sentence production. And since this is the only factor of significant influence, it is the only remaining explanation. It also makes sense, in a way, as participants might change to correct articles more readily after repetitions. They might remember a correct article after repeating the nouns more and more, and be reluctant to change back to an incorrect article.

The Compatible nouns showed the same pattern as the incompatible nouns, with only a main effect of Production ($F(1.8, 29.5) = 5.196$, $p = 0.013$, $\eta_p^2 = 0.245$; *Greenhouse-Geisser correction*), but not of Input ($F(1, 16) = 0.285$, $p = 0.601$, $\eta_p^2 = 0.017$), nor an interaction effect between Production and Input ($F(2.0, 31.2) = 74.8$, $p = 0.478$, $\eta_p^2 = 0.045$; *Greenhouse-Geisser correction*). In contrast to the Incompatible category, the pairwise comparisons for the Compatible category showed that only the 2nd ($M = 47.1$, $SE = 2.7$) production moment and posttest ($M = 54.3$, $SE = 3.0$, $\Delta M = 7.2$, $p = 0.045$) differed significantly from one another (for all other contrasts: $p > 0.07$). Again, this points to an effect of repetition rather than input. This could be due to participants only

changing their answer when they remember a correct article, changing over repetitions towards more correct article usage. Another possibility could be the methodological difference of the posttest and the other production moments, although the clear absence of a difference between 3rd (M = 49.1, SE = 3.3) production moment and posttest (ΔM = 5.2, p = 0.549) is not consistent with this explanation.

So, a repetition effect seems to be the most likely explanation presently available. Regrettably, any explanation at this point is purely speculative, because unfortunately the differences between production moments in the pairwise comparisons are small and do not provide a clearly interpretable picture. Still, a repetition effect is consistent with the unclear picture, and is even consistent with the previous inexplicable effect of the posttest in the Follow group. In fact, all presently found results are in line with the repetition-effect explanation. This does lead to an important question: How does the proposed effect hold up in a German (L1) sample? In previous research, no such repetition-like effect was found, so this remains an open question for future research. Recommendations on how to tackle this problem will be provided in the discussion section below.

Learning the mapping: Delayed posttest

The last question to be put to the test in the present study is whether any potential effects of learning can still be observed after a longer period of time. This was tested in a delayed posttest, about six weeks after the lab session. Because the participants voluntarily responded to the email prompt six weeks after the lab session, these participants were likely more highly motivated than the other participants. This might have influenced the results for the delayed posttest, so this should be kept in mind when comparing these results to the other results in the experiment.

Table 4 summarizes the mean, SD and range of a subgroup of participants (N = 18) that participated in the delayed posttest. This means that there were 5 instead of the former 4 production moments: the 1st, 2nd and 3rd (after input) production moments, the posttest (4th) and the delayed posttest (5th). Of the total participant group of 18, only 4 were from the original NoFollow group. Because the NoFollow group was so much smaller than the Follow group for this measurement, the subgroup-specific effects will not be discussed. The Figure 7 is a plot of the means and SD as described in Table A, showing the development over production moments for each word category (Incompatible with Input (1); Incompatible without Input (2); Compatible with Input (3); Compatible without Input (4)).

Table 4. *Means, Standard Deviation and Range of percentage article correct for all participants together (N = 18), for all Word Categories (Incompatible and Compatible) over all production moments including the Delayed Posttest (1, 2, 3, Posttest and Delayed Posttest).*

	Production Moment									
	Production 1		Production 2		Production 3		Posttest		Delayed Posttest	
	M (SD)	*Range*	*M (SD)*	*Range*	*M (SD)*	*Range*	*M (SD)*	*Range*	*M (SD)*	*Range*
Incompatible with Input	26.3 (8.7)	13.0 – 39.1	29.1 (12.5)	5.3 – 47.8	43.9 (18.2)	21.1 – 100	40.9 (15.9)	13.6 – 72.7	43.4 (14.0)	13.6 – 63.6
Incompatible without Input	27.2 (11.1)	10.5 – 48.0	26.1 (10.9)	0 – 48.0	27.6 (10.3)	13.0 – 52.0	37.4 (12.6)	14.8 – 66.7	37.9 (15.1)	14.8 – 63.0
Compatible with Input	57.9 (12.6)	31.8 – 81.8	60.6 (13.8)	27.3 – 81.8	67.3 (15.9)	35.7 – 100	65.2 (11.4)	43.5 – 87.0	65.9 (15.2)	43.5 – 91.3
Compatible without Input	65.6 (14.9)	26.3 – 91.7	67.0 (14.8)	27.8 – 84.2	64.9 (11.0)	27.8 – 84.2	64.9 (11.0)	44.0 – 84.0	69.3 (13.4)	62.0 – 88.0

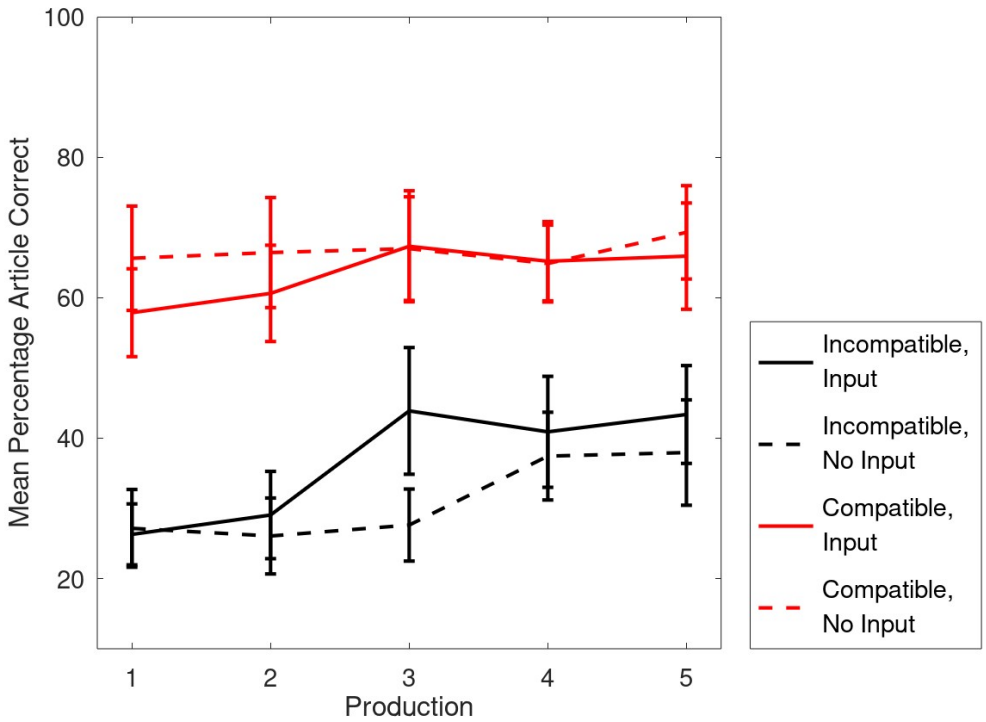

Figure 7. The Mean and SD of all participants that participated in the delayed posttest are plotted per word category and production moment (1, 2, 3, posttest (4) and delayed posttest (5)).

The pattern in Figure 7 is similar to the one in Figure 5, which showed the results of the Follow group. This is not really surprising as the majority of the participants summarized in Figure A are members of the Follow group. The delayed posttest analyses follow the same general strategy as was used before, but now over 5 instead of 4 production moments, and without separately discussing the Follow and NoFollow groups.

Consistent with the analyses of the Follow group before, there was a significant main effect of Word Category ($F(1, 17) = ,8.681$ $p < 0.001$, $\eta_p^2 = 0.873$), as well as Production ($F(1.8, 29.9) = 6.562$, $p = 0.006$, $\eta_p^2 = 0.279$; *Greenhouse-Geisser correction*), but not for Input ($F(1, 17) = 0.71$, $p = 0.411$, $\eta_p^2 = 0.040$). Additionally, there was a significant interaction effect for Input and Word Category ($F(1, 17) = 24.351$, $p < 0.001$, $\eta_p^2 = 0.589$) and Input and Production ($F(3.0, 51.6) = 6.394$, $p < 0.001$, $\eta_p^2 = 0.273$; *Greenhouse-Geisser correction*), but not for Word Category and Production ($F(1.8, 31.0) = 1.990$, $p = 0.157$, $\eta_p^2 = 0.105$; *Greenhouse-Geisser correction*), nor for the threeway interaction of Word Category, Input and Production ($F(4, 68) = 1.331$, $p = 0.267$, $\eta_p^2 = 0.073$). Since there was a significant difference between the Incompatible and Compatible categories, these were again analysed separately.

For the Incompatible category the results are, again, completely in line with the previous results of the Follow group. A main effect of Input ($F(1, 17) = 14.328$, $p = 0.001$, $\eta_p^2 = 0.457$) and of Production ($F(2.0, 33.3) = 7.691$, $p = 0.002$, $\eta_p^2 = 0.321$; *Greenhouse-Geisser correction*) were found, as well as an interaction effect between Input and Production ($F(4, 68) = 5.259$, $p < 0.001$, $\eta_p^2 = 0.236$). Again, since the Input and No Input conditions showed significant differences from one another, both conditions are discussed separately in the following.

For the Incompatible with Input category, a significant main effect of Production was observed ($F(2.651, 45.067) = 7.808$, $p < 0.001$, $\eta_p^2 = 0.315$; *Greenhouse-Geisser correction*). The pairwise comparisons showed no difference between the 1st ($M = 26.3$, $SE = 2.1$) and 2nd ($M = 29.1$, $SE = 2.9$, $\Delta M = 2.8$, $p = 1$) production moments , while the 3rd ($M = 43.9$, $SE = 4.3$) production moment differed significantly from both the 1st ($\Delta M = 17.6$, $4.2p = 0.006$) and 2nd ($\Delta M = 14.8$, $p = 0.033$) production moments. This is very similar to the input effects we have seen before in the Follow group.

The posttest and the delayed posttest did not differ significantly from each other ($p = 1$) or the 3rd production moment ($p = 1$), but while the posttest ($M = 64.1$, $SE = 3.2$) differed from both the 1st ($\Delta M = 13.1$, $SE = 3.6$, $p = 0.019$) and 2nd ($\Delta M = 15.9$, $SE = 4.6$, $p = 0.028$) production moments, the delayed posttest ($M = 85$, $SE = 3.1$) only showed a significant difference from the 1st ($\Delta M = 14.2$, $SE = 5.3$, $p = 0.155$), but not the 2nd ($\Delta M = 12.3$, $SE = 5.5$, $p = 0.401$) production moment.

The Incompatible without Input category also, unexpectedly but again similar to the previous analysis of the Follow group, showed a significant main effect of production ($F(1.924, 32.709) =$

5.892, $p = 0.007$, $\eta_p^2 = 0.257$; *Greenhouse-Geisser correction*), and it seemed to stem, again, from the posttest. The posttest ($M = 37.4$; $SE = 3$) was significantly different from the 1st ($M = 27.2$; $SE = 2.6$; $\Delta M = 10.3$; $p = 0.027$), 2nd ($M = 26.1$; $SE = 2.6$; $\Delta M = 11.4$; $p = 0.005$) and 3rd ($M = 27.6$; $SE = 2.4$; $\Delta M = 9.8$; $p = 0.012$) production moments, but not the delayed posttest ($M = 38$; $SE = 3.6$; $\Delta M = 0.5$; $p = 1$). No other production moments were significantly different from one another (all $p > 0.21$).

For the Compatible categories, the results were largely inconsistent with the previous results. Although there was a consistent absence of an input effect ($F(1, 17) = 3.704$, $p = 0.071$, $\eta_p^2 = 0.179$), the absence of both a main effect of production ($F(1.6, 27.6) = 1.097$, $p = 0.336$, $\eta_p^2 = 0.061$; *Greenhouse-Geisser correction*) and an absence of an interaction effect between input and production ($F(2.5, 42.3) = 1.856$, $p = 0.160$, $\eta_p^2 = 0.098$; *Greenhouse-Geisser correction*) were not found previously in the Follow group. None of the pairwise comparisons showed a significant difference (all $p > 0.18$). Although these results were inconsistent, the previous effects of the compatible categories were quite small compared to the incompatible categories. Because of this, there is not much that changes in terms of interpretation given this new set of results. Especially because the delayed posttest was performed on a much smaller sample (N = 18) compared to the previous Follow group analyses (N = 35), there is no good reason to suddenly change any previous interpretation in the light of the absence of an effect in the delayed posttest.

Consistent responses

In addition to the analyses above, it was important to control for the possibility that the effects were driven by the inconsistent responses in the first two production moments, that is by items on which a participant produced a different article in the 1st and 2nd production moment (Brandt, Lemhöfer & Schriefers, 2021). These inconsistent, or unstable, responses could mean participants first simply guess an article, then changing it only to the correct article when they hear the input. The stable responses, by contrast, could imply that a participant already had a stable association between a word and an article, and if this is the incorrect article, it might be more difficult to change this by input. By looking only at the stable responses and comparing these to the effects that were found across all responses (i.e. independent of in-/consistency in the 1st and 2nd production moments), any effect that is primarily driven by inconsistent answers should become apparent. Additionally, the amount of inconsistencies in and of itself will give an indication of whether the added article category in German would make Dutch participants be less likely to give consistent responses. An added article gives one more possible answer category, which would result in less consistency when guessing because of more possibility for variance.

The consistency was calculated by first selecting only those productions that were consistently intelligible across the 1st and 2nd production moments, so that the interpretation would be as straightforward as possible. Only after this criterion was met, the consistency was calculated: if the article was the same in both the 1st and 2nd production moment for one participant, it was counted as a consistent, or stable, response. If not, it was an inconsistent or unstable response. This resulted in a consistency of 82.4%: out of all selected nouns, the produced article was the same in the 1st and 2nd production moment 82.4% of the time. These trials were selected to replicate some of the analyses above, to see if the effects reported above might primarily have been driven by the (now absent) inconsistent responses.

Table 5 gives the mean, SD and range for all consistent responses, calculated in the same way as in the analyses of the Over Block comparisons. Specifically, this table can directly be contrasted to Table 3 in terms of content. The figures and statistical analyses to go with Table 3 can be found in Appendix E. The mean, SD and confidence interval are plotted in Figure E1 for the Follow group and Figure E2 for the NoFollow group.

Table 5. *Means, Standard Deviation and Range for the Follow and NoFollow subgroups, for all Word Categories (Incompatible and Compatible) over all production moments (1, 2, 3 and Posttest), for only the consistent responses.*

	Production Moment					
	Production 1 & 2 (identical)		Production 3		Posttest	
	M (SD)	*Range*	*M (SD)*	*Range*	*M (SD)*	*Range*
Incompatible with Input	29.0 (13.3)	0 – 57.1	47.5 (19.3)	6.3 – 100	39.4 (16.7)	6.3 – 72.2
Follow	26.4 (14.0)	0 – 50.0	51.2 (20.3)	6.3 – 100	38.6 (16.0)	6.3 – 71.4
NoFollow	34.2 (10.2)	12.5 – 57.1	39.7 (14.7)	20.0 – 78.6	40.9 (18.3)	10.0 – 72.2
Incompatible without Input	30.6 (13.8)	0 – 72.7	33.1 (13.9)	11.1 – 77.3	41.1 (14.0)	6.3 – 78.3
Follow	29.8 (15.3)	0 – 72.7	31.0 (14.7)	11.1 – 77.3	42.0 (14.0)	6.3 – 76.2
NoFollow	32.4 (10.0)	10.5 – 46.2	37.4 (11.5)	15.8 – 57.9	39.3 (14.2)	12.5 – 78.3
Compatible with Input	62.9 (16.3)	27.8 – 89.5	66.5 (15.7)	27.3 – 100	65.3 (15.7)	17.6 – 86.4
Follow	70.4 (10.0)	53.3 – 89.5	74.5 (13.1)	50.0 – 100	71.3 (10.5)	47.4 – 86.4
NoFollow	47.5 (15.9)	27.8 – 76.5	49.8 (15.7)	27.3 – 75.0	52.7 (50.0)	17.6 – 84.2
Compatible without Input	66.5 (17.3)	30.8 – 92.9	65.5 (16.4)	25.0 – 87.5	67.7 (13.8)	33.3 – 94.7
Follow	74.2 (13.2)	31.3 – 92.9	73.4 (11.2)	37.5 – 87.5	72.9 (10.6)	43.8 – 94.7
NoFollow	50.7 (13.8)	30.8 – 77.3	49.3 (13.3)	25.0 – 68.8	57.0 (13.8)	33.3 – 81.8

The statistical analyses for the consistent responses (see Appendix E) lead the same results as the previous analyses, except for one deviation. This deviation specifically concerns the Compatible without Input category of the Follow group's results. This effect might indeed have been caused by the inconsistent responses that were included in the previous analysis, though the difference in interpretation was minimal, like the delayed posttest's interpretation of a similar deviation.

Concerning the percentage of consistent responses found in the present study – 82.4% of the responses was consistent – it was interesting to note that the study of Brandt and colleagues (2021) found a very similar consistency percentage, namely 81%. Given the high degree of convergence between analyses across all responses on the one hand and only consistent responses on the other hand, the general discussion will be restricted to the analyses across all items, independent of their (in-)stability.

Discussion

From the outset of this experiment, there was one central question: What does the L1-L2 transfer effect of Dutch (L1) articles to German (L2) articles look like? Considering that this paper is based on the work of Brandt and colleagues (2021, submitted), their previous findings served as a basis for the expected effect in Dutch (L1) participants. Before an accurate comparison between results can be made, however, a few categories need to be defined first. Let us start with the categorical distinction between incompatible and compatible nouns. Many nouns in the German and Dutch languages are almost identical, such as '*Mund*' (in German) and '*mond*' (in Dutch; '*mouth*' in English). These similar nouns are called *cognates* (e.g. Lemhöfer, Spalek & Schriefers, 2008, Lemhöfer, Schriefers & Hanique, 2010). In addition, all Dutch and German nouns have an associated article that is either neuter (das_{neu} for German, het_{neu} for Dutch) or common gender (der_{masc}/die_{fem} for German, de_{com} for Dutch), These two aspects of Dutch and German made these languages perfect candidates for examining the L1-L2 transfer in this study. The exact reason is that the neuter words in one language usually are neuter words in the other language as well, and the same holds for common gender words. Words that have the same category of article (neuter or common) in both languages were called *compatible* nouns (Lemhöfer et al., 2008, 2010; Lemhöfer, Schriefers & Indefrey, 2014; Brandt, Schriefers & Lemhöfer, 2021, submitted), and the large majority of cognate nouns fall within the compatible category.

Crucially, however, there are exceptions to the rule of compatibility. There are some cognates that do not share this neuter or common category in both languages. In Dutch they might be neuter, while in German they are common, or vice versa. An example is the case of '*das Radio*' in German and '*de radio*' in Dutch ('*the radio*' in English); identical words that do not share their article category. These are called *incompatible* nouns. Furthermore, for the interpretation of the results it is imperative to know that compatible and incompatible cognates are indistinguishable without prior knowledge, and that there are far more compatible than incompatible cognates in German and Dutch (e.g. Lemhöfer et al., 2008, 2010). In general: Unless you know which words are the exception to the rule of compatibility, there is no way you can predict which cognate pairs are incompatible[10].

Getting that distinction out of the way, the expected L1-L2 transfer, or *mapping*, effect can now be described. The effect of mapping found by Brandt and colleagues (2021; see also Lemhöfer et al., 2008, 2010, 2014) in German (L1) speakers learning Dutch (L2), means that experienced Dutch (L2) speakers used their knowledge of German (L1) articles when encountering incompatible

10 One possible caveat here is the reported strategy that words that 'end in -e are always female' in German (such as '*die Kanone*' ('*het kanon'* in Dutch; '*the cannon'* in English), thus necessitating the use of the article die_{com} for these nouns. Although there is no evidence this strategy actually influenced the study's results, this was a strategy that multiple participants reported and could be expanded upon in future research.

nouns. This means that when a German (L1) speaker who already is experienced with Dutch (L2) encounters a cognate noun in Dutch, they tend to use their knowledge of German articles to provide an article for the noun. Since this strategy fails when encountering incompatible nouns, experienced German (L1) speakers make reliable errors when they provide an article for incompatible nouns (Lemhöfer et al., 2010).

This leads to another important categorical distinction: experienced and less experienced participants. The mapping effect as described above has conclusively been shown in experienced L2 speakers of Dutch, but there is only limited evidence of this effect in less experienced Dutch (L2) speakers (Brandt et al., submitted). In previous research, experience was measured by estimating how much exposure to Dutch (L2) a German (L1) native speaker had received prior to the experiment, by way of questionnaire and task-based assessments, or by selection criteria upfront (Brandt et al., 2021, submitted). As all high school students in the Netherlands were necessarily exposed to German in class (as this is required by law), but not all to the same degree, this method of estimation could not be accurately performed in the Dutch (L1) sample, as participants could not effectively be selected on previous experience, nor on the previously used 'immersion score' (Brandt et al., 2021, submitted, Mickan & Lemhöfer, 2020).

To make sure a distinction between experienced and less experienced participants was meaningful, the compatible cognates were used as a proxy of the effect: if participants indeed showed they were consistently using their Dutch (L1) article knowledge when prompted with a German (L2) noun at the start of the experiment, thus showing they used the mapping strategy, they were considered to be *experienced* participants. In contrast, if participants showed no such evidence, choosing their articles in a way indistinguishable from chance, then they were considered to be *inexperienced* (or less experienced) participants. Across all analyses, the resulting groups (referred to as *Follow* versus *NoFollow* groups) consistently showed different behavioural patterns, so indeed this way of dividing participants was meaningful. Therefore, their results will be discussed separately.

Lastly, the effects of learning from input were assessed, which was the main focus of previous experiments (Brandt, Schriefers and Lemhöfer, 2021, submitted). This brings us to the last important distinction: *input* or *no input*. This allows to see whether Dutch (L1) participants, too, could learn which article was appropriate for each noun from only one instance of correct input. For this purpose, the participants were exposed to one instance of native input of German articles (the participants' L2) for half of the nouns they encountered, after which they would have to produce the nouns once more. In this way, it was measured whether participants changed their use of articles after this input moment.

Time course of learning: learning by mere exposure

In Brandt and colleagues (submitted), the authors found an unexpected effect: in this study of inexperienced German (L1) participants, who showed no evidence of L1-L2 transfer (or mapping strategy) at the start of the experiment, started to show some evidence of L1-L2 transfer at the end of the experiment. More specifically, during the experiment (i.e. the '*dialogue*' part of the experiment), no clear evidence of participants using their L1 knowledge for their L2 was found, but in the posttest the participants seemingly started making consistent use of their German (L1) article knowledge to guide which articles they used in Dutch (L2). It is important to keep in mind that the previous study (Brandt et al., submitted) did not incorporate a time course analysis other than a contrast between the experimental dialogue game (or *incidental learning task*) and the posttest following the dialogue game. Instead of looking at effects over time, the analyses were done over *production moments*. For each noun, a total of four production moments were uttered by the participant. The first three were in relatively quick succession, separated only by a few trials per every word. The last was a production in the posttest, 15 minutes after the last trial of the dialogue game was completed.

This means that some words had their first three productions far removed from their posttest production, while others were still produced in the last part of the dialogue game, after which the posttest immediately followed. While this was fine for the analyses of input, it lacked temporal resolution in terms of design. Such temporal resolution was available in the present study by a slight change in methods which made it possible to look at when exactly the L1-L2 transfer effect took place in the inexperienced participants. The present experiment was made up of four separate blocks, and in each block a subset of the critical nouns were introduced. By looking at the percentage of article productions that were congruent with a mapping strategy per each separate block of the experiment, a time course was construed throughout the experiment. These congruent responses are referred to as *mapping responses*. In addition, to ensure there was no possible influence of correct article input for any given noun pair, only the first two production moments were used in this analysis, as half of the noun pairs received correct input of articles after the second production moment. So, the percentage of mapping responses of the first two production moments of all nouns was calculated per block of the experiment. As the posttest was different in set-up from the dialogue game (as it contained all words at once), the posttest was left out of the time course analysis.

Previous research in German (L1) participants only found an increase in mapping responses in the posttest compared to the rest of the experiment in inexperienced participants (Brandt et al., submitted), but did not find this effect in experienced participants (Brandt et al., 2021). This was, then, also the expectation for Dutch (L1) participants: an effect was only expected in the

inexperienced participants, not in the experienced participants.

Experienced participants

The majority of participants were part of the experienced category, who already showed evidence of using the mapping as a strategy at the start of the experimental session. In these participants, indeed, no increase in mapping responses over time was found. The percentage of mapping responses stayed stable throughout the experiment. In addition, it was clear that the experienced participants used this mapping strategy immediately in the first block of the experiment. This was concluded from the fact that the categories of incompatible and compatible nouns were treated as different by the group immediately. The percentage of mapping responses was significantly higher for the compatible category than for the incompatible category. When considering that, together with an absence of correct input, these two categories can only be distinguished when the participants already know at the start of the experiment that an incompatible noun is an exception to the rule of compatibility. Thus, the experienced group must have used their prior knowledge of German (L2) to immediately distinguish the compatible and incompatible categories. This is in agreement with the previous findings of Brandt and colleagues (2021), although it was striking to see such a consistent difference between compatible and incompatible nouns, as they were thought to be indistinguishable and had not yet been directly contrasted in previous research.

Inexperienced participants

The inexperienced group, which showed no evidence of a mapping strategy at the start of the experiment, generally confirmed the previous finding of Brandt and colleagues (submitted). The participants showed an increase in mapping responses over time, as was expected. However, interestingly, only the compatible nouns showed this increase while no change was observed for the incompatible nouns. Because compatible and incompatible cognates can only be disentangled by prior knowledge, as otherwise they were indistinguishable, this implies that even the inexperienced group must have had prior knowledge that only became available after the first block of the experiment. In fact, the first block showed no difference between incompatible and compatible nouns in terms of percentage of mapping responses, only in the second block the difference between these categories became apparent.

The second interesting finding was the time course of the effect: there was an increase in percentage of mapping responses selectively for the compatible category from the first to the second block. The third and fourth block were indistinguishable from the second block in terms of percentage of mapping responses. This meant that the prior knowledge that the inexperienced group

apparently did possess only became available after the first block, and that no further learning occurred afterwards. Before any strong conclusions are drawn, it is important to realise that individual differences were not taken into account in the analysis. It might well be that only some of the participants recovered their previous experience with German, but others did not recover or even receive enough previous experience to show any effect at all.

Conclusion

The current results support the findings of previous research, showing that Dutch (L1) speakers seem to use the mapping in the same way as their German (L1) speaking counterparts.. The experienced participants showed no change in percentage of mapping responses over time, as was found in Brandt and colleagues (2021), while the inexperienced participants did show an increase in percentage of mapping responses over time in the compatible category, as was found in Brandt and colleagues (submitted). This change was further examined in terms of its development over time and surprisingly there was only an effect of learning early on in the experiment, while no change was observed after this initial learning effect. In addition, the selective effect for the compatible nouns – and not for the incompatible nouns – suggests that previous experience was (partially) accountable for this effect, which seemingly only became available after some exposure to German articles.

Future research could focus on two aspects to shed more light on these effects. The first option is to look at the individual differences of (in)experienced participants, to see whether this can predict the learning effect or learning trajectory. It is expected that participants with less (or no) experience should have a harder time distinguishing between compatible and incompatible nouns, and should therefore show less difference in learning patterns between these categories, nor will it be limited to only one category. With some previous experience, exposure to German might make their prior knowledge more easily available which should then mirror the selective learning effect that was found in the inexperienced group in this study. The experienced participants should show no learning effects over time, as was found in this study as well, as they will be more familiar with both the rule of compatibility and its exceptions.

The other direction for future research is to use this analysis technique on inexperienced German (L1) speaking participants, as they will not have had the same amount of formal schooling and therefore should not differentiate between compatible and incompatible nouns at allh. If the participants did already have some experience – though perhaps limited – in Dutch (L2), a replication of the current effect is expected. As no time course analysis has of yet been performed on German (L1) speaking participants, it could be a valuable addition to future work. The same holds for a formal analysis of the compatible category. This category has been analysed before

(Brandt et al., submitted), but was not integrated in the methods in the same way as was done in the present study, so the results were limited in scope. Since the current study found large and consistent differences between compatible and incompatible categories for all analyses in the Dutch (L1) sample, there is a strong recommendation to look at both categories in future research.

Input-driven learning: learning implicitly from input

A second objective of the present paper was to see whether Dutch (L1) participants responded to native German input in the same way as German (L1) participants did to native Dutch input in previous studies (Brandt et al., 2021, submitted). This was a separate effort from the time course analyses above, as the time course analysis was completely novel and therefore exploratory in nature. In contrast, the input-driven learning analyses were meant to replicate the findings of Brandt and colleagues (2021). The analysis was not so much performed over the time course of the experiment. Instead, the input-driven learning that resulted from the multiple productions of each cognate pair in the study was targeted. Each cognate pair was produced a total of 4-5 times by each participant: three times throughout the dialogue game, once in the posttest after the game had ended, and for a minority of participants once more after six weeks in a delayed posttest. The delayed posttest will be discussed separately due to its different nature from the other analyses.

The input-driven learning was derived from the three productions within the dialogue game. The first two times a cognate was produced, with their article, served as a baseline: which article were participants inclined to use for this noun? Then, for half the words, native German input was provided for the cognates and their articles, in the form of a standard sentence. In this way, half of the words for each participants fell into the *input* category, and the other half fell into the *no input* category. After this (absence of) input, the participants had to produce the word once more within the dialogue game, which served as a measure of whether the participants had changed their article usage compared to the first two production moments. For each word category, the percentage of correct article responses was calculated at each production moment, which served as the dependent variable for this analysis.

The choice of *percentage of article responses* over *percentage of mapping responses* is an important change in dependent variable from the previous analysis. This change was necessary due to the nature of German as a target language instead of Dutch (as in previous studies, see Lemhöfer et al., 2008, 2010, 2014**;** Brandt, Schriefers & Lemhöfer, 2021, submitted). When usin

g Dutch as the target language, there were only two articles to take into account: de_{com} and het_{neu}. This meant that there were only two answer categories (*de* and *het*) which corresponded to their German counterparts (der_{masc}/die_{fem} and das_{neu}) for the compatible nouns. So, when providing an article for compatible cognates in Dutch, '*following the mapping*' would always lead to a correct

article. For incompatible cognates, following the mapping would inversely always result in an incorrect article. In this way, when Dutch is the target language, mapping responses and correct article responses are directly related, and the choice of dependent variable is up to personal preference.

For German as a target language, the choice of dependent variable was more impactful. Because German has two articles (*der*$_{masc}$, *die*$_{fem}$) corresponding to *de*$_{com}$ in Dutch, a response can be both a mapping response and incorrect in the case of compatible nouns (e.g. '*Der* Mund*', where '*Die Mund*' ('*de mond*' in Dutch; '*the mouth*' in English) would be correct), or not a mapping response and incorrect in the case of incompatible nouns (e.g. *'Der* Gefahr*', where '*Die Gefahr*' ('*het gevaar*' in Dutch; '*the danger*' in English) would be correct).

Since the goal of the input-based learning analyses was to show whether Dutch (L1) participants could learn at all from only one instance of correct article input from a native German speaker, and since a correct input-driven change within the same category (e.g. a correct change from *der* to *die* or vice versa) is also an indication of learning, the *percentage of correct article responses* was used as the dependent article for the input-driven learning analyses, as it was sensitive to every kind of input-driven learning. The new dependent variable still gave an indication of the strength of the mapping, as a stronger effect would mean more correct responses for compatible nouns and more incorrect responses would be expected for incompatible nouns.

The participant groups were identical to the groups defined in the time course analysis. For the experienced participants, the same results for the incompatible cognates were expected as previously found in the German (L1) sample (Brandt et al., 2021): initially two stable production moments, followed by an increase in correct responses after input. If there was no input, no change from the initial production moments was expected. Simply put: correct article usage should go up only after receiving input. The previous study did not examine the compatible category, but there was no reason to assume a priori that input should have a different effect on the compatible nouns.

Based on of previous research (Brandt et al., submitted), inexperienced participants' learning effects were expected to be smaller than the experienced group, but it is difficult to compare the current inexperienced group to the former. German (L1) speakers do not get the same amount of classroom exposure to Dutch (L2) as Dutch (L1) speakers get to German (L2). This means that Dutch (L1) speakers are likely to be more experienced than the German (L1) participants. Furthermore, the groups in the present study were defined differently from previous participant groups, and the experienced group was much larger than the inexperienced group. Because of this, the results of the inexperienced group in terms of learning from input were explorative in nature, to be expanded upon by future research.

Experienced participants

For the most part, the experienced participants behaved as expected. The percentage of correct articles for the compatible cognates especially was exactly conform expectations: an increase in correct article usage after receiving input. The effect was quite small, which was likely caused in part by a ceiling effect. The percentage of correct articles for the compatible cognates was already quite high at the first production moment, and there is not much to learn from input when you already correctly produced the article in the first place. The learning effect of input was not very stable, however, as the posttest showed a similar percentage of correct articles as in the first two production moments. The nouns that did not receive any input, showed no change over productions, also confirming the predictions.

Now, we turn to the incompatible cognates. There was indeed a replication of the expected input-driven learning effect found previously in German (L1) speakers (Brandt, Schriefers & Lemhöfer, 2021). This means that the experienced Dutch (L1) participants used more correct articles for the incompatible nouns after they were exposed to native German input, precisely as the German (L1) speakers did in previous research. This effect was, in line with the compatible nouns, not stable over time (as the posttest did not differ from the first two production moments), which meant that one instance of input did not seem adequate to induce a long-term learning effect.

There was also an unexpected finding. A learning effect was not only found for the incompatible cognates that received input, but also for those that *did not* receive any input. This effect was not found previously (Brandt et al., 2021), and as of yet there is no theoretical explanation in the literature. Still, the effect must have come from somewhere. The most likely explanation – though still speculative – is that it was a repetition effect. It could have, for example, stemmed from participants remembering more of the correct articles after each production moment. This would lead them to switch from an incorrect article to a correct article more often than vice versa, as the recollection of correct articles would accumulate over production moments. In turn, this would explain the effect, as for each production moment a relative increase in correct articles would be observed. Regrettably, the results of the experienced group are not sufficient to clearly see a pattern confirming a repetition effect, as only the posttest showed a detectable increase in percentage of correct articles. However, when combined with the results of the inexperienced group (see below), the explanation becomes more likely, as they showed a similar effect that started already before the posttest. At least, the results of the experienced group do not falsify the repetition effect hypothesis outright, as each production moment had a numerically higher percentage of correct article usage than the one before, even if this could not be detected by the formal statistical analysis.

When looking closer at the posttest's results, the suspicion arises that it might be the posttest (as an explicit test of article knowledge) caused the pattern observed in this postetst, rather than

being just a reflection of learning from input. The almost exact convergence of scores between the input and no input categories in both the compatible and incompatible categories seem to point in this correction: this kind of convergence is not seen in any other production moment, and unlikely to be an effect of input. The change in behavioural pattern could either be due to a later activation of what the participants had learned in the experiment (through an as of yet unknown mechanism), or there might be some methodological confound that elicited a different response pattern than the responses in the dialogue game. It is striking that the German (L1) participants did not show this convergence of compatible and incompatible categories in the posttest, and neither did they show the proposed repetition effect (Brandt et al., 2021). It is likely, then, that the classroom-based experience of the Dutch (L1) participants played some role in these new findings.

Inexperienced participants

The inexperienced participants showed quite different results from the experienced group. In contrast to the experienced participants, they showed no learning effect based on input at all. This was interesting, as it shows that some sort of familiarity with the target language is necessary before incidental learning from input is possible. Still, the inexperienced participants changed their behaviour throughout the experiment. They did not so much show a learning effect of input, but did improve in their percentage of correct articles over repetitions. In fact, for all categories, the participants gave most correct article responses at the third production moment – i.e. at the third time they produced a noun with its article. This kind of learning looked more like a repetition effect, rather than input-driven.

In the experienced participants, a similar effect could be seen in the incompatible cognates that did not receive input. This effect was limited to one category and only detectable in the posttest. For the inexperienced participants, neither was the case. The repetition effect was present in all categories of words, and was already demonstrably established in the third production moment, which was – in contrast to the posttest – methodologically identical to the first two production moments and not separated in time from them at all, as they were distributed throughout the experiment. This shows that even, or especially, less experienced participants retain more correct articles when confronted with a noun multiple times.

Although the specifics behind this effect are unclear, the same explanation might be given as for the experienced participants. Participants might remember the correct article for a noun at any production moment, which accumulates in more correct article responses over repetitions. Although this was not expected of the inexperienced group a priori, it is in line with the finding that they could in fact distinguish between incompatible and compatible cognates. This, in turn, was proof that they had access to some form of prior knowledge, at least at the group level.

Delayed posttest

The delayed posttest was an additional test that was sent to all participants by email, six weeks after the original testing session. It was similar in set-up to the posttest that was conducted 15-30 minutes after the dialogue game. The purpose for the delayed posttest was to see (1) whether the posttest was reliable over a longer period of time and (2) whether the input effects found during the experimental session in the lab were stable over time.

The delayed posttest can only be seen as an explorative extension of the other results, as only a small subset of participants actually responded to the email invitation. Another complication of this small participant group was that there were so few inexperienced participants that completed the delayed posttest, that the original distinction between the experienced and inexperienced groups had to be dropped. So, the delayed posttest's participant group is best interpreted as a different subgroup altogether, which was most similar to the experienced participants group in former analyses. Indeed, the group showed the same learning effects of input (when going from production moment 2 to production moment 3) in the compatible and incompatible categories as the experienced groups in the analyses reported above. The compatible cognates that did not receive any input showed no change in correct article usage over time, and the unexpected finding that incompatible cognates that did not receive input showed an increase of correct article usage in the posttest was also replicated for the subgroup that took part in the delayed posttest.

The delayed posttest did indeed show the same results as the posttest. This means that after six weeks participants showed the same scores (for interpretation's sake)as found in the initial posttest. Similarity between posttest and delayed posttest results implies that the measurement itself was stable over time, which is good in terms of reliability of the posttest and delayed posttest. However, this reliability was not necessarily expected, regarding the previously found results in the experienced participants. The results of the delayed posttest's analyses largely warrant the same interpretation as the experienced group's before. This was as expected for all categories but one: the incompatible cognates that did not receive any input.

As was found previously in the experienced group, the incompatible cognates that did not receive input showed an unexpected increase in correct article responses, most noticeably in the posttest. This was best explained by an 'effect of repetition', if not an effect driven by the changed set-up of the posttest. This explanation will further be referred to as the '*repetition hypothesis*'. Now we move onto the problem with this hypothesis. The problem is that there was no change in correct article usage after an additional production moment after six weeks (the delayed posttest). In contrast to the prediction of the repetition hypothesis (an increase in correct articles as production moments increase), no such further increase was observed in the delayed posttest. This means that

either the repetition effect must suddenly have become stable after the posttest, or the methodology of the (delayed) posttest did indeed drive this unexpected finding. The posttest and delayed posttest were very similar in set-up to one another, presenting a multiple choice of articles with their respective nouns that the participants had to choose from, rather than the participants generating the articles themselves in the context of a German sentence. This change in methodology might explain both the initial unexpected finding and the stability of this effect over a time of six weeks.

However, likely due to low power due to the lower amount of participants in the delayed posttest, it was difficult to interpret whether the answers were indeed as stable as proposed. There was no statistical difference between the delayed posttest and the second production moment, for example, while the posttest did differ from the second production moment. This is an indication that the results are not reliable enough to warrant strong conclusions, and the 'repetition hypothesis' might therefore not yet be convincingly falsified. As stated before, the delayed posttest's results are best regarded as explorative and further research is necessary for any definitive conclusions.

It is important to note that the findings in the delayed posttest's analyses, which were mostly drawn from participants of the 'experienced group', do not diminish the proposed repetition effect in the inexperienced participants. The inexperienced participants already showed an increase in correct article usage for all categories, independent of input, in the third production moment – before the posttest was administered. Thus, the best explanation for their results remains a repetition effect. Additionally it is important to keep in mind that only the 'incompatible words that did not receive input' showed problematic results in the delayed posttest. There is no reason to change the interpretation of any of the other categories (compatible with/without input or incompatible with input), as none of those categories showed an unexpected increase or decrease of correct article usage in either of the posttests.

An additional important note is that the German (L1) participants of previous research did not show any of the effects described above (Brandt et al., 2021). Not only was there no increase of correct article usage for the incompatible cognates that did not receive input in the posttest (so no reason to assume a 'repetition hypothesis'), but the initial learning effects of input were retained in the posttest and only collapsed in the delayed posttests. In that former study, the same reliability of results of posttest and delayed posttest was not observed, nor did the posttest seem to drive any of the effects in the experiment. So it is likely that previous classroom-based knowledge played some role in the unexpected behavioural pattern seemingly caused by the posttest and delayed posttest..

General discussion

The main purpose of the present study was to replicate Brandt and colleagues' (2021) study with a change of population. Instead of native German (L1) speaking participants tested in Dutch

(L2), the participants were now native Dutch (L1) speaking participants tested in German (L2). The most critical finding to replicate was the effect of a native speaker's input on the production of articles. The relatively experienced participants, which made up the largest subgroup of the experiment, did indeed show a clear effect of input. This implies that the majority of the Dutch (L1) native population indeed is capable of learning the correct articles from a single instance of a German (L1) native speaker's input.

At the start of the present study, this was not yet proven beyond doubt. As German has an extra article ($der_{masc}/die_{fem}/das_{neu}$) compared to Dutch ($de_{com}/het_{neu}$), as well as an intricate case system that Dutch lacks (Baten, 2013), it was a real possibility that Dutch (L1) participants would have been unable to learn the mapping as readily as German (L1) participants. The simple fact that experienced participants could be identified on the basis of using the mapping, in combination with the finding that even inexperienced participants adopted a mapping bias only after relatively brief exposure to the experimental stimuli, conclusively proved that Dutch (L1) speakers can use the mapping just as well as German (L1) speakers.

Because Dutch (L1) speakers do in fact learn to use the mapping, the incidental learning paradigm seems valid to use on a Dutch (L1) population, but with the caveat that less experienced participants might not show the same behavioural patterns, as they did not show any change due to native German (L1) input. It is unclear why less experienced participants did not respond to this input. Perhaps these participants were more rigid in their responses , and perhaps more like the '*default article users*' as identified in Brandt and colleagues' study (submitted), who showed a tendency to predominantly use only one article. However, this is inconsistent with the finding that the inexperienced participants in fact did learn both over time and by what looked like a repetition effect. Both of these effects were unexpected.

The repetition effect was especially surprising as there was no similar previous finding, although the learning of the L1-L2 transfer over time was found previously in one of the studies of Brandt and colleagues (submitted) in inexperienced German L1 participants. Perhaps the most intriguing finding in the present paper was that the L1-L2 transfer, or *mapping* of Dutch (L1) knowledge on German (L2) articles, did not develop linearly throughout the experiment, but was rather fully realised within the first quarter of the experiment and did not change any more in the remaining parts of the experiment. This pattern only occurred for the compatible cognates, while the pattern of results for the incompatible cognates did not change at all. This is in clear contrast to the German (L1) participants in Brandt and colleagues (submitted) who also showed an increase in correct article accuracy for compatible nouns, but in addition a decrease in accuracy for incompatible nouns.

A main difference between the German L1 population tested by Brandt and colleagues

(submitted) and the Dutch (L1) population tested here, concerns their prior classroom based knowledge. Only the Dutch (L1) group had received systematic classroom training in the relevant L2 (in their case German), as this is required by Dutch law. In contrast, the German participants in Brandt and colleagues' (2021) study did have prior knowledge of Dutch, but mostly by being immersed in a Dutch L2 language environment. While none of the Dutch L1 participants lived, worked or studied in Germany, the German L1 participants in Brandt and colleagues (submitted) almost all lived in the Netherlands, and all of them went to a Dutch university on a daily basis. The Dutch (L1) participants also seemed more keenly aware of the possibility that they could learn the correct articles from the native stimuli and input than the German (L1) participants. This suspicion was supported by the interviews regarding the debriefing in the present study compared to the previous study of Brandt and colleagues (2021). It remains unclear whether this actually had an effect on the results, but it (again) points to the possibility that the potential differences in L2-knowledge of the Dutch (L1) and German (L1) groups might lead to a different approach in learning from new stimuli.

This in turn raises the question: How does classroom-based L2 learning, in contrast to an immersive L2 context, shape the further learning of this L2 outside a classroom setting? Both the results for the experienced and the inexperienced groups support that the kind of prior knowledge shaped the form of learning in the experiment. The experienced group, first of all, clearly showed that they differentiated between the compatible and incompatible nouns from the very start of the experiment. This clearly indicates that they used some prior knowledge for producing the articles. At this moment it remains unclear whether this differentiates the Dutch (L1) participants from their German (L1) counterparts, as the compatible-incompatible noun contrast was not fully measured in previous research on German (L1) speakers. What did separate them, was that the effect of input that was found in both the German (L1) group in previous research (Brandt et al., 2021) and the experienced Dutch (L1) group in the present study disappeared in the posttest selectively for the Dutch (L1) speaking group. The German (L1) participants retained this effect until at least after the posttest (Brandt, Schriefers & Lemhöfer, 2021).

In the Dutch (L1) inexperienced group, the main finding in support of prior knowledge shaping the learning effect was the speed with which the learning of the mapping was established. After the first quarter of the experiment, there was no further increase in the usage of mapping. It is unknown how the learning of the mapping develops over time in German (L1) speakers, as only the effect itself was detected in previous research and not its development over time (Brandt, Schriefers and Lemhöfer, submitted). Still, it seems that the speed of the observed learning of the mapping was only achieved due to the previous classroom experience of the Dutch (L1) participants. This idea receives further support by the difference between incompatible and compatible cognates after

showing some learning of the mapping.

Given the fast time inexperienced Dutch (L1) participants learn the mapping, as well as the different behavioural patterns for the experienced Dutch (L1) and German (L1) participants, prior classroom based learning seems to shape the learning behaviour of participants in a non-trivial way in the future. Especially striking was the divergence of posttest results between experienced German (L1) (see Brandt, Schriefers & Lemhöfer, 2021) and experienced Dutch (L1) participants, which showed in a direct comparison that adult German (L1) and Dutch (L1) participants indeed learn L2 gender differently in an incidental learning context. With the data at hand so far, the most likely explanation of this difference is the classroom based nature of the prior knowledge of the Dutch (L1) participants which was not present in German (L1) participants.

Classroom-based language learning already has received quite some attention (e.g. Norris & Ortega, 2000; Spada & Tomita, 2010; Spada, 2011; Miller & Pan, 2012; Belliveau & Kim, 2013; Plonsky & Brown, 2015; Brown, 2016; Brown, Plonsky & Teimouri, 2018), and most Europeans report learning their L2 in a formal setting (TNS Opinion & Social, 2012). However, there has not yet been any other evidence – to the knowledge of the author – of prior classroom-based learning affecting the way languages are picked up in an incidental context (except perhaps in the domain of phonology, see Best & Tyler, 2007). Or how classroom-based might influence results of language research. So for future research, this interplay between previous classroom learning and incidental language learning seems like an interesting direction for follow-up in future research. In any case, the current evidence points to the conclusion that 'learning language in the wild' is influenced by and likely influences classroom-learning of language.

Recommendations for Future Research

Most recommendations for future research, including problems to solve as well as the most interesting follow-up questions, coalesce into a few distinct solutions and suggestions for future research design. So, instead of repeating over and over what the right solution is for each problem, this section provides an overview of the most important recommendations

First, the present study is the first study that employed a time course analysis using the '*incidental learning paradigm*' (or: '*L2 dialogue game*'), as well as a full analysis (including a contrast between an input and a no input condition) of compatible cognates for both experienced and inexperienced participants. As such, no data on the temporal development of the learning of the mapping in German (L1) participants exist. Neither is there any full comparison of input versus no input conditions of German (L1) participants regarding compatible cognates. This made comparisons between the present study's participant groups and those of previous research difficult. Due to its small size and the changed selection process, especially the Dutch (L1) inexperienced

group was difficult to compare with previous research (Brandt, Schriefers & Lemhöfer, submitted). Because of this lack of direct comparison and data, a lot can be learned from either replicating the present study with German (L1) participants or, perhaps better still, running an experiment with a Dutch (L1) and a German (L1) group and comparing their results directly.

Secondly, all analyses performed up until this point were performed at the group level, both in the present study and in previous research (Lemhöfer, Spalek & Schriefers, 2008; Lemhöfer, Schriefers & Hanique, 2010; Lemhöfer, Schriefers & Indefrey, 2014; Brandt, Schriefers & Lemhöfer, 2021, submitted). There are still some open questions regarding the individual differences in how participants pick up a different language outside of the classroom, or why some participants immediately show evidence of mapping while others do not. Especially since both previous research has shown that (individual) experience changes the manner in which participants pick up languages 'in the wild' (Brandt, Lemhöfer & Schriefers, 2021, submitted) and the current evidence backs this up, research on individual differences would be a great way to learn more on this topic.

Therefore, a careful study of individual differences in prior (classroom) experience should be a fruitful endeavor for future research. The L2 dialogue game paradigm lends itself very well to investigating individual differences, as each participant has at least 3 measurement points for each target word, and 96 words were tested in total in the current study (half of which were incompatible words). This amounts to a rich dataset in which individual learning trajectories can be detected with relative ease. The main remaining challenge is to find a good way of measuring prior knowledge and experience, for which a test battery could be used (as was partly implemented in Brandt and colleagues' earlier studies (2021, submitted)) and be formally analysed. If the proposed repetition effect (as found especially in the inexperienced group of the present study) were taken into account as well in future research, the amount of measurement points for each word could be extended by simply adding more production moments into the design. This might lead to more complexities in list construction and a longer duration of the experimental procedure, but there is no reason to assume that an addition of measurement moments will otherwise have a negative effect on the results.

Last is the posttest, which showed a disappearance of the positive effect of native input on correct article accuracy in the experienced Dutch (L1) sample of the present study, while in the German (L1) sample of previous research the positive effect of native input was still clearly visible in the posttest (Brandt and colleagues, 2021). The same held for the delayed posttest, which could be interpreted relatively clearly in earlier studies (Brandt, Schriefers & Lemhöfer, 2021, submitted) but was difficult to interpret in the present study.

There are two main explanatory factors for these differences in behaviour: the method of the

posttests that was different from the dialogue game's production moments, and the difference in prior knowledge between the samples (classroom based (Dutch) vs. immersion based (German)). These explanations are not mutually exclusive, but the method of the posttest is easiest to manipulate. By changing the method of the posttest to be more like the other production moments and testing a sample of Dutch (L1) speakers again, it can be seen whether the method had any effect on the behaviour at all. The easiest way to change the method is by simply asking participants to generate their own articles for every noun in the posttest (either spoken or written), perhaps in a sentence context, instead of letting them choose from a prompt on the screen which article is correct. If the results change, this shows that the methods did indeed impact the results of the study, which leaves the question why it seemed to selectively impact Dutch (L1) speakers. If the results are the same as in the present study, it implies that somehow the effect is driven solely by the change of sample, pointing to prior classroom experience as the driving factor.

Concluding remarks

The current experiment has conclusively shown that the experimental paradigm as construed by Brandt and colleagues (2021) can effectively be used in other participant groups, like in a Dutch population. In addition, it showed that the method is open for experimental tweaks, such as increasing the amount of compatible nouns and also examining them on its own. A time course analysis analysis over the experimental block was also successfully performed on both experienced and inexperienced participants, and in the latter group showed the unexpected finding that even these participants could use prior classroom knowledge but only after a first exposure to the experimental stimuli. Whether this also holds for German participants is unclear, but can easily be found out in further research by including a time course analysis of both compatible and incompatible nouns, or by focussing on individual differences.

More surprising findings were that classroom-based learning (e.g. Norris & Ortega, 2000; Spada & Tomita, 2010; Spada, 2011; Miller & Pan, 2012; Belliveau & Kim, 2013; Plonsky & Brown, 2015; Brown, 2016; Brown, Plonsky & Teimouri, 2018) seemed to make participants behave differently in the experimental context when compared to the original German (L1) findings (Brandt, Schriefers & Lemhöfer, 2021), not just because of their prior knowledge attained in the classroom, but also through an interplay between the incidental learning paradigm and the prior classroom-based knowledge. These findings give an opening for further research, specifically when focusing on individual differences. Inexperienced participants seemed to show some form of a repetition effect across production moments, unlike the participants in earlier research (Brandt, Schriefers & Lemhöfer, 2021, submitted). Even though there was an indication for a partial repetition effect in the experienced group, the evidence was so limited that no clear conclusion

could be drawn.

In general, most Dutch (L1) participants did learn from only one instance of input, but this learning did not persist into the posttest, with the latter finding being in contrast to their German (L1) counterparts. This is good news for the incidental learning paradigm, as it can indeed be generalised to other populations. Perhaps some tweaks to the design are warranted, at least to confirm whether the posttest and delayed posttest are working as intended, but the base of the paradigm seems solid for producing reliable effects for item-specific learning. The information that comes from the experimental set-up is quite rich, allowing for both time course, within- and between-subjects analyses, and lends itself well to a transition into research focusing on individual differences. I hope this paper has shown that the possibilities within this paradigm are numerous and should be employed in the future of incidental L2 learning research.

References

Antón-Méndez, I. (2011). Whose? L2-English speakers' possessive pronoun gender errors. *Bilingualism: Language and Cognition*, *14*(3), 318-331.

Audacity Team (2017). Audacity®: Free Audio Editor and Recorder [Computer software]. Version 2.2.1. Retrieved from https://audacityteam.org/.

Baten, K. (2013). *The acquisition of the German case system by foreign language learners* (Vol. 2). John Benjamins Publishing.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (Release 2) [CD-ROM]*. Philadelphia, PA: Linguistic Data Consortium, Unisersity of Pennsylvania.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The development of speech perception: The transition from speech sounds to spoken words*, *167*(224), 233-277.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O.-S. Bohn (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13-34). Amsterdam: John Benjamins.

Brandt, A. C., Schriefers, H., & Lemhöfer, K. (2021). A laboratory study of naturalistic second language learning: Acquiring grammatical gender from simple dialogue. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Brandt, A. C., Schriefers, H., & Lemhöfer, K. (2022). *Incidental Acquisition of Grammatical Gender in Beginning Second Language Learners* [Manuscript submitted for publication]. Donders Institute for Brain, Cognition and Behaviour, Radboud University.

Brown, B. (2016). Other things. In *Other Things*. University of Chicago Press.

Brown, A. V., Plonsky, L., & Teimouri, Y. (2018). The use of course grades as metrics in L2 research: A systematic review. *Foreign Language Annals*, *51*(4), 763-778.

Belliveau, G., & Kim, W. (2013). Drama in L2 learning: A research synthesis. *Scenario: A Journal of Performative Teaching, Learning, Research*, *7*(2), 7-27.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*(2), B13-B25.

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied linguistics*, *27*(2), 164-194.

Ellis, N. C. (2011). Implicit and explicit SLA and their interface. *Implicit and explicit language learning: Conditions, processes, and knowledge in SLA and bilingualism*, *35*, 47.

Embarcadero Technologies, Inc. (2013). Delphi®: RAD Studio [Computer software]. Version XE5 Update 2. Retrieved from https://www.embarcadero.com/products/delphi.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London, UK: Sage Publications Ltd.

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second language research*, *29*(3), 311-343.

Gooskens, C., Kürschner, S., & Van Bezooijen, R. (2011). Intelligibility of standard German and Low German to speakers of Dutch. *Dialectologia: revista electrònica*, 35-63.

Hinskens, F., & Taeldeman, J. (Eds.). (2013). *Dutch* (Vol. 30). Walter de Gruyter.

Hopp, H. (2010). Ultimate attainment in L2 inflection: Performance similarities between non-native and native speakers. *Lingua*, *120*(4), 901-931.

Hulstijn, J. (2003). Incidental and Intentional Learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition*, (pp. 349-381). London, UK: Blackwell Publishing Ltd.

Klein, W., & Perdue, C. (1997). The Basic Variety (or: Couldn't natural languages be much simpler?). *Second language research*, *13*(4), 301-347.

Ladefoged, P., & Maddieson, I. (1998). The sounds of the world's languages. *Language*, *74*(2), 374-376.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior research methods*, *44*(2), 325-343.

Lemhöfer, K., Schriefers, H., & Hanique, I. (2010). Native language effects in learning second-language grammatical gender: A training study. *Acta Psychologica*, *135*(2), 150-158.

Lemhöfer, K., Schriefers, H., & Indefrey, P. (2014). Idiosyncratic grammars: Syntactic processing in second language comprehension uses subjective feature representations. *Journal of Cognitive Neuroscience*, *26*(7), 1428-1444.

Lemhöfer, K., Spalek, K., & Schriefers, H. (2008). Cross-language effects of grammatical gender in bilingual word recognition and production. *Journal of Memory and Language*, *59*(3), 312-330.

Mickan, A., & Lemhöfer, K. (2020). Tracking syntactic conflict between languages over the course of L2 acquisition: A cross-sectional event-related potential study. *Journal of Cognitive Neuroscience*, *32*(5), 822-846.

Miller, P. C., & Pan, W. (2012). Recasts in the L2 classroom: A meta-analytic review. *International Journal of Educational Research*, *56*, 48-59.

Neurobehavioural Systems, Inc. (2017). Presentation® [Computer software]. Version 18.0. Retrieved from: https://neurobs.com/.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language learning*, *50*(3), 417-528.

Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta-analyses of L2 research (Or why 18 meta-analyses of feedback have different results). *Second Language Research*, *31*(2), 267-278.

Spada, N. (2011). Beyond form-focused instruction: Reflections on past, present and future research. *Language Teaching*, *44*(2), 225-236.

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language learning*, *60*(2), 263-308.

The Document Foundation (2017). LibreOffice Calc [Computer software]. Version 5.4.0.1.

TNS Opinion & Social (2012). Media use in the European Union. *Standard Eurobarometer, 78.*

Tyler, M. D. (2019). PAM-L2 and Phonological Category Acquisition in the Foreign Language Classroom. In Anne Mette Nyvad, Michaela Hejná, Anders Højen, Anna Bothe Jespersen & Mette Hjortshøj Sørensen (Eds.)*, A Sound Approach to Language Matters – In Honor of Ocke-Schwen Bohn*, (pp. 607-630). Dept. of English, School of Communication & Culture, Aarhus University.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Volkswagen Nederland. (2011, September 30). *Volkswagen "oud vrouwtje" commercial* [Video]. Youtube. https://www.youtube.com/watch?v=aCbQoqtMjL8

White, L. (2003). Fossilization in steady state L2 grammars: Persistent problems with inflectional morphology. *Bilingualism: language and cognition*, *6*(2), 129-141.

Qualtrics (2017). Qualtrics XM Platform [Online survey software]. Retrieved in 2017 from https://qualtrics.com/.

Appendices

Appendix A. Participants' exposure to regional languages in the Netherlands

The region of Nijmegen in the Netherlands is situated quite close to the German border, and is surrounded by multiple dialects and languages within these border regions, the most common of which were expected to be Limburgian ('*Limburgs*' in Dutch; '*Limbörgs/Lèmbörgs*' in Limburgian) and West Low German ('*Nedersaksisch*' in Dutch; '*Niedersächsisch*' in German). As previous research has found that residents of these border-areas might have more proficiency in German (Gooskens, Kürschner & Van Bezooijen, 2011), the participants were asked whether they had experience with these languages or other Dutch-related languages or dialects, which might share similarities with German not present in standard Dutch (Hinskens & Taeldeman, 2013). Three participants reported speaking either Limburgian or West Low German, while 26 reported they have encountered either or both in their direct environment. Furthermore, 8 participants reported speaking a different dialect (or Dutch-like language) themselves, namely Brabants (5), Rotterdams (1), Flemish (1; '*Vlaams*' in Dutch), Zeeuws (1) and Frisian (1; '*Fries*' in Dutch; '*Frysk*' in Frisian), while two more participants encountered Frisian in their direct environment.

Appendix B – Complete list of Stimuli Words

Table B1. *All German incompatible cognate nouns used in the experiment and their translations in Dutch and English.*

German	Dutch	English translation
Der Abfall	Het afval	*The garbage*
Der Anker	Het anker	*The anchor*
Der Balkon	Het balkon	*The balcony*
Der Chor	Het koor	*The choir*
Der Deckel	Het deksel	*The lid*
Der Hirsch	Het hert	*The deer*
Der Kanal	Het kanaal	*The channel*
Der Kompass	Het kompas	*The compass*
Der Park	Het park	*The park*
Der Plan	Het plan	*The plan*
Der Sand	Het zand	*The sand*
Der Schaum	Het schuim	*The foam*
Der Schirm	Het scherm	*The screen*
Der Speck	Het spek	*The bacon*
Der Strand	Het strand	*The beach*
Der Zirkus	Het circus	*The circus*
Die Adresse	Het adres	*The address*
Die Gardine	Het gordijn	*The curtain*
Die Gefahr	Het gevaar	*The danger*
Die Kanone	Het kanon	*The cannon*
Die Kreide	Het krijt	*The chalk*
Die Leiche	Het lijk	*The corpse*

Die Maske	Het masker	*The mask*
Die Nummer	Het nummer	*The number*
Die Orgel	Het orgel	*The organ*
Die Palette	Het palet	*The palette*
Die Pistole	Het pistool	*The pistol*
Die Schnur	Het snoer	*The cord*
Die Schrift	Het schrift	*The journal*
Die Terrasse	Het terras	*The terrace*
Die Toilette	Het toilet	*The toilet*
Die Uniform	Het uniform	*The uniform*
Das Armband	De armband	*The bracelet*
Das Auto	De auto	*The car*
Das Cello	De cello	*The cello*
Das Cockpit	De cockpit	*The cockpit*
Das Foto	De foto	*The photo*
Das Handtuch	De handdoek	*The towel*
Das Horn	De hoorn	*The horn*
Das Kanu	De kano	*The canoe*
Das Knie	De knie	*The knee*
Das Öl	De olie	*The oil*
Das Pony	De pony	*The pony*
Das Puzzel	De puzzel	*The puzzle*
Das Sofa	De sofa	*The sofa*
Das Studio	De studio	*The studio*
Das Taxi	De taxi	*The taxi*

Das Zebra	De zebra	*The zebra*

Table B2. *All German compatible cognate nouns used in the experiment and their translations in Dutch and English.*

German	Dutch	English translation
Der Arm	De arm	*The arm*
Der Baum	De boom	*The tree*
Der Berg	De berg	*The mountain*
Der Brief	De brief	*The letter*
Der Finger	De vinger	*The finger*
Der Fisch	De vis	*The fish*
Der Hund	De hond	*The dog*
Der Kaffee	De koffie	*The coffee*
Der Käse	De kaas	*The cheese*
Der Mund	De mond	*The mouth*
Der Stein	De steen	*The stone*
Der Stempel	De stempel	*The stamp*
Der Stern	De ster	*The star*
Der Stuhl	De stoel	*The chair*
Der Vogel	De vogel	*The bird*
Der Wein	De wijn	*The wine*
Die Banane	De banaan	*The banana*
Die Bank	De bank	*The couch*
Die Blume	De bloem	*The flower*
Die Flasche	De fles	*The bottle*

Die Hand	De hand	*The hand*
Die Katze	De kat	*The cat*
Die Küche	De keuken	*The kitchen*
Die Kirche	De kerk	*The church*
Die Mauer	De muur	*The wall*
Die Nase	De neus	*The nose*
Die Treppe	De trap	*The stairs*
Die Trommel	De trommel	*The drum*
Die Trompete	De trompet	*The trumpet*
Die Tür	De deur	*The door*
Die Wolke	De wolk	*The cloud*
Die Zigarette	De sigaret	*The cigarette*
Das Bett	Het bed	*The bed*
Das Blatt	Het blad	*The leaf*
Das Brot	Het brood	*The bread*
Das Dach	Het dak	*The roof*
Das Ei	Het ei	*The egg*
Das Feuer	Het vuur	*The fire*
Das Geld	Het geld	*The money*
Das Gesicht	Het gezicht	*The face*
Das Gewehr	Het geweer	*The rifle*
Das Gewicht	Het gewicht	*The weight*
Das Glas	Het glas	*The glass*
Das Gras	Het gras	*The glass*
Das Kloster	Het klooster	*The monastary*

Das Kreuz	Het kruis	*The cross*
Das Orchester	Het orkest	*The orchestra*
Das Pferd	Het paard	*The horse*

Table B3. *All German incompatible noncognate filler nouns used in the experiment and their translations in Dutch and English.*

German	Dutch	English translation
Der Bahnhoff	Het treinstation	*The train station*
Der Bleistift	Het potlood	*The pencil*
Der Flughafen	Het vliegveld	*The airport*
Der Knochen	Het bot	*The bone*
Der Kopf	Het hoofd	*The head*
Der Körper	Het lichaam	*The body*
Der Platz	Het plein	*The (town) square*
Der Schreibtisch	Het bureau	*The desk*
Der Teller	Het bord	*The plate*
Die Decke	Het plafond	*The ceiling*
Die Dose	Het blik	*The can*
Die Insel	Het eiland	*The island*
Die Narbe	Het litteken	*The scar*
Die Uhr	Het horloge	*The watch*
Das Brett	De plank	*The plank*
Das Fahrrad	De fiets	*The bicycle*
Das Gemüse	De groente	*The vegetable*
Das Gepäck	De bagage	*The baggage*

Das Klavier	De piano	*The piano*
Das Rohr	De pijp	*The pipe*
Das Streichholz	De lucifer	*The match*
Das Tor	De poort	*The gate*

Table B4. *All German incompatible noncognate filler nouns used in the experiment and their translations in Dutch and English.*

German	Dutch	English translation
Der Korb	De mand	*The basket*
Der Pilz	De paddestoel	*The mushroom*
Der Regenschirm	De paraplu	*The umbrella*
Der Schrank	De kast	*The closet*
Der Stiefel	De laars	*The boot*
Der Tisch	De tafel	*The table*
Der Tropfen	De druppel	*The drop*
Der Zug	De trein	*The train*
Die Axt	De bijl	*The axe*
Die Gabel	De vork	*The fork*
Die Höhle	De grot	*The cave*
Die Kartoffel	De aardappel	*The potato*
Die Scheune	De schuur	*The shed*
Die Schwelle	De drempel	*The speed bump*
Die Zeitung	De krant	*The newspaper*
Das Fenster	Het raam	*The window*
Das Früshstück	Het ontbijt	*The breakfast*
Das Gehirn	Het brein	*The brain*

Das Herz	Het hart	*The heart*
Das Loch	Het gat	*The hole*
Das Rad	Het wiel	*The wheel*
Das Segel	Het zeil	*The sail*
Das Ziel	Het doel	*The target*

Appendix C – Subgroups and Chi-Squares

In the Netherlands, learning the basics of the German language is a mandatory part of high school education. This means that every participant in the present experiment must have had some experience with German previously in their life. Because of this, it was likely that at least some participants would use their Dutch (L1) article knowledge to predict the correct German (L2) article, as was established in previous studies to be the case for experienced German (L1) speakers learning Dutch (L2). This behaviour resulted from the proposed L1-L2 transfer effect, and was referred to as *mapping* in the main text of this paper. To see whether Dutch (L1) indeed showed this mapping behaviour at the start of the experiment, the articles of the participants' first block were subject to an extra analysis. This analysis consisted of a chi-squares analysis (Field, 2013), where a predetermined *expected value* was compared to the participants' responses, or the *true value*.

Since incompatible nouns were explicitly included in the experiment to induce mistakes in the participants' behaviour (Lemhöfer, Schriefers & Hanique, 2010), and participants with more experience were both expected to make more of these errors *and* to know more of these words and their articles, the incompatible nouns were unsuitable for this analysis. The compatible nouns, in contrast, were straightforward to interpret in terms of mapping responses: experienced participants were expected to make less errors and make more use of the mapping. Only the first block was assessed because Brandt, Schriefers and Lemhöfer (submitted) found evidence of inexperienced participants learning to answer according to the mapping solely due to exposure to the experiment. Due to the possible effect of mere exposure to the stimuli, it was necessary to focus solely on the productions where a minimal amount of stimuli had been presented: the first block (or first quarter) of the experiment.

Additionally, only the productions *before* any explicit input was received were analysed, as explicit native input of the correct article was thought to effect the participant's behaviour. As this entire analysis was meant to provide a baseline for mapping knowledge in the experiment, it would have been counterproductive to include productions after the experimental manipulation. So: only the first two production moments in the first block[11] of each participant were assessed in the analysis.

Having defined the exact productions to be analysed, the next question is how to define the *expected value*. The expected value was the response pattern that was expected from completely naive participants. A priori, it was decided that naive participants would likely have no other strategy than to guess randomly which article would be correct for any given noun, given that a

11 All decisions were made based on this first block analysis, and this paper does not report on any of the other three blocks. However, the same chi-square analyses for each participant's other blocks do remain available upon requests, as they have been readily calculated. They were deemed too uninformative for the present paper to include, but if at all relevant for future research, they can still be provided.

naive participant per definition would have no prior knowledge to build on. So, the expected value was defined as 'the amount of mapping responses a randomly guessing (naive) participant would give in the first two productions of all compatible cognates in the first experimental block'. The expected value was, then, a simple value of one variable: an expected amount of mapping responses.

However, there was still one problem to solve before analysis could begin. The participants could encounter one out of a total of four[12] experimental lists in the first block. These lists were distributed among participants equally. Every list consisted of a different amount of compatible words with a specific article. Because each article had a different random chance for having a corresponding mapping response, each of these four lists had a separate expected value of mapping responses. Now having defined what the task was – calculating an expected amount of mapping responses for each experimental list – we could move on to calculating these values.

For these calculations, it was necessary to know for each article how likely they were to elicit a mapping response when choosing an article at random. Luckily, this calculation was quite straightforward, as only the compatible nouns were considered. There were two possible values of an article: 'following the mapping (*a mapping response*; 1)' or 'not following the mapping (*no mapping response*; 0)'. As a shorthand, these responses are referred to respectively as *mapping-congruent* (for a *mapping response*; 1) and *mapping-incongruent* (for *no mapping response*; 0) responses. Since participants answered in German, not Dutch, there were a total of three possible responses: 'der_{masc}', 'die_{fem}' or 'das_{neu}'. Then, to figure out whether a response is mapping-congruent or not, we turn to the Dutch articles. Whenever a noun's correct article is 'de_{com}' in Dutch, the mapping-congruent response is either 'der_{masc}' or 'die_{fem}' in German. In contrast, whenever a noun's correct article is 'het_{neu}' in Dutch, the mapping-congruent answer is 'das_{neu}' in German. So, while a *de*-word has *two* mapping-congruent responses in German, a *het*-word only has *one*. So, when a participant would be naively guessing, responses to a *der*-word or a *die*-word would be mapping-congruent in 2/3rd of all possible responses, but a response to a *das*-word would be mapping-congruent for only 1/3rd of all possible responses.

This way, an expected value for each list can be calculated. Each *der*-word and *die*-word got a value of 2/3 or 0.67, while each *das*-word got a value of 1/3 or 0.33. Over all 24 compatible words in the list these values were summed. This total value was, then, the total amount of expected responses in the first block for each participant, which was what was defined as the expected value. The expected value for mapping-congruent responses was, however, not quite enough for the chi-square analysis to be completed. A second expected value is also necessary to be defined: the expected value of incongruent-responses. This was calculated in a the same general way as before,

12 If all blocks are considered, not just the first, 8 total experimental lists were used in the experiment.

with the only difference that the values per word were reversed. So for the expected incongruent value, the *der*- and *die*-words got a value of 1/3 or 0.33 and the *das*-words got a value of 2/3 or 0.67.

There was still one last hurdle to jump before the chi-square analysis could be performed. Sometimes an article or word was unintelligible, becoming a missing value, which posed an issue for the expected values: Does an unintelligible (or otherwise missing) response count as a mapping-incongruent response? The answer was no, because allowing unintelligible responses to be counted as mapping-incongruent responses would unnecessarily inflate the amount of participants who were thought to have *no knowledge* of the mapping. This followed from the expected values remaining unchanged, which would then set a higher threshold of mapping-congruent responses. It was deemed more important to include all participants who showed any evidence of already using their Dutch (L1) article knowledge when speaking German (L2). So, to circumvent this problem, all intelligible (or otherwise missing) trials were deleted from both the expected values and the true values, and were subtracted from the total amount of words. A minimum of 18 productions was required (a maximum of 25% missing utterances) to be included in this analysis, which all participants adhered to, except one that would have been excluded for other reasons as well.

Now, we can move on to the true values. In itself the true value was not difficult to define: the actual responses of participants were coded for mapping-congruent responses (1) or mapping-incongruent responses (0) and counted together. This was done by hand by the first author based on the audio recordings. The resulting frequency of mapping-congruent responses was the true value of the participant. The frequency of mapping-incongruent responses, too, was necessary for the analyses, so in total we now have the four values for each participant: the *congruent expected value* (*congruent EV*); the *incongruent expected value* (*incongruent EV*); the *congruent true value* (*congruent TV*); and *incongruent true value* (*incongruent TV*) With expected and true values in hand, the chi-square analyses were performed separately for each participant. For each analysis, the *congruent EV*s was first subtracted from the *congruent TV*. This difference (the *congruent difference*) was then squared. Then, the *incongruent EV* was similarly subtracted from the *incongruent TV*, of which the resulting difference (the *incongruent difference*) was also squared.. The squared *congruent difference* was then divided by the *congruent EV*. The squared *incongruent difference*was then likewise divided by the *incongruent EV*. Finally, adding these values together created the *chi-square value*. This value was contrasted with a value on the chi-square distribution corresponding to a significance value of $p = 0.05$ (Field, 2013). If the found *chi-square value* for a participant was higher than the distribution's defined value, this would depict a significantly larger deviation from the originally defined expected value, which was a reflection of a randomly guessing participant's response pattern. In other words: the participant significantly differed from the

expected behavioural pattern. A significant difference meant in practical terms that the participant wasn't simply guessing *in terms of mapping*, but was using some sort of strategy or previous experience.

The chi-square analysis was run separately for each participant. All significant values were in the expected direction, which meant that each participant that differed from the expected response patterns differed by making *more* mapping-congruent responses than expected. There was one exception to this finding, which concerned a participant who used one article almost exclusively, and would already have been deleted from all analyses for having too many missing values.

The final result of the chi-square analyses were two groups of participants. In total, 18 participants of the 56 showed no evidence of using the mapping mapping in the first block, while 38 did show evidence of using the mapping. The 18 who showed no evidence of prior mapping knowledge were dubbed the 'NoFollow', or 'inexperienced' subgroup, for showing no signs of following the mapping prior to the experiment, and the 38 who did show evidence were dubbed the 'Follow' or 'experienced' group for following the mapping more than by random chance.

Appendix D. Explicit learning: mapping reports.

In this appendix, the Tables and Figures of the section on the 'explicit learning from input' from the main paper are repeated but with an alternative dependent variable. This variable is '*percentage of mapping responses*' rather than the '*percentage of correct article responses*' that was used in the main text. The purpose for this appendix is to provide a different perspective on the descriptive data, but not to delve into a lengthy discussion. Most of the conclusions are the same, or might have been partially masked due to a possible ceiling-like effect: in the Follow group, responses on compatible nouns especially were following the mapping in >85% in all production moments. Still, these descriptive statistics might give new insights that were not discussed in the main text, so they are nonetheless provided in this appendix.

One important reminder is that the *mapping* effects will be in the opposite direction for incompatible nouns, as a correct article for an incompatible noun means that the mapping was *not* followed. This was one of the reasons for choosing 'correct article usage' as the dependent variable in the main text.

The corresponding tables and figures of the main text are provided in the description of each table or figure. The delayed posttest (N = 18) results on explicit learning, seeing how participants behave after a period of six weeks after testing, are also provided below the results of all participants (N = 52).

Table D1. *Means, Standard Deviation and Range for the Follow and NoFollow subgroups, for all Word Categories (Incompatible and Compatible) over all production moments (1, 2, 3 and Posttest). This table corresponds to Table 3 in the main text.*

	Production Moment							
	Production 1		Production 2		Production 3		Posttest	
	M (SD)	*Range*	*M (SD)*	*Range*	*M (SD)*	*Range*	*M (SD)*	*Range*
Incompatible with Input								
Follow	65.0 (12.4)	40.9 – 89.5	65.6 (13.1)	47.6 – 91.3	43.7 (19.1)	0 – 88.2	50.6 (14.4)	21.7 – 83.3
NoFollow	51.4 (12.8)	33.3 – 69.6	53.3 (12.8)	33.3 – 71.4	46.0 (13.8)	21.7 – 69.2	47.3 (14.0)	23.8 – 77.3
Total	60.6 (14.0)	33.3 – 89.5	61.6 (14.2)	33.3 – 91.3	44.4 (17.5)	0 – 88.2	49.5 (14.2)	21.7 – 3.3
Incompatible without Input								
Follow	62.3 (12.8)	28.0 – 87.0	63.4 (12.6)	28.0 – 87.0	60.5 (13.2)	24.0 – 82.6	52.1 (12.3)	25.0 – 78.9
NoFollow	56.3 (14.1)	33.3 – 81.8	53.4 (13.8)	33.3 – 81.8	49.4 (13.9)	33.3 – 77.3	51.9 (14.9)	13.0 – 86.4
Total	60.4 (13.4)	28.0 – 87.0	60.2 (13.7)	28.0 – 87.0	56.9 (14.3)	24.0 – 82.6	51.9 (13.0)	13.0 – 86.4
Compatible with Input								
Follow	88.1 (8.0)	72.7 – 100	89.1 (8.0)	63.6 – 100	90.1 (7.2)	72.7 – 100	85.9 (8.1)	68.2 – 100
NoFollow	67.4 (15.1)	36.4 – 86.4	67.8 (13.7)	36.4 – 84.2	72.0 (17.6)	36.4 – 94.7	72.0 (13.0)	42.9 – 94.7
Total	81.3 (14.5)	36.4 – 100	82.1 (14.3)	36.4 – 100	84.2 (14.3)	36.4 – 100	81.4 (11.9)	42.9 – 100
Compatible without Input								
Follow	88.9 (10.6)	47.4 – 100	88.6 (9.8)	52.6 – 100	87.7 (10.6)	52.6 – 100	86.3 (8.1)	68.4 – 100
NoFollow	68.3 (14.6)	33.3 – 92.0	67.2 (15.7)	33.3 – 92.0	67.5 (16.8)	33.3 – 87.5	71.7 (10.6)	52.2 – 88.0
Total	82.1 (15.4)	33.3 – 100	81.6 (13.7)	33.3 – 100	81.1 (15.9)	33.3 – 100	81.2 (11.3)	52.2 – 100

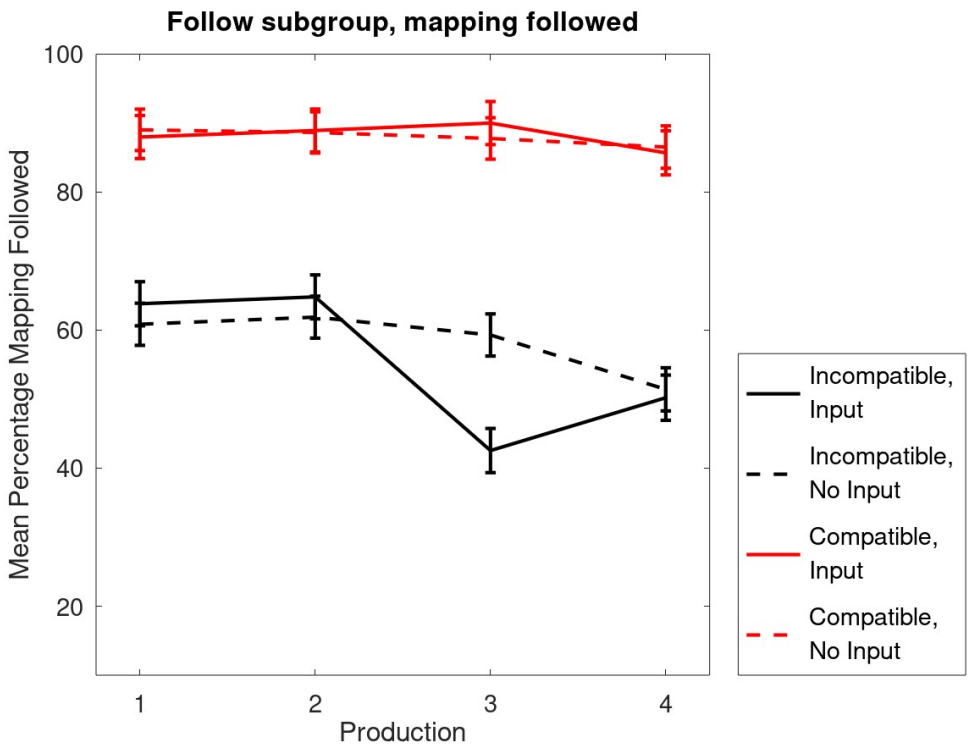

Figure D1. Mean percentage of mapping productions of only the Follow subgroup, given per Production moment (1, 2, 3 and Posttest), and per Word Category (Incompatible and Compatible), with the confidence intervals as error bars. This figure corresponds to Figure 5 in the main text.

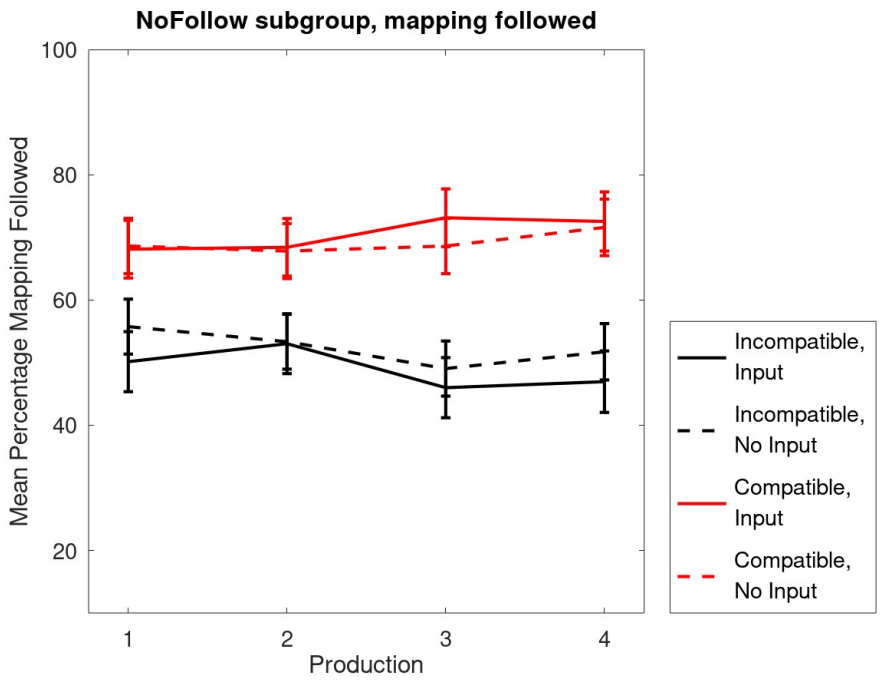

Figure D2. Mean percentage of mapping productions of only the NoFollow subgroup, given per Production moment (1, 2, 3 and Posttest), and per Word Category (Incompatible and Compatible), with the confidence intervals as error bars. This figure corresponds to Figure 6 in the main text.

Delayed Posttest

Below will follow the descriptive statistics and visualisations of the '*percentage of mapping followed'* dependent variable for the delayed posttest.

Table D2. *Means, Standard Deviation and Range for all participants together (N = 18), for all Word Categories (Incompatible and Compatible) over all production moments including the Delayed Posttest (1, 2, 3, Posttest and Delayed Posttest). This table corresponds to Table 4 in the main text.*

	Production Moment									
	Production 1		Production 2		Production 3		Posttest		Delayed Posttest	
	M (SD)	*Range*	*M (SD)*	*Range*	*M (SD)*	*Range*	*M (SD)*	*Range*	*M (SD)*	*Range*
Incompatible with input	66.0 (11.9)	33.3 – 85.7	64.1 (13.5)	33.3 – 88.9	48.2 (18.5)	0 – 69.2	52.9 (15.1)	21.7 – 83.3	51.8 (15.3)	30.0 – 85.7
Incompatible without input	64.0 (12.4)	30.4 – 78.9	64.6 (11.3)	30.4 – 79.2	62.8 (11.2)	30.4 – 79.2	52.4 (16.3)	13.0 – 86.4	54.5 (16.3)	26.3 – 83.3
Compatible with input	85.2 (12.4)	50.0 – 100	85.7 (13.9)	50.0 – 100	85.9 (13.7)	42.9 – 100	82.4 (11.3)	57.1 – 100	81.8 (13.4)	59.1 – 100
Compatible without input	86.9 (12.4)	57.9 – 100	87.1 (13.0)	52.6 – 100	84.1 (13.0)	50.0 – 96.0	81.5 (9.4)	59.1 – 94.7	85.0 (13.2)	54.5 – 100

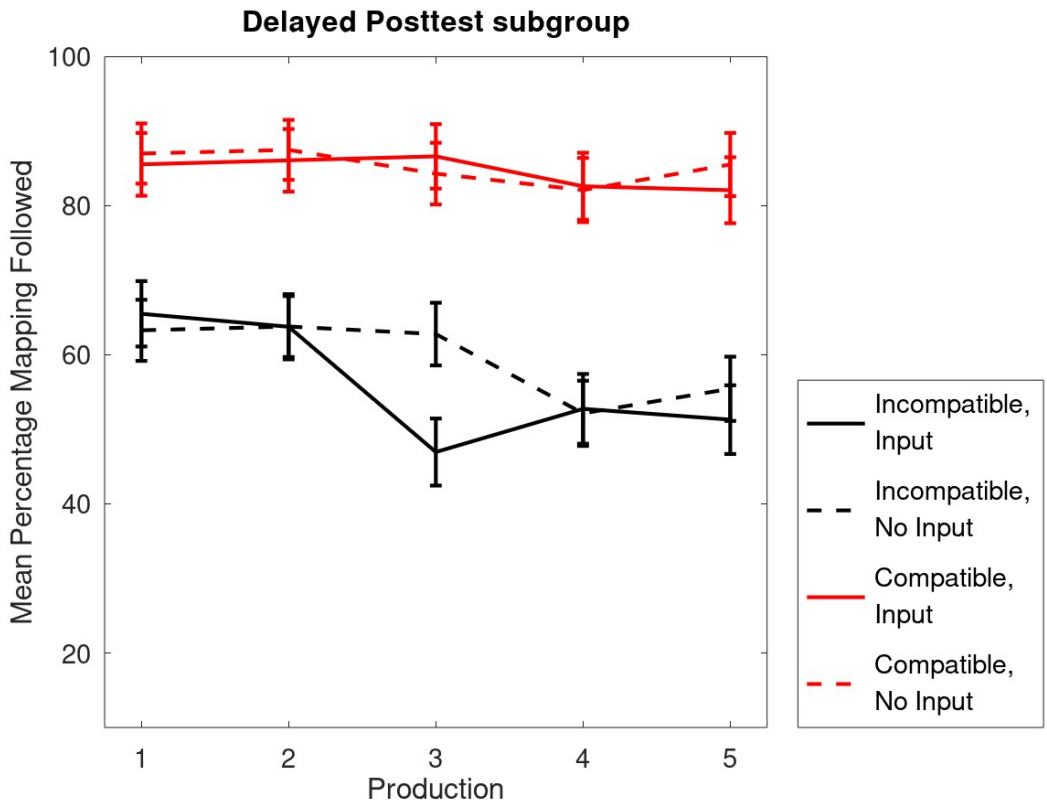

Figure D3. The Mean and SD of all participants that participated in the delayed posttest are plotted per noun category (Incompatible/Compatible Input/No Input) and Production moment (1, 2, 3, Posttest (4) and Delayed Posttest (5)). This figure corresponds to Figure 7 in the main text.

Appendix E. Consistency reports

To verify whether the main results were not caused by the inconsistencies in response patterns of participants, the previously reported analyses were repeated over the responses that were deemed to be *consistent*. *Consistent responses* were defined as only those responses where the participant used the same article for the 1st and 2nd production moments.

Only the most specific contrasts were repeated, as these were the most important for the conclusions drawn in the main text. One difference from the analyses in the main text, however, is that due to the selection of only consistent responses, the 1st and 2nd production moments were now identical. To compensate for this fact, the 1st production moment was dropped from the model. In this way, the results remained easily interpretable and straightforward. The remaining production moments were the 2nd production moment, 3rd production moment and posttest. The delayed posttest was not included in this appendix because of its limited scope and exploratory nature.

Figure E1 gives two plots, the right one is the same as Figure 5 in the main text, while the left plots the same contrasts but for only the consistent responses. It concerns the correct article analyses of the Follow subgroup, across all production moments. Although the 1st production moment *is* plotted in the left-hand plot, this production moment was dropped in the analyses for only consistent responses. The reason for still plotting it, is because this makes it easier to see the similarities and differences with the right-side plot of the results in the main text.

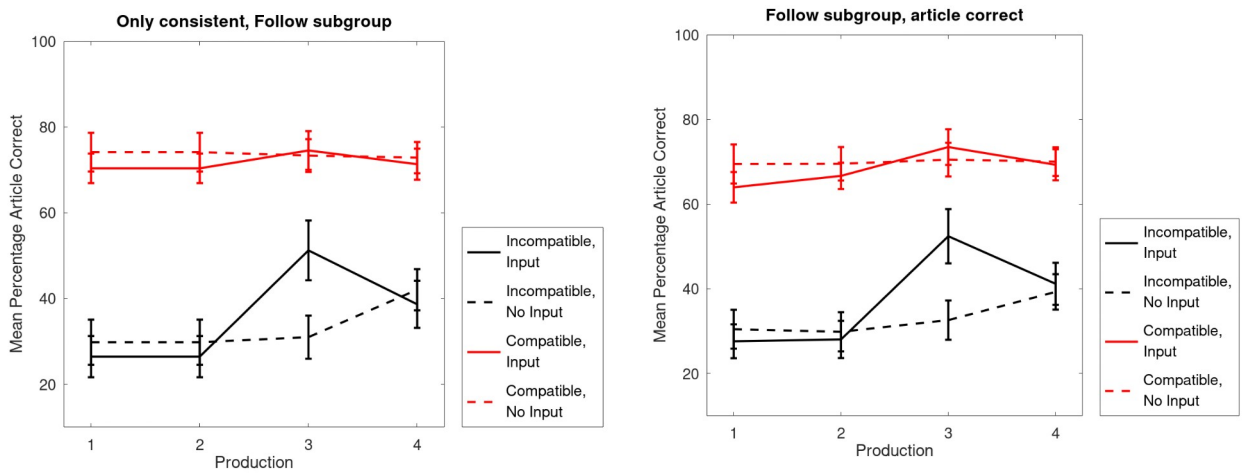

Figure E1. Left panel: The means of consistent responses are plotted per Word Category (Incompatible/Compatible Input/No Input) and Production (1, 2, 3, Posttest (4)) for the Follow subgroup. Right panel: the same information is plotted for all responses of the Follow subgroup (as in *Figure 5*).

The Follow group's Incompatible with Input category showed a significant main effect of Production ($F(2, 68) = 48.584$, $p < 0.001$, $\eta_p^2 = 0.588$). This finding is very similar to the findings in the main text, with only a lower effect size that does not lead to a difference in interpretation. The Bonferroni pairwise comparisons confirm the main text's findings as well, with the 2nd ($M = 26.4$, $SE = 2.4$) production moment showing a significant difference from the 3rd ($M = 51.2$, $SE = 3.4$, $\Delta M = 24.8$, $SE = 2.8$, $p < 0.001$) and 4th ($M = 38.6$, $SE = 2.7$, $\Delta M = 12.6$, $SE = 2.2$, $p < 0.001$), and the 3rd and 4th also showed a significant difference from one another ($p < 0.001$). This confirms that the analyses lead to the same conclusions regardless of the consistency of responses.

For the Incompatible without Input category, a similar significant main effect ($F(1.592, 54.115) = 28.022$, $p < 0.001$, $\eta_p^2 = 0.452$, *Greenhouse-Geisser corrected*) was found to the main text's analysis, including the unexpected significantly different posttest that drove the main effect. The effect size is larger, but does not warrant a different interpretation than before. Only the posttest ($M = 29.8$, $SE = 2.6$) was significantly different from the 2nd ($M = 29.8$, $SE = 2.6$, $\Delta M = 12.3$, $SE = 1.9$, $p < 0.001$) and 3rd ($M = 31.0$, $SE = 2.5$, $\Delta M = 11.1$, $SE = 2.1$, $p < 0.001$) production moments, while the 2nd and 3rd did not differ significantly from each other ($p = 1$). This, again, yields the exact same interpretations as in the previous analysis.

For the Compatible with Input category, the main effect of Production was not significant ($F(2, 68) = 2.691$, $p = 0.075$, $\eta_p^2 = 0.073$), which was essentially the same result as in the main text's analysis. There was also no significant main effect for the Compatible without Input category ($F(1.689, 57.426) = 0.383$, $p = 0.648$, $\eta_p^2 = 0.011$, *Greenhouse-Geisser corrected*). While for the Compatible without Input category an absence of an effect was initially expected, this absence was inconsistent with the main text's analysis. So, the previous Compatible without Input effect might have indeed been caused, or at least influenced, by the inconsistent responses. However, since the previous effect was rather small, the finding has a minimal effect on the conclusion. Nonetheless, this inconsistency is a warning for future research to tread carefully when investigating this effect further.

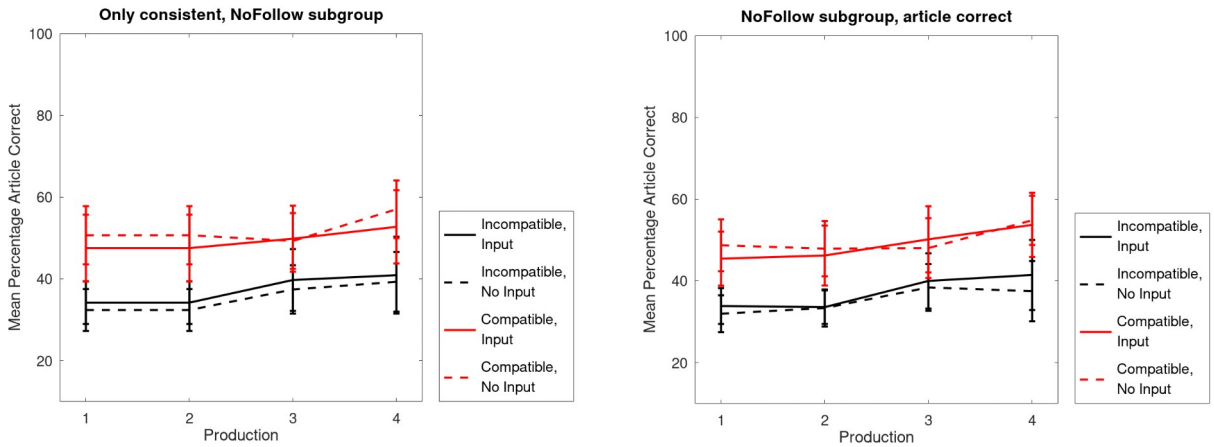

Figure E2. Left panel: The means of consistent responses are plotted per Word Category (Incompatible/Compatible Input/No Input) and Production (1, 2, 3, Posttest (4)) for the NoFollow subgroup. Right panel: the same information is plotted for all responses of the NoFollow subgroup (as in *Figure 6*).

For the NoFollow group, the analyses were not split in Input and No Input conditions as in the main text's analyses no main effect of Input was found, and this split similarly did not occur. For the Incompatible category, only a significant effect of Production was found ($F(1.4, 21.9) = 4.002$, $p = 0.047$, $\eta_p^2 = 0.2$, *Greenhouse-Geisser corrected*), but not of Input ($F(1, 16) = 0.487$, $p = 0.495$, $\eta_p^2 = 0.03$) and the interaction effect was also found to not be significant ($F(1.4, 22.2) = 0.016$, $p = 0.953$, $\eta_p^2 = 0.001$, *Greenhouse-Geisser corrected*). The pairwise-comparisons also show the same pattern, with the 1st and 2nd ($M = 33.3$, $SE = 1.8$) production moments being significantly different only from the 3rd ($M = 38.6$, $SE = 2.7$, $\Delta M = 5.3$, $SE = 1.4$, $p = 0.006$) production moment (all other $p > 0.08$). These conclusions are the same as was found in the main text's analysis.

For the Compatible category, only a significant effect of production was detected ($F(1.4, 22.3) = 4.568$, $p = 0.033$, $\eta_p^2 = 0.222$, *Greenhouse-Geisser corrected*), but not of input ($F(1, 16) = ,$ $p = 0.272$, $\eta_p^2 = 0.075$). The interaction effect between input and production was also not significant ($F(1.4, 22.8) = 0.899$, $p = 0.389$, $\eta_p^2 = 0.053$, *Greenhouse-Geisser corrected*), which means the main conclusions were again the same as in the main text. However, in contrast to the main text's analysis, the pairwise comparisons did not show any significant differences (all $p > 0.1$). In the main text's analysis, the detected difference was also close to the critical p-value, which implies that this effect was not very stable, or the 'only consistent' analysis needed more power to detect the exact differences. Nonetheless, the main conclusions from the main text's analysis remain the same, as the differences were minimal.